РОССИЙСКАЯ АКАДЕМИЯ НАУК
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР

СООБЩЕНИЯ ПО ПРИКЛАДНОЙ МАТЕМАТИКЕ

В. В. СТРИЖОВ

МЕТОДЫ ИНДУКТИВНОГО ПОРОЖДЕНИЯ РЕГРЕССИОННЫХ МОДЕЛЕЙ

ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР РАН
МОСКВА, 2008

УДК 519.584

Ответственный редактор
канд. физ.-матем. наук К. В. Воронцов

При решении задач линейной или нелинейной регрессии, искомая регрессионная модель может быть назначена аналитиком на основе предположений о характере решаемой задачи или выбрана из некоторого множества различных моделей. При выборе моделей встают вопросы о том, какова должна быть структура модели, ее сложность, устойчивость и точность. В монографии рассматриваются проблемы индуктивного порождения и выбора моделей, состоящих из суперпозиций параметрических функций.

Работа поддержана грантом РФФИ 07-07-00181 “Развитие теории поиска регрессионных моделей в неявно заданном множестве”.

Рецензенты: Ю. В. Чехович,
А. Г. Дьяконов

Научное издание

1. Введение

Регрессионный анализ — метод моделирования измеряемых данных и исследования их свойств. Данные состоят из пар значений зависимой переменной (переменной отклика) и независимой переменной (объясняющей переменной). Регрессионная модель есть функция независимой переменной и параметров с добавленной случайной переменной. Параметры модели настраиваются таким образом, что модель наилучшим образом приближает данные. Критерием качества приближения (целевой функцией) обычно является среднеквадратичная ошибка: сумма квадратов разности значений модели и зависимой переменной для всех значений независимой переменной в качестве аргумента.

Предполагается, что зависимая переменная есть сумма значений некоторой модели и случайной величины. Относительно характера распределения этой величины делаются распределения, называемые гипотезой порождения данных. Для подтверждения или опровержения этой гипотезы выполняются статистические тесты, называемые анализом регрессионных остатков. При этом предполагается, что независимая переменная не содержит ошибок.

Регрессионный анализ используется для прогноза, анализа временных рядов, тестирования гипотез и выявления скрытых взаимосвязей в данных. Регрессионный анализ — раздел математической статистики машинного обучения.

1.1. Определение регрессии

Регрессия — зависимость математического ожидания (например, среднего значения) случайной величины от одной или нескольких других случайных величин (свободных переменных), то есть $E(y|\mathbf{x}) = f(\mathbf{x})$. Регрессионным анализом называется поиск такой функции f, которая описывает эту зависимость. Регрессия может быть представлена в виде суммы неслучайной и случайной составляющих

$$y = f(\mathbf{x}) + \nu,$$

где f — функция регрессионной зависимости, а ν — аддитивная случайная величина с нулевым матожиданием. Предположение о характере распределения этой величины называется гипотезой порождения

данных. Обычно предполагается, что величина ν имеет гауссово распределение с нулевым средним и дисперсией σ_ν^2.

Задача нахождения регрессионной модели нескольких свободных переменных ставится следующим образом. Задана выборка — множество $\{\mathbf{x}_1, ..., \mathbf{x}_N | \mathbf{x} \in \mathbb{R}^M\}$ значений свободных переменных и множество $\{y_1, ..., y_N | y \in \mathbb{R}\}$ соответствующих им значений зависимой переменной. Эти множества обозначаются как D, множество исходных данных $\{(\mathbf{x}, y)_i\}$. Задана регрессионная модель — параметрическое семейство функций $f(\mathbf{w}, \mathbf{x})$ зависящая от параметров $\mathbf{w} \in \mathbb{R}$ и свободных переменных $\mathbf{x}$. Требуется найти наиболее вероятные параметры $\bar{\mathbf{w}}$:

$$\bar{\mathbf{w}} = \arg \max_{\mathbf{w} \in \mathbb{R}^W} p(y|x, \mathbf{w}, f) = p(D|\mathbf{w}, f).$$

Функция вероятности p зависит от гипотезы порождения данных и задается Байесовским выводом или методом наибольшего правдоподобия.

1.2. Линейная регрессия

Линейная регрессия предполагает, что функция f зависит от параметров $\mathbf{w}$ линейно. При этом линейная зависимость от свободной переменной $\mathbf{x}$ необязательна,

$$y = f(\mathbf{w}, \mathbf{x}) + \nu = \sum_{j=1}^{W} w_j g_j(\mathbf{x}) + \nu.$$

В случае, когда функция $g \equiv \mathrm{id}$ линейная регрессия имеет вид

$$y = \sum_{j=1}^{N} w_j x_j + \nu = \langle \mathbf{w}, \mathbf{x} \rangle + \nu,$$

здесь x_j — компоненты вектора $\mathbf{x}$.

Значения параметров в случае линейной регрессии находят с помощью метода наименьших квадратов. Использование этого метода обосновано предположением о гауссовском распределении случайной переменной. При этом одна из важных оценок критерия качества полученной зависимости определена как

$$\mathrm{SSE} = \|f(\mathbf{x}_i) - y_i\|_2 = \sum_{i=1}^{N} (y_i - f(\mathbf{w}, \mathbf{x}_i))^2$$

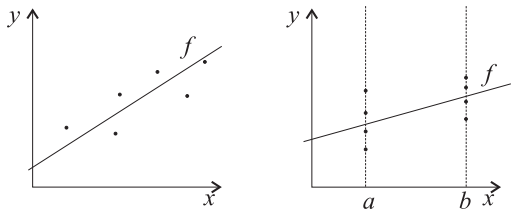

Рис. 1. Выборка может быть не функцией, а отношением. Например, данные для построения регрессии могут быть такими: $\{(0,0),(0,1),(0,2),(1,1),(1,2),(1,3)\}$ (рис. справа). В такой выборке одному значению переменной x соответствует несколько значений переменной y.

и называется сумма квадратов ошибок, SSE — Sum of Squared Errors. Разность между фактическим и вычисленным значением зависимой переменной $f(\mathbf{x}_i) - y_i$ называется невязкой. Вектор невязок обозначается $\mathbf{f}(\mathbf{x}) - \mathbf{y}$. Этот вектор также называется вектором регрессионных остатков (residuals). Анализу этого вектора посвящен отдельный раздел данной области, который называется «анализ регрессионных остатков». В частности в него входит вычисление дисперсии остатков:

$$\bar{\sigma}_\nu^2 = \frac{\text{SSE}}{N-2} = \text{MSE}.$$

Здесь MSE — Mean Square Error, среднеквадратичная ошибка.

На графиках представлены выборки, обозначенные синими точками, и регрессионные зависимости, обозначенные сплошными линиями. По оси абсцисс отложена свободная переменная, а по оси ординат — зависимая. Все три зависимости линейны относительно параметров.

1.3. О терминах

Термин «регрессия» был введен Фрэнсисом Гальтоном в конце 19-го века. Гальтон обнаружил, что дети родителей с высоким или низким ростом обычно не наследуют выдающийся рост и назвал этот феномен «регрессия к посредственност». Сначала этот термин использовался

исключительно в биологическом смысле. После работ Карла Пирсона этот термин стали использовать и в статистике.

В статистической литературе различают регрессию с участием одной свободной переменной и с несколькими свободными переменными — *одномерную* и *многомерную* регрессию. Предполагается, что мы используем несколько свободных переменных, то есть, свободная переменная — вектор $\mathbf{x} \in \mathbb{R}^N$. В частных случаях, когда свободная переменная является скаляром, она будет обозначаться x. Различают *линейную* и *нелинейную* регрессию. Если регрессионную модель не является линейной комбинацией функций свободных переменных, то говорят о нелинейной регрессии. При этом модель может быть произвольной суперпозицией функций свободных переменных g из некоторого набора. Нелинейными моделями являются, экспоненциальные, тригонометрические, и другие (например, радиальные базисные функции или персептрон Розенблатта), полагающие зависимость между параметрами и зависимой переменной нелинейной.

Различают *параметрическую* и *непараметрическую* регрессию. Строгую границу между этими двумя типами регрессий провести сложно. Сейчас нет существует общепринятого критерия отличия одного типа моделей от другого. Например, считается, что линейные модели являются параметрическими, а модели, включающие усреднение зависимой переменной по пространству свободной переменной — непараметрическими. Пример параметрической регресионной модели: линейный предиктор, многослойный персептрон. Примеры смешанной регрессионной модели: функции радиального базиса. Непараметрическая модель — скользящее усреднение в окне некоторой ширины. В целом, непараметрическая регрессия отличается от параметрической тем, что зависимая переменная зависит не от одного значения свободной переменной, а от некоторой заданной окрестности этого значения.

Есть различие между терминами: “приближение функций”, “аппроксимация”, “интерполяция”, и “регрессия”. Оно заключается в следующем.

Приближение функций. Дана функция u дискретного или непрерывного аргумента. Требуется найти функцию f из некоторого параметрическую семейства, например, среди алгебраических полиномов заданной степени. Параметры функции f должны доставлять мини-

мум некоторому функционалу, например,

$$\rho(f,u) = \left(\frac{1}{b-a}\int_a^b |f(x)-u(x)|^2 dx\right)^{\frac{1}{2}}.$$

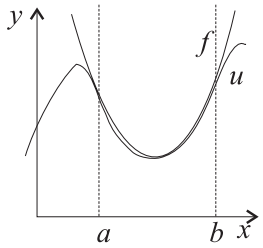

Рис. 2. Аппроксимация функций: непрерывная функция f приближает непрерывную или дискретную функцию u

Термин *аппроксимация* — синоним термина "приближение функций". Чаще используется тогда, когда речь идет о заданной функции, как о функции дискретного аргумента. Здесь также требуется отыскать такую функцию f, которая проходит наиболее близко ко всем точкам заданной функции. При этом вводится понятие *невязки* — расстояния между точками непрерывной функции f и соответствующими точками функции u дискретного аргумента.

Интерполяция функций — частный случай задачи приближения, когда требуется, чтобы в определенных точках, называемых узлами интерполяции совпадали значения функции u и приближающей ее функции f. В более общем случае накладываются ограничения на значения некоторых производных f производных. То есть, дана функция u дискретного аргумента. Требуется отыскать такую функцию f, которая проходит через все точки u. При этом метрика обычно не используется, однако часто вводится понятие"гладкости" искомой функции.

Регрессия и классификация тесно связаны друг с другом. Термин *алгоритм* в классификации мог бы стать синонимом термина *модель* в регрессии, если бы алгоритм не оперировал с дискретным множеством ответов-классов, а модель — с непрерывно-определенной свободной переменной.

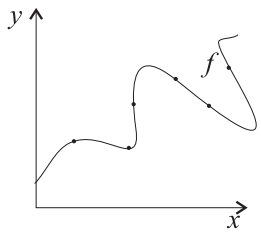

Рис. 3. Интерполяция: функция f задана значениями узловых точек

1.4. Регрессионная модель

Термину *регрессионная модель*, используемому в регрессионном анализе, можно сопоставить синонимы: «теория», «гипотеза». Эти термины пришли из статистики, в частности из раздела «проверка статистических гипотез». Регрессионная модель есть прежде всего гипотеза, которая должна быть подвергнута статистической проверке, после чего она принимается или отвергается.

Регрессионная модель $f(\mathbf{w}, \mathbf{x})$ — это параметрическое семейство функций, задающее отображение

$$f : W \times X \longrightarrow Y,$$

где $\mathbf{w} \in W$ — пространтсво параметров, $\mathbf{x} \in X$ — пространство свободных переменных, Y — пространство зависимых переменных.

Так как регрессионный анализ предполагает поиск зависимости матожидания случайной величины от свободных переменных $E(y|\mathbf{x}) = f(\mathbf{x})$, то в её состав входит аддитивная случайная величина ε:

$$y = f(\mathbf{w}, \mathbf{x}) + \varepsilon.$$

Предположение о характере распределения случайной величины ν называются гипотезой порождения данных. Эта гипотеза играет центральную роль в выборе критерия оценки качества модели и, как следствие, в способе настройки параметров модели.

Модель является настроенной (идентифицированной, обученной) когда зафиксированы её параметры, то есть модель задаёт отображение

$$f|_{\bar{\mathbf{w}}} : X \longrightarrow Y$$

для фиксированного значения $\bar{\mathbf{w}}$ (выражение читается «сужение области определения функции f»).

Различают «математическую модель» и «регрессионную модель». Математическая модель предполагает участие аналитика в конструировании функции, которая описывает некоторую известную закономерность. Математическая модель является интерпретируемой — объясняемой в рамках исследуемой закономерности. При построении математической модели сначала создаётся параметрическое семейство функций, затем с помощью измеряемых данных выполняется «идентификация модели» — нахождение её параметров. Известная функциональная зависимость объясняющей переменной и переменной отклика — основное отличие математического моделирования от регрессионного анализа. Недостаток математического моделирования состоит в том, что измеряемые данные используются для верификации, но не для построения модели, вследствие чего можно получить неадекватную модель. Также затруднительно получить модель сложного явления, в котором взаимосвязано большое число различных факторов.

Регрессионная модель объединяет широкий класс универсальных функций, которые описывают некоторую закономерность. При этом для построения модели в основном используются измеряемые данные, а не знание свойств исследуемой закономерности. Такая модель часто неинтерпретируема, но более точна. Это объясняется либо большим числом моделей-претендентов, которые используются для построения оптимальной модели, либо большой сложностью модели. Нахождение параметров регрессионной модели называется «обучением модели».

Недостатки регрессионного анализа: модели, имеющие слишком малую сложность, могут оказаться неточными, а модели, имеющие избыточную сложность, могут оказаться «переобученными».

Примеры регрессионных моделей: линейные функции, алгебраические полиномы, ряды Чебышёва, нейронные сети без обратной связи, например, однослойный персептрон Розенблатта, радиальные базисные функции и прочее.

И регрессионная, и математическая модель, как правило, задают непрерывное отображение. Требование непрерывности обусловлено классом решаемых задач: чаще всего это описание физических, химических и других явлений, где требование непрерывности выставляется естественным образом. Иногда на отображение f накладываться ограничения монотонности, гладкости, измеримости, и некоторые

другие. Теоретически, никто не запрещает работать с функциями произвольного вида, и допускать в моделях существование не только точек разрыва, но и задавать конечное, неупорядоченное множество значений свободной переменной, то есть, превращать задачи регрессии в задачи классификации.

При решении задач регрессионного анализа встают следующие вопросы.

- Как выбрать тип и структуру модели, какому именно семейству она должна принадлежать?
- Какова гипотеза порождения данных, каково распределение случайной переменной?
- Какой целевой функцией оценить качество аппроксимации?
- Каким способом отыскать параметры модели, каков должен быть алгоритм оптимизации параметров?

2. Линейные методы

Метод наименьших квадратов метод нахождения оптимальных параметров моделей линейных регрессии, таких, что сумма квадратов невязок регрессионных остатков минимальна. Метод заключается в минимизации евклидова расстояния $\|A\mathbf{w} - \mathbf{y}\|$ между двумя векторами.

2.1. Метод наименьших квадратов

Задача метода наименьших квадратов состоит в выборе вектора $\mathbf{w}$, минимизирующего ошибку (длину вектора невязки) $S = \|A\mathbf{w} - \mathbf{y}\|^2$. Эта ошибка есть расстояние от вектора $\mathbf{y}$ до вектора $A\mathbf{w}$. Вектор $A\mathbf{w}$ лежит в простанстве столбцов матрицы A, так как $A\mathbf{w}$ есть линейная комбинация столбцов этой матрицы с коэффициентами $w_1, ..., w_N$. Отыскание решения $\mathbf{w}$ по методу наименьших квадратов эквивалентно задаче отыскания такой точки $\mathbf{p} = A\mathbf{w}$, которая лежит ближе всего (евклидовой метрике) к $\mathbf{y}$ и находится при этом в пространстве столбцов матрицы A. Таким образом, вектор $\mathbf{p}$ должен быть проекцией $\mathbf{y}$

на пространство столбцов и вектор невязки $A\mathbf{w} - \mathbf{y}$ должен быть ортогонален этому пространству. Ортогональность состоит в том, что каждый вектор в пространстве столбцов есть линейная комбинация столбцов с некоторыми коэффициентами $v_1, ..., v_N$, то есть это вектор $A\mathbf{v}$. Для всех v в пространстве $A\mathbf{v}$, эти векторы должны быть перпендикулярны невязке $A\mathbf{w} - \mathbf{y}$:

$$(A\mathbf{v})^T(A\mathbf{w} - \mathbf{y}) = \mathbf{v}^T(A^TA\mathbf{w} - A^T\mathbf{y}) = 0.$$

Так как это равентсво должно быть справедливо для произвольного вектора $\mathbf{v}$, то

$$A^TA\mathbf{w} - A^T\mathbf{y} = 0.$$

Решение по методу наименьших квадратов несовместной системы $A\mathbf{w} = \mathbf{y}$, состоящей из M уравнений с N неизвестными, есть уравнение

$$A^TA\mathbf{w} = A^T\mathbf{y},$$

которое называется *нормальным уравнением*. Если столбцы матрицы A линейно независимы, то матрица A^TA обратима и единственное решение

$$\mathbf{w} = (A^TA)^{-1}A^T\mathbf{y}.$$

Проекция вектора $\mathbf{y}$ на пространство столбцов матрицы имеет вид

$$\mathbf{p} = A\mathbf{w} = A(A^TA)^{-1}A^T\mathbf{y} = P\mathbf{y}.$$

Матрица $P = A(A^TA)^{-1}A^T$ называется матрицей проектирования вектора $\mathbf{y}$ на пространство столбцов матрицы A. Эта матрица имеет два основных свойства: она идемпотентна, $P^2 = P$, и симметрична, $P^T = P$. Обратное также верно: матрица, обладающая этими двумя свойствами есть матрица проектирования на свое пространство столбцов.

2.2. Пример постоения линейной регрессии

Задана выборка — таблица

$$D = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \dots & \dots \\ x_M & y_M \end{pmatrix}.$$

Задана регрессионная модель — квадратичный полином

$$f = w_3 x^2 + w_2 x + w_1 = \sum_{j=1}^{3} w_j x^{j-1}.$$

Назначенная модель является линейной. Для нахождения оптимального значения вектора параметров $\mathbf{w} = \langle w_1, ..., w_3 \rangle^T$ выполняется следующая подстановка:

$$x_i^0 \mapsto a_{i1}, x_i^1 \mapsto a_{i2}, x_i^2 \mapsto a_{i3}.$$

Тогда матрица A значений подстановок свободной переменной x_i будет иметь вид

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \cdots & \cdots & \cdots \\ a_{M1} & a_{M2} & a_{M3} \end{pmatrix}.$$

Задан критерий качества модели: функця ошибки

$$S = \sum_{i=1}^{M} (f(\mathbf{w}, x_i) - y_i)^2 = \|A\mathbf{w} - \mathbf{y}\|^2 \longrightarrow \min.$$

Здесь вектор $\mathbf{y} = \langle y_1, \ldots, y_M \rangle$. Требуется найти такие параметры $\mathbf{w}$, которые бы доставляли минимум этому функционалу,

$$\mathbf{w} = \arg \min_{\mathbf{w} \in \mathbb{R}^3} (S).$$

Требуется найти такие параметры $\mathbf{w}$, которые доставляют минимум S — норме вектора невязок $A\mathbf{w} - \mathbf{y}$.

$$\begin{aligned} S = \|A\mathbf{w} - \mathbf{y}\|^2 = (A\mathbf{w} - \mathbf{y})^T (A\mathbf{w} - \mathbf{y}) = \\ = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T A\mathbf{w} - \mathbf{w}^T A^T \mathbf{y} + \mathbf{w}^T A^T A\mathbf{w} = \\ = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T A\mathbf{w} + \mathbf{w}^T A^T A\mathbf{w}. \end{aligned}$$

Для того, чтобы найти минимум функции невязки, требуется приравнять ее производные к нулю. Производные данной функции по $\mathbf{w}$ составляют

$$\frac{\partial S}{\partial \mathbf{w}} = -2A^T\mathbf{y} + 2A^T A\mathbf{w} = 0.$$

Это выражение также называется нормальным уравнением. Решение этой задачи должно удовлетворять системе линейных уравнений

$$A^T A\mathbf{w} = A^T\mathbf{y},$$

то есть,

$$\mathbf{w} = (A^T A)^{-1}(A^T\mathbf{y}).$$

После получения весов можно построить график найденной функции.

При обращении матрицы $(A^T A)^{-1}$ предполагается, что эта матрица невырождена.

2.3. Сингулярное разложение

Сингулярное разложение (Singular Value Decomposition, SVD) — декомпозиция вещественной матрицы с целью ее приведения к каноническому виду. Сингулярное разложение является удобным методом при работе с матрицами. Оно показывает геометрическую структуру матрицы и позволяет наглядно представить имеющиеся данные. Сингулярное разложение используется при решении самых разных задач — от приближения методом наименьших квадратов и решения систем уравнений до сжатия изображений. При этом используются разные свойства сингулярного разложения, например, способность показывать ранг матрицы, приближать матрицы данного ранга. SVD позволяет вычислять обратные и псевдообратные матрицы большого размера, что делает его полезным инструментом при решении задач регрессионного анализа.

Для любой вещественной $(n \times n)$-матрицы A существуют две вещественные ортогональные $(n \times n)$-матрицы U и V такие, что $U^T AV$ — диагональная матрица Λ,

$$U^T AV = \Lambda.$$

Матрицы U и V выбираются так, чтобы диагональные элементы матрицы Λ имели вид

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > \lambda_{r+1} = ... = \lambda_n = 0,$$

где r — ранг матрицы A. В частности, если A невырождена, то

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n > 0.$$

Индекс r элемента λ_r есть фактическая размерность собственного пространства матрицы A.

Столбцы матриц U и V называются соответственно левыми и правыми сингулярными векторами, а значения диагонали матрицы Λ называются сингулярными числами.

Эквивалентная запись сингулярного разложения — $A = U\Lambda V^T$.

Например, матрица

$$A = \begin{pmatrix} 0.96 & 1.72 \\ 2.28 & 0.96 \end{pmatrix}$$

имеет сингулярное разложение

$$A = U\Lambda V^T = \begin{pmatrix} 0.6 & 0.8 \\ 0.8 & -0.6 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0.8 & -0.6 \\ 0.6 & 0.8 \end{pmatrix}^T$$

Легко увидеть, что матрицы U и V ортогональны,

$$U^T U = UU^T = I, \text{также } V^T V = VV^T = I,$$

и сумма квадратов значений их столбцов равна единице.

2.3.a. Геометрический смысл SVD

Пусть матрице A поставлен в соответствие линейный оператор. Сингулярное разложение можно переформулировать в геометрических терминах. Линейный оператор, отображающий элементы пространства $\mathbb{R}^n$ в себя представим в виде последовательно выполняемых линейных операторов вращения, растяжения и вращения. Поэтому компоненты сингулярного разложения наглядно показывают геометрические изменения при отображении линейным оператором A множества векторов из векторного пространства в себя или в векторное пространство другой размерности.

2.3.b. Пространства матрицы и SVD

Сингулярное разложение позволяет найти ортогональные базисы различных векторных пространств разлагаемой матрицы

$$A_{(n\times n)} = U_{(n\times n)}\Lambda_{(n\times n)}V^T_{(n\times n)}.$$

Для прямоугольных матриц существует так называемое экономное представление сингулярного разложения матрицы.

$$A_{(m\times n)} = U_{(m\times m)}\Lambda_{(m\times n)}V^T_{(n\times n)}$$

Согласно этому представлению при $m > n$, диагональная матрица Λ имеет пустые строки (их элементы равны нулю), а при $m < n$ — пустые столбцы. Поэтому существует еще одно экономное представление

$$A_{(m\times n)} = U_{(m\times r)}\Lambda_{(r\times r)}V^T_{(r\times n)},$$

в котором $r = \min(m, n)$.

Нуль-пространство матрицы A — набор векторов $\mathbf{x}$, для которого справедливо высказывание $A\mathbf{x} = \mathbf{0}$. Собственное пространство матрицы A — набор векторов $\mathbf{b}$, при котором уравнение $A\mathbf{x} = \mathbf{b}$ имеет ненулевое решение для $\mathbf{x}$. Обозначим $\mathbf{u}k$ и $\mathbf{v}k$ — столбцы матриц U и V. Тогда разложение $A = U\Lambda V^T$ может быть записано в виде: $A = \sum_{k=1}^{r} A_k$, где $A_k = \mathbf{u}_k\lambda_k\mathbf{v}_k^T$. Если сингулярное число $\lambda_k = 0$, то $A\mathbf{v}_k = \mathbf{0}$ и $\mathbf{v}_k$ находится в нуль-пространстве матрицы A, а если сингулярное число $\lambda_k \neq 0$, то вектор $\mathbf{u}_k$ находятся в собственном пространстве матрицы A. Следовательно, можно сконструировать базисы для различных векторных подпространств, определенных матрицей A. Набор векторов $\mathbf{v}_1, \ldots, \mathbf{v}_k$ в векторном пространстве V формирует базис линейного пространства|базис для V, если любой вектор $\mathbf{x}$ из V можно представить в виде линейной комбинации векторов $\mathbf{v}_1, \ldots, \mathbf{v}_k$ единственным способом. Пусть V_0 будет набором тех столбцов $\mathbf{v}k$, для которых $\lambda_k \neq 0$, а V_1 — все остальные столбцы $\mathbf{v}k$. Также, пусть U_0 будет набором столбцов $\mathbf{u}k$, для которых $\lambda_k \neq 0$, а U_1 — все остальные столбцы $\mathbf{u}k$, включая и те, для которых $k > n$. Тогда, если r — количество ненулевых сингулярных чисел, то имеется r столбцов в наборе V_0 и $n - r$ столбцов в наборе V_1 и U_1, а также $m - n + r$ столбцов в наборе U_0. Каждый из этих наборов формирует базис векторного пространства матрицы A:

- V_0 — ортонормальный базис для ортогонального комплементарного нуль-пространства A,
- V_1 — ортонормальный базис для нуль-пространства A,
- U_0 — ортонормальный базис для собственного пространства A,

- U_1 — ортонормальный базис для ортогонального комплементарного нуль-пространства A.

2.3.c. SVD и собственные числа матрицы

Сингулярное разложение обладает свойством, которое связывает задачу отыскания сингулярного разложения и задачу отыскания собственных векторов. Собственный вектор $\mathbf{x}$ матрицы A — такой вектор, при котором выполняется условие $A\mathbf{x} = \lambda\mathbf{x}$, число λ называется собственным числом. Так как матрицы U и V ортогональные, то

$$AA^T = U\Lambda V^T V\Lambda U^T = U\Lambda^2 U^T,$$
$$A^T A = V\Lambda U^T U\Lambda V^T = V\Lambda^2 V^T.$$

Умножая оба выражения справа соответственно на U и V получаем

$$AA^T U = U\Lambda^2,$$
$$A^T AV = V\Lambda^2.$$

Из этого следует, что столбцы матрицы U являются собственными векторами матрицы AA^T, а квадраты сингулярных чисел $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_r)$ — ее собственными числами. Также столбцы матрицы V являются собственными векторами матрицы $A^T A$, а квадраты сингулярных чисел являются ее собственными числами.

2.3.d. SVD и норма матриц

Рассмотрим изменение длины вектора $\mathbf{x}$ до и после его умножения слева на матрицу A. Евклидова норма вектора определена как

$$\|\mathbf{x}\|_E = \mathbf{x}^T\mathbf{x}.$$

Если матрица A ортогональна, длина вектора $A\mathbf{x}$ остается неизменной. В противном случае можно вычислить, насколько матрица A растянула вектор $\mathbf{x}$.

Евклидова норма матрицы есть максимальный коэффициент растяжения произвольного вектора $\mathbf{x}$ заданной матрицей A

$$\|A\|_E = \max_{\|\mathbf{x}\|=1} \left(\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \right).$$

Альтернативой Евклидовой норме является норма Фробениуса:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}.$$

Если известно сингулярное разложение, то обе эти нормы легко вычислить. Пусть $\lambda_1, \ldots, \lambda_r$ — сингулярные числа матрицы A, отличные от нуля. Тогда

$$\|A\|_E = \lambda_1,$$

и

$$\|A\|_F = \sqrt{\sum_{k=1}^{r} \lambda_k^2}.$$

Сингулярные числа матрицы A — это длины осей эллипсоида, заданного множеством

$$\{A\mathbf{x} | \, \|\mathbf{x}\|_E = 1\}.$$

2.3.e. Нахождение псевдообратной матрицы с помощью SVD

Если $(m \times n)$-матрица A является вырожденной или прямоугольной, то обратной матрицы A^{-1} для нее не существует. Однако для A может быть найдена псевдообратная

матрица A^+ — такая матрица, для которой выполняются условия

$$\begin{aligned} &A^+A = I_n, \\ &AA^+ = I_m, \\ &A^+AA^+ = A^+, \\ &AA^+A = A. \end{aligned}$$

Пусть найдено разложение матрицы A вида

$$A = U\Lambda V^T,$$

где $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_r)$, $r = \min(m, n)$ и $U^TU = I_m, VV^T = I_n$. Тогда матрица $A^+ = V^T\Lambda^{-1}U$ является для матрицы A псевдообратной. Действительно, $A^+A = V\Lambda^{-1}U^TU\Lambda V^T = I_n$, $AA^+ = U\Lambda V^TV\Lambda^{-1}U^T = I_m$.

2.3.f. Метод наименьших квадратов и число обусловленности

Задача наименьших квадратов ставиться следующим образом. Даны действительная $(m\times n)$-матрица A и действительный (m)-вектор Y. Требуется найти действительный (n)-вектор $\mathbf{w}$, минимизирующий Евклидову длину вектора невязки,

$$\|Y - A\mathbf{w}\|_E \longrightarrow \min .$$

Решение задачи наименьших квадратов —

$$\mathbf{w} = (A^T A)^{-1}(A^T Y).$$

Для отыскания решения $\mathbf{w}$ требуется обратить матрицу $A^T A$. Для квадратных матриц A число обусловленности $æ(A)$ определено отношением

$$æ(A) = \|A\|_E \|A^{-1}\|_E.$$

Из формулы Евклидовой нормы матрицы и предыдущей формулы следует, что число обусловленности матрицы есть отношение ее первого сингулярного числа к последнему.

$$æ(A) = \frac{\lambda_1}{\lambda_n}.$$

Следовательно, число обусловленности матрицы $A^T A$ есть квадрат числа обусловленности матрицы A. Это высказывание справедливо и для вырожденных матриц, если полагать число обусловленности как отношение λ_1/λ_r, r — ранг матрицы A. Поэтому для получения обращения, устойчивого к малым изменениям значений матрицы A, используется усеченное SVD.

2.3.g. Усеченное SVD при обращении матриц

Пусть матрица A представлена в виде $A = U\Lambda V^T$. Тогда при нахождении обратной матрицы $A^+ = V\Lambda^{-1}U^T$ в силу ортогональности матриц U и V и в силу условия убывания диагональных элементов матрицы $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$, псевдообратная матрица A^+ будет более

зависеть от тех элементов матрицы Λ, которые имеют меньшие значения, чем от первых сингулярных чисел. Действительно, если матрица A имеет сингулярные числа $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$, то сингулярные числа матрицы A^+ равны

$$\Lambda^{-1} = \text{diag}(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_n}) \text{ и } \frac{1}{\lambda_1} \leq \frac{1}{\lambda_2} ... \leq \frac{1}{\lambda_n}.$$

Считая первые s сингулярных чисел определяющими собственное пространство матрицы A, используем при обращении матрицы A первые s сингулярных чисел, $s \leq \text{rank}A$. Тогда обратная матрица A^+ будет найдена как $A^+ = V\Lambda_s^{-1}U^T$.

Определим усеченную псевдообратную матрицу A_s^+ как

$$A_s^+ = V\Lambda_s^{-1}U^T,$$

где $\Lambda_s^{-1} = \text{diag}(\lambda_1^{-1}, ..., \lambda_s^{-1}, 0, ..., 0)$ — $(n \times n)$-диагональная матрица.

2.4. Использование SVD для анализа временных рядов

Рассмотрим пример-иллюстрацию использования сингулярного разложения. Жизнь биосистемы описывается набором параметров, образующих фазовое пространство. Например пусть x_1, x_2 — концентрация кислорода в крови и частота сердечных сокращений пациента. Эти параметры, изменяясь во времени, образуют траекторию его жизни. Фазовое пространство разбито на три непересекающихся области: жизни $\mathcal{A}$ — *alive*, смерти $\mathcal{D}$ — *dead* и границу между ними $\mathcal{B}$ — *boundary*, рис. 4. Гипотеза: в точке, максимально удаленной от границ $\mathcal{B}$ внутри области $\mathcal{A}$ энтропия системы максимальна; в то время как у границы поведение системы становится ригидным, жестким, эффективная размерность траектории снижается.

Под эффективной размерностью матрицы будем понимать количество сингулярных чисел, превосходящих заданное λ_r. Для выяснения эффективной размерности траектории на интервале времени используется сингулярное разложение как наиболее удобный инструмент. Строки разлагаемой матрицы — последовательные вектора состояний системы в фазовом пространстве. Количество сингулярных чисел, больших λ_r есть эффективная размерность сегмента траектории. На рис. 9 показан пример — траектория системы с аттрактором

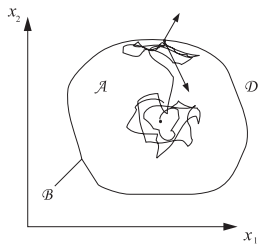

Рис. 4. Поведение биосистемы в экстремальных условиях

Лоренца и одно из подмножеств ее сегментов, находящихся в пространстве меньшей размерности.

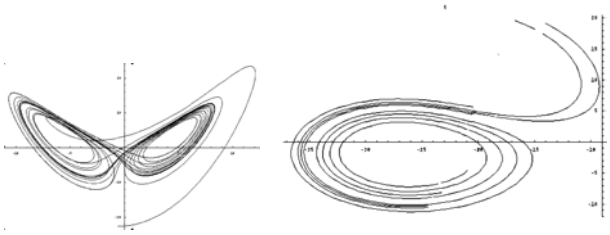

Рис. 5. Траектория системы с аттрактором Лоренца и подмножество ее сегментов

Разбиение множества $\{\mathbf{a}_{i\cdot}\}$ на кластеры зависит от размерности подпространства $\mathbb{R}^r \in \mathbb{R}^n$, в котором находятся кластеры, и от задаваемых требований к искомой кластеризации. Пусть Дана $(m \times n)$-матрица A упорядоченных векторов-строк $\{\mathbf{a}_{i\cdot}\}, i \in I = \{1, \ldots, m\}$. Матрица соответствует фазовой траектории дискретного времени i. Требуется найти разбиение фазовой траектории на сегменты, находящиеся в подпространстве заданной размерности r.

Обозначим A_S множество векторов-строк $\{\mathbf{a}_{i\cdot}\}$ с индексами $i \in S$. Множество $\{S(1), \ldots, S(k)\}$, k — число сегментов, $1 \leqslant k \leqslant m$, есть разбиение I такое, что

$$S(\xi) \subset I, \quad \bigcup_{\xi=1}^{k} S(\xi) = I, \quad \bigcap_{\xi=1}^{k} S(\xi) = \varnothing.$$

Требуется найти такое разбиение $\mathcal{S}$, что $\mathrm{rank}(A_{S(\xi)}) \leqslant r$ для всех элементов этого разбиения, $\xi = 1, \ldots, k$.

Пусть на первой итерации каждый вектор из A включен в отдельный сегмент размерности ноль,

$$k = m, \quad \xi = i = 1, \ldots, k-1.$$

Далее на каждой итерации выполняем следующую последовательность действий. Начиная с первого вектора присоедниняем к кластеру последующие векторы, при условии, что

$$\mathrm{rank}_{\lambda_r} A_{S(\xi) \cup \ldots \cup S(\xi+\iota)} \leqslant r, \iota = 1, \ldots, k-\xi, \quad \xi = 1, \ldots, k$$

При выполнении условия кластер на следующей итерации определяется индексами

$$S^*(\xi) = S(\xi) \cup \ldots \cup S(\xi + \iota),$$

в противном случае он остается без изменения,

$$S^*(\xi) = S(\xi).$$

Итерации повторяются до тех пор, пока удается присоединить хотя бы один последующий кластер.

Этот алгоритм является эвристическим. Он не доставляет единственного разбиения траектории на сегменты. Разбиение зависит от порядка присоединения кластеров. Если на парах векторов $(\mathbf{a}_{i\cdot}, \mathbf{a}_{j\cdot})$, определена метрика $\rho(\mathbf{a}_{i\cdot}, \mathbf{a}_{j\cdot})$, то алгоритм можно изменить следующим образом. На каждой итерации к кластеру присоединяется предшествующий или предыдущий, в зависимости от расстояния между соседними точками данной пары кластеров.

3. Метод группового учета аргументов

Метод группового учета аргументов, МГУА (Group Method of Data Handling, GMDH)[1] — метод порождения и выбора регрессионных моделей оптимальной сложности. Под сложностью модели в МГУА понимается число параметров. Для порождения используется , подмножество элементов которой должно входить в искомую модель. Для

[1] Альтернативные названия метода: Group Method for Data Handling, Polynomial Neural Networks, Abductive and Statistical Learning Networks.

выбора моделей используются внешние критерии, специальные функционалы качества моделей, вычисленные на тестовой выборке.

МГУА рекомендуется к использованию и в том случае, когда выборка содержит всего несколько элементов. Тогда при построении регрессионных моделей использовать статистические гипотезы о плотности распределения, плотности распределения например, гипотезу о Гауссовском распределении, невозможно. Поэтому используется индуктивный подход, согласно которому последовательно порождаются модели возрастающей сложности до тех пор, пока не будет найден минимум некоторого критерия качества модели. Этот критерий качества называется *внешний критерий*, так как при настройке моделей и при оценке качества моделей используются разные данные. Достижение глобального минимума внешнего критерия при порождении моделей означает, что модель, доставляющая такой минимум, является искомой.

Один из авторов этого метода А. Г. Ивахненко пишет: «Осуществляется целенаправленный перебор многих моделей-претендентов различной сложности по ряду критериев. В результате находится модель оптимальной структуры в виде одного уравнения или системы уравнений. Минимум критерия селекции определяет модель оптимальной структуры».

3.1. Описание алгоритма МГУА

Индуктивный алгоритм отыскания модели оптимальной структуры в состоит из следующих основных шагов.

1. Пусть задана выборка $D = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$, $\mathbf{x} \in \mathbb{R}^m$. Выборка разбивается на обучающую и тестовую. Обозначим ℓ, C — множества индексов из $\{1, \ldots, N\} = W$. Эти множества удовлетворяют условиям разбиения $\ell \cup C = W, \ell \cap C = \emptyset$. Матрица X_ℓ состоит из тех векторов-строк $\mathbf{x}_n$, для которых индекс $n \in \ell$. Вектор $\mathbf{y}_\ell$ состоит из тех элементов y_n, для которых индекс $n \in \ell$. Разбиение выборки представляется в виде

$$X_W = \begin{pmatrix} X_\ell \\ \hline X_C \end{pmatrix}, \mathbf{y}_W = \begin{pmatrix} \mathbf{y}_\ell \\ \hline \mathbf{y}_C \end{pmatrix}, \mathbf{y}_W \in \mathbb{R}^{N\times 1}, X_W \in \mathbb{R}^{N\times m}, |\ell|+|C| = N.$$

2. Назначается базовая модель. Эта модель описывает отношение между зависимой переменной y и свободными переменными $\mathbf{x}$. На-

пример, используется функциональный ряд Вольтерра, называемый также полиномом Колмогорова-Габора:

$$y = w_0 + \sum_{i=1}^{m} w_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} w_{ijk} x_i x_j x_k + \dots .$$

В этой модели $\mathbf{x} = \{x_i | i = 1, \dots, m\}$ — множество свободных переменных и $\mathbf{w}$ — вектор параметров — весовых коэффициентов

$$\mathbf{w} = \langle w_i, w_{ij}, w_{ijk}, \dots | i, j, k, \dots = 1, \dots, m \rangle.$$

В некоторых случаях имеет смысл увеличить число элементов вектора свободной переменной $\mathbf{x}$ за счет добавления нелинейных преобразований отдельных переменных. Например, задано конечное множество нелинейных функций $G = \{g | \mathbb{R} \longrightarrow \mathbb{R}\}$. Дополнительная свободная переменная получается путем применения некоторого преобразования из G к одной или к нескольким переменным из множества $\{x\}$. Базовая модель линейна относительно параметров w и нелинейна относительно свободных переменных x.

3. Исходя из поставленных задач выбирается целевая функция — внешний критерий, описывающий качество модели. Ниже описаны несколько часто используемых внешних критериев.

4. Индуктивно порождаются модели-претенденты. При этом вводится ограничение на длину полинома базовой модели. Например, степень полинома базовой модели на не должно превышать заданное число R. Тогда базовая модель представима в виде линейной комбинации заданного числа F_0 произведений свободных переменных:

$$y = f(x_1, x_2, \dots, x_1^2, x_1 x_2, x_2^2, \dots, x_m^R),$$

здесь f — линейная комбинация. Аргументы этой функции переобозначаются следующим образом:

$$x_1 \mapsto a_1, x_2 \mapsto a_2, \dots, x_1^2 \mapsto a_\alpha, x_1 x_2 \mapsto a_\beta, x_2^2 \mapsto a_\gamma, \dots, x_m^q \mapsto a_{F_0},$$

то есть,

$$y = f(a_1, a_2, \dots, a_{F_0}).$$

Для линейно входящих коэффициентов задается одноиндексная нумерация $\mathbf{w} = \langle w_1, \dots, w_{F_0} \rangle$. Тогда модель может быть представлена в

виде линейной комбинации

$$y = w_0 + \sum_{i=1}^{F_0} w_i a_i = w_0 + \mathbf{w} \cdot \mathbf{a}.$$

Каждая порождаемая модель задается линейной комбинацией элементов $\{(w_i, a_i)\}$, в которой множество индексов $\{i\} = s$ является подмножеством $\{1, \ldots, F_0\}$.

5. Настраиваются параметры моделей. Для настройки используется *внутренний критерий* — критерий, вычисляемый с использованием обучающей выборки. Каждому элементу вектора $\mathbf{x}_n$ — элемента выборки D ставится в соответствие вектор $\mathbf{a}_n$, алгоритм построения соответствия указан выше. Строится матрица A_W — набор векторов-столбцов $\mathbf{a}_i$. Матрица A_W разбивается на подматрицы A_ℓ и A_C. Наименьшую невязку $\|\mathbf{y} - \hat{\mathbf{y}}\|$, где $\hat{\mathbf{y}} = A\hat{\mathbf{w}}$ доставляет значение вектора параметров $\hat{\mathbf{w}}$, который вычисляется методом наименьших квадратов:

$$\hat{\mathbf{w}}_G = (A_G^T A_G)^{-1} A_G^T \mathbf{y}_G, \text{ где } G \in \{\ell, C, W\}.$$

При этом в качестве внутреннего критерия выступает среднеквадратичная ошибка

$$\varepsilon_G^2 = \|\mathbf{y}_G - A_G \hat{\mathbf{w}}_G\|^2.$$

В соответствии с критерием $\varepsilon_G^2 \longrightarrow \min$ происходит настройка параметров $\mathbf{w}$ и вычисление ошибки на тестовой подвыборке, обозначенной G, здесь $G = \ell$. При усложнении модели внутренний критерий не дает минимума для моделей оптимальной сложности, поэтому для выбора модели он не пригоден.

6. Для выбора моделей вычисляется качество порожденных моделей. При этом используется контрольная выборка и назначенный внешний критерий. Ошибка на подвыборке H обозначается

$$\Delta^2(H) = \Delta^2(H \backslash G) = \|\mathbf{y}_H - A_H \hat{\mathbf{w}}_G\|^2,$$

где $H \in \{\ell, C\}$, $H \cap G = \emptyset$. Это означает что ошибка вычисляется на подвыборке H при параметрах модели, полученных на подвыборке G.

7. Модель, доставляющая минимум внешнему критерию, считается оптимальной.

Если значение внешнего критерия не достигает своего минимума при увеличении сложности модели или значение функции качества

неудовлетворительно, то выбирается лучшая модель из моделей заданной сложности. Под сложностью модели подразумевается число настраиваемых параметров модели. Существуют следующие причины, по которым глобальный минимум может не существовать:

- данные слишком зашумлены,
- среди данных нет необходимых для отыскания модели переменных,
- неверно задан критерий выбора,
- при анализе временных рядов существует значительная временная задержка отыскиваемой причинно-следственной связи.

3.2. Внешние критерии

Авторами метода рассмотрены весьма большое число различных критериев выбора моделей. Значительная часть этих критериев опубликована на сайте http://www.gmdh.net.

Критерий выбора модели может быть назван внешним, если он получен с помощью дополнительной информации, не содержащейся в данных которые использовались при вычислении параметров моделей. Например, такая информация содержится в дополнительной, тестовой выборке.

Алгоритм МГУА использует и внутренний критерий и внешний. Внутренний критерий используется для настройки параметров модели, внешний критерий используется для выбора модели оптимальной структуры. Возможен выбор моделей по нескольким внешним критериям.

3.2.а. Критерий регулярности

Критерий регулярности $\Delta^2(C)$ включает среднеквадратичную ошибку на обучающей подвыборке C полученную при параметрах модели, настроенных на тестовой подвыборке ℓ.

$$\Delta^2(C) = \|\mathbf{y}_C - A_C\hat{\mathbf{w}}_\ell\|^2 = (\mathbf{y}_C - A_C\hat{\mathbf{w}}_\ell)^T(\mathbf{y}_C - A_C\hat{\mathbf{w}}_\ell),$$

где

$$\hat{\mathbf{w}}_\ell = (A_\ell^T A_\ell)^{-1}(A_\ell^T \mathbf{y}_\ell)$$

и

$$\hat{\mathbf{y}}_C(\ell) = A_C \hat{\mathbf{w}}_\ell.$$

Другие модификации критерия регулярности:

$$\Delta^2(C) = \frac{\|\mathbf{y}_C - A_C \hat{\mathbf{w}}_\ell\|^2}{\|\mathbf{y}_C\|^2}$$

и

$$\Delta^2(C) = \frac{\|\mathbf{y}_C - A_C \hat{\mathbf{w}}_\ell\|^2}{\|\mathbf{y}_C - \bar{\mathbf{y}}_C\|^2},$$

где $\bar{\mathbf{y}}$ — среднее значение вектора $\mathbf{y}$.

Критерий $\Delta^2(C)$ также обозначается $\Delta^2(C\backslash\ell)$, то есть ошибка на подвыборке C, при параметрах, полученных на подвыборке ℓ.

3.2.b. Критерий минимального смещения

Иначе критерий непротиворечивости модели: модель которая имеет на обучающей выборке одну невязку, а на контрольной — другую, называется противоречивой. Этот критерий включает разность между зависимыми переменными модели, вычисленными на двух различных выборках ℓ и C. Критерий не включает ошибку модели в явной форме. Он требует, чтобы оценки коэффициентов в оптимальной модели, вычисленные на множествах ℓ и C, различались минимально.

Критерий непротиворечивости как критерий минимума смещения имеет вид

$$\eta^2_{\text{bs}} = \|A_W \hat{\mathbf{w}}_\ell - A_W \hat{\mathbf{w}}_C\|^2 = (\hat{\mathbf{w}}_\ell - \hat{\mathbf{w}}_C)^T A_W^T A_W (\hat{\mathbf{w}}_\ell - \hat{\mathbf{w}}_C).$$

Другие модификации этого критерия:

$$\eta^2_{\text{bs}} = \frac{\|A_W \hat{\mathbf{w}}_\ell - A_W \hat{\mathbf{w}}_C\|^2}{\|\mathbf{y}_C - \bar{\mathbf{y}}_C\|^2}$$

и

$$\eta^2_a = \|\hat{\mathbf{w}}_\ell - \hat{\mathbf{w}}_C\|^2,$$

где $\hat{\mathbf{w}}_\ell$ и $\hat{\mathbf{w}}_C$ — векторы коэффициентов, полученные с использованием подвыборок ℓ и C. При использовании последнего варианта следует помнить, что число элементов вектора параметров $\mathbf{w}$ в различных моделях может быть различно.

3.2.c. Критерий “absolute noise-immune”

Утверждается, что с помощью этого критерия, из сильно зашумленных данных возможно найти скрытые физические закономерности.

$$V^2 = (A_W\hat{\mathbf{w}}_\ell - A_W\hat{\mathbf{w}}_W)^T(A_W\hat{\mathbf{w}}_W - A_W\hat{\mathbf{w}}_C) =$$
$$= (\hat{\mathbf{w}}_\ell - \hat{\mathbf{w}}_W)^T A_W^T A_W(\hat{\mathbf{w}}_W - \hat{\mathbf{w}}_C).$$

где $\hat{\mathbf{w}}_W$ — вектор коэффициентов, полученный на всей выборке W.

3.2.d. Критерий предсказательной способности

Является модификацией критерия регулярности. Этот критерий включает среднеквадратичную ошибку для отдельной экзаменационной выборки B, которая не была использована ни при нахождении коэффициентов, ни при выборе моделей. В этом случае выборка делится не на две, а на три части:

$$X_W = \left(\frac{X_\ell}{\frac{X_C}{X_B}}\right), \mathbf{y}_W = \left(\frac{\mathbf{y}_\ell}{\frac{\mathbf{y}_C}{\mathbf{y}_B}}\right).$$

Критерий предсказательной способности имеет вид

$$\Delta^2(W\backslash B) = \frac{\|\mathbf{y}_W - A_W\hat{\mathbf{w}}_B\|^2}{\|\mathbf{y}_W - \bar{\mathbf{y}}_W\|^2}.$$

3.2.e. Комбинированный критерий

Этот критерий позволяет использовать при выборе моделей линейную комбинацию нескольких критериев. Комбинированный критерий

$$k^2 = \sum_{i=1}^{K} \alpha_i k_i^2, \text{при условии нормировки } \sum_{i=1}^{K} \alpha_i = 1.$$

Здесь k_i — принятые на рассмотрение критерии, а α_i — веса этих критериев, назначенные в начале вычислительного эксперимента.

Используются также нормализованные значения критериев. При этом предыдущая формула имеет вид

$$k^2 = \sum_{i=1}^{K} \alpha_i \frac{k_i^2}{k_{i\max}^2}.$$

Максимальное значение критерия $k_{i\,\max}^2$ берется по вычисленным значениям критериев для всех порожденных моделей. В данном случае оптимальная модель может быть найдена только после завершения настройки параметров всех моделей.

Пример распространенного комбинированного критерия — смещение плюс ошибка аппроксимации.

$$c_1^2 = \bar{\eta}_{\text{bs}}^2 + \bar{\varepsilon}^2(W) = \frac{\eta_{\text{bs}}^2}{\eta_{\text{bs max}}^2} + \frac{\varepsilon^2}{\varepsilon_{\text{max}}^2},$$

где $\bar{\varepsilon}^2(W)$ — нормализованная среднеквадратичная ошибка аппроксимации на всей выборке $W = \ell \cup C$ с использованием коэффициентов, полученных также на W.

Второй пример комбинированного критерия — смещение плюс регулярность.

$$c_2^2 = \bar{\eta}_{\text{bs}}^2 + \bar{\Delta}^2(C).$$

Третий пример — смещение плюс ошибка на тестовой выборке.

$$c_3^2 = \bar{\eta}_{\text{bs}}^2 + \bar{\Delta}^2(B \backslash W).$$

Такой критерий обеспечивает выбор наиболее несмещенных, устойчивых и точных моделей. Здесь $\Delta(C \backslash W)$ — среднеквадратичная ошибка, вычисленная на выборке C, с весами, настроенными на всей выборке W.

Обычно при вычислении критерия c_3 выборку делят на три части в пропорциях $\ell = 40\%$, $C = 40\%$ и $B = 20\%$. Выборки ℓ и C используются для вычисления критерия минимального смещения, а выборка B — для вычисления ошибки предсказания. Для критериев c_1 и c_2 выборка обычно делится на две равные части.

3.2.f. Парето-оптимальный фронт в пространстве критериев

Парето-оптимальный фронт — альтернатива комбинированным критериям. Выбирается множество внешних критериев, условиям оптимальности которых должна удовлетворять модель. Каждой модели ставится в соответствие вектор в пространстве выбранных критериев. Отыскиваются векторы, принадлежащие Парето-оптимальному

фронту множества всех векторов, соответствующих порожденным моделям. При создании комбинированного критерия рассматриваются модели, критерии которых принадлежат полученному Парето-оптимальному фронту.

3.3. Алгоритм порождения моделей МГУА

Целю МГУА является получение модели в результате перебора моделей из индуктивно-порождаемого множества. Параметры каждой модели настраиваются так, чтобы доставить минимум выбранному внешнему критерию. Различают два основных типа алгоритмов МГУА — однорядный и многорядный.

Все алгоритмы МГУА воспроизводят схему массовой селекции: последовательно порождаются модели возрастающей сложности. Каждая модель настраивается — методом наименьших квадратов находятся значения параметров. Из моделей-претендентов выбираются лучшие в соответствии с выбранным критерием. Многорядные алгоритмы могут вычислять остатки регрессионных моделей после каждого ряда селекции или не вычислять; при этом используются исходные данные.

Каждая полиномиальная модель однозначно определяется набором индексов s входящих в нее мономов

$$y = a_0 + \mathbf{w}(s) \cdot \mathbf{a}(s).$$

Элементы вектора $\mathbf{w}$ — коэффициенты при мономе полинома Колмогорова-Габора; элементы вектора $\mathbf{a}$ — результат произведения свободных переменных соответствующих мономов. Индексы $s \subseteq \{1, \ldots, F_0\}$ есть индексы мономов, входящих в модель. Иначе, произвольная модель

$$y = w_0 + \mathbf{w}(s) \cdot \mathbf{a}(s)$$

порождается набором индексов $s \subseteq \{1, \ldots, F_0\}$, включающих соотвествующие элементы векторов

$$\mathbf{w} = \langle w_1, \ldots, w_{F_0} \rangle \text{ и } \mathbf{a} = \langle a_1, \ldots, a_{F_0} \rangle.$$

При ограничении степени полинома числом R, число мономов по-

линома равно

$$F_0 = \sum_{r=1}^{R} \bar{C}_r^P = \sum_{r=1}^{R} \frac{(r+P-1)!}{P!(r-1)!},$$

а число моделей первого ряда соответственно равно 2^{F_0}. Здесь $\bar{C}_r^P$ — число сочетаний с повторениями из P по r, P — число свободных переменных — элементов вектора $\mathbf{x}$.

3.3.а. Комбинаторный алгоритм

Комбинаторный (однорядный) алгоритм использует только один ряд выбора. При этом порождаются все возможные линейные комбинации ограниченной сложности. Так как под сложностью понимается число линейно входящих параметров w, то сложность не превосходит заданное значение F_0. Пусть, как и ранее

$$y = w_0 + w_1 a_1 + w_2 a_2 + w_3 a_3, \ldots, w_{F_0} a_{F_0}.$$

Алгоритм выполняет следующие шаги. Для всех комбинаций входных аргументов строятся модели-претенденты неубывающей сложности. Например,

$$\begin{aligned}
y_1 &= w_{10} + w_{11} a_1, \\
y_2 &= w_{20} + w_{22} a_2, \\
y_3 &= w_{30} + w_{31} a_1 + w_{32} a_2, \\
y_4 &= w_{40} + w_{43} a_3, \\
y_5 &= w_{50} + w_{51} a_1 + w_{53} a_3, \\
y_6 &= w_{60} + w_{62} a_2 + w_{63} a_3, \\
y_7 &= w_{70} + w_{71} a_1 + w_{72} a_2 + w_{73} a_3.
\end{aligned}$$

Параметры каждой модели настраиваются методом наименьших квадратов по обучающей выборке. Наилучшая модель выбирается исходя из минимума значения внешнего критерия. Как вариант — назначается порог и выбираются несколько моделей, значения критерия для которых не превышает этот порог.

При программировании данного алгоритма удобно ввести переменную выбора — вектор $\mathbf{c}(s) = \langle c_1, \ldots, c_{F_0} \rangle$. Его элемент $c_i \in \{0, 1\}$ принимает значение 1, если $i \in s$, в противном случае 0. Тогда модель имеет вид

$$y = \mathbf{c}(s) \cdot \mathbf{w} \cdot \mathbf{a}.$$

Последовательность векторов **c** для предыдущего примера выглядит как

$$\begin{matrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ 0 & \cdots & 1 & 1 \\ \vdots & \ddots & \vdots & \vdots \\ 1 & \cdots & 1 & 1 \end{matrix}.$$

Так как при порождении моделей необходимо выбирать из 2^{F_0} моделей, что может повлечь недопустимо большое время вычислений, предложены несколько эвристических алгоритмов, позволяющих сократить время вычислений, без уменьшения максимальной сложности моделей.

3.3.b. Многорядный алгоритм

На первом ряде алгоритма порождения моделей задано множество из F_0 переменных a_i. Порождаются модели как линейные комбинации всевозможных пар переменных

$$y_{(ij)} = w_0 + w_i a_i + w_j a_j, i, j = 1, \ldots, F_0, i \neq j.$$

Число моделей первого ряда M_1 есть число сочетаний $C_2^{F_0} = \frac{1}{2} F_0(F_0 - 1) = M_1$. Каждая модель, порождаемая на ряде задается парой индексов (i, j). После порождения моделей их параметры настраиваются с использованием внутреннего критерия. Затем выбираются F_1 наилучших моделей с использованием внешнего критерия. Эти модели используются в следующем ряде. Множество выбранных моделей задано множеством пар индексов $\{(i, j)\}$. На первом ряде индексы выбранных моделей i, j принадлежат множеству $s_1 \subset \{1, \ldots, F_0\}$.

На каждом последующем ряде новые модели — порождаются как суммы всевозможных пар выбранных моделей предыдущего ряда. Например, для второго ряда множество моделей

$$z_{(pquv)} = y_{(pq)} + y_{(uv)} = w_0 + w_p a_p + w_q a_q + w_u a_u + w_v a_v,$$

$$p, q, u, v = 1, \ldots, s_1,\ p \neq q \neq\ u \neq\ v.$$

Порожденные модели снова настраиваются, выбираются наилучшие. Индексы, задающие выбранные модели принадлежат множеству $\{(p, q, u, v)\}$, где $p, q, u, v \in s_2 \subset s_1 \subset \{1, \ldots, F_0\}$.

Таким образом на l-м ряде с помощью вышеприведенного алгоритма выбирается множество моделей, задаваемых множеством наборов индексов $\{p, \ldots, v\}$, которые принадлежат полученному множеству $s_l \subset s_{l-1} \subset s_1 \subset \{1, \ldots, F_0\}$. При этом индексация элементов моделей остается сквозной,

$$z_{(p\ldots v)} = y_{(p\ldots q)} + y_{(u\ldots v)} = w_0 + w_p a_p + \ldots + w_v a_v.$$

Остановка порождения моделей на последующих рядах происходит в том случае, когда с увеличением номера слоя, то есть, с усложнением моделей, происходит увеличение внешнего критерия лучшей модели.

4. Нелинейные методы

Нелинейные регрессионные модели — модели вида

$$y = f(\mathbf{w}, \mathbf{x}) + \varepsilon,$$

которые не могут быть представлены в виде скалярного произведения

$$f(\mathbf{w}, \mathbf{x}) = (\mathbf{w}, \mathbf{g}(\mathbf{x})) = \sum_{i=1}^{W} w_i g_i(\mathbf{x}).$$

Здесь $\mathbf{w} = [w_1, \ldots, w_W]$ — параметры регрессионной модели, $\mathbf{x}$ свободная переменная из пространства $\mathbb{R}^N$, y — зависимая переменная, ε — случайная величина и $\mathbf{g} = [g_1, \ldots, g_W]$ — функция из некоторого заданного множества.

4.1. Часто используемые регрессионные модели

Ниже приведены модели, которые используются при регрессионном анализе измеряемых данных. Параметры моделей обозначены латинскими и греческими буквами: $\{a, b, c, ..., \chi, \psi, \omega\}$, x, y — свободная и зависимая переменные. Все параметры и переменные принадлежат действительным числам. При соединении параметров в вектор $\mathbf{w}$, для представления модели в виде $y = f(\mathbf{w}, \mathbf{x}) + \varepsilon$, параметры присоединяются в лексикографическом порядке, то есть в том порядке, в котором

они появляются, если представить формулу регрессионной модели в виде строки.

В список не вошли универсальные параметрические модели, например, нейронная сеть — многослойный персептрон, функции радиального базиса, полиномы Лагранжа, полиномы ЧебышёваТакже не вошли непараметрические модели. Оба эти класса требуют специального описания.

4.1.a. Нелинейные модели

1. Экспонента, $y = e^b x$, с линейным коэффициентом, $y = ae^b x$. Распространена двухкомпонентная экспоненциальная модель, $y = ae^b x + ce^d x$. Модель может быть использована, в частности, если коэффициент изменения величины свободной переменной пропорционален ее начальной величине.

2. Ряд Фурье, $y = a_0 + \sum_{i=1}^{n} \big(a_i \cos(i\omega x) + b_i \sin(i\omega x)\big)$. Используется для описания периодических сигналов.

3. Сумма гауссианов, $y = \sum_{i=1}^{n} a_i \exp(-\frac{(x-b_i)^2}{c_i})$. Используется для аппроксимации пиков. Коэффициент a_i является амплитудой, b_i — смещение, коэффициент c_i отражает ширину пика. Всего в сумме может быть до n пиков.

4. Моном, $y = x^b$, с линейным коэффициентом, $y = ax^b$. Используется при моделировании размерности физических или химических величин. Например, количество некоторого реагирующего в химической реакции вещества как правило, пропорциональна концентрации этого вещества, возведенного в некоторую степень.

5. Рациональный полином, $y = \frac{\sum_{i=0}^{n} a_i x^i}{x^m + \sum_{i=0}^{m-1} b_i x^i}$. Принято считать коэффициент перед x^m единицей. Например, если $m = n$, такое соглашение позволит получить уникальные числитель и знаменатель.

6. Сумма синусов, $y = \sum_{i=1}^{n} a_i \sin(b_i x + c_i)$. Здесь a_i — амплитуда, b_i — частота, c_i — фаза некоторого периодического процесса.

7. Распределение Вейбулла, двухпараметрическое, $y = abx^{b-1}\exp(-ax^b)$. Параметр a является масштабирующим, а параметр b определяет форму кривой. Трехпараметрическое распределение Вейбулла, со смещением c, $y = abx^{b-1}\exp(-a(x-c)^b)$.

8. Логарифмическая сигмоида, $\frac{1}{1+\exp(-n)}$, используются в нейронных сетях, например в MLP, в качестве функций активации.

9. Тангенциальная сигмоида, $y = \frac{2}{1+\exp(-2n)} - 1$, также используются в качестве функций активации.

4.1.b. Линейные модели

1. Полином, $y = \sum_{i=1}^{n} a_i x^{i-1}$ и его частный случай прямая $y = ax+b$. Следует помнить, что полиномы высоких степеней крайне неустойчивы и могут неадекватно описывать измеряемые данные.

2. Гипербола, $y = k/x$, а также прочие нелинейные функции с линейно-входящими параметрами: тригонометрические функции $\sin(x), \arcsin(x)$, гиперболический синус $\mathrm{sh}(x)$, корневые $\sqrt{x}$ и обратно-корневые функции. Эти функции используются в финансовом анализе и других приложениях.

Этот список не является жестко заданным. Выбираемая регрессионная модель зависит прежде всего от экспертных предположений относительно моделируемого явления.

4.2. Символьная регрессия

Символьная регрессия — метод построения нелинейных регрессионных моделей путем перебора различных произвольных суперпозиций функций из некоторого заданного набора. Суперпозиция функций при этом называется «программой», а выполнение стохастического оптимизационный алгоритма, который строит такие суперпозиции называется генетическим программированием.

Генетическое программирование — модификация генетического алгоритма. Различие заключается в том, что для решения задач символьной регрессии необходима изменяющаяся длина хромосом, опи-

сывающих суперпозиции. Так как подобные алгоритмы являются переборными и требуют значительных вычислительных ресурсов, то публикации по данной теме стали появляться в 90-х годах, а значительное развитие они получили после 2000-го года. Наиболее известным исследователем является Джон Коза.

Задача отыскания оптимальной структуры регрессионной модели нескольких свободных переменных следующим образом. Задана выборка — множество $\{\mathbf{x}_1, \ldots, \mathbf{x}_N | \mathbf{x} \in \mathbb{R}^M\}$ значений свободных переменных и множество $\{y_1, \ldots, y_N | y \in \mathbb{R}\}$ соответствующих им значений зависимой переменной. Обозначим оба эти множества как множество исходных данных D. Также задано множество $G = \{g | g : \mathbb{R} \times \ldots \times \mathbb{R} \longrightarrow \mathbb{R}\}$ гладких параметрических функций $g = g(\mathbf{b}, \cdot, \cdot, \ldots, \cdot)$. Первый аргумент функции g — вектор-строка параметров $\mathbf{b}$, последующие — переменные из множества действительных чисел, рассматриваемые как элементы вектора свободных переменных. Рассмотрим произвольную суперпозицию, состоящую из не более чем r функций g. Эта суперпозиция задает параметрическую регрессионную модель $f = f(\mathbf{w}, \mathbf{x})$. Регрессионная модель f зависит от вектора свободных переменных $\mathbf{x}$ и от вектора параметров $\mathbf{w}$. Вектор $\mathbf{w} \in \mathbb{R}^W$ состоит из присоединенных векторов-параметров функций $g_1, \ldots, g_r$, то есть, $\mathbf{w} = \mathbf{b}_1 \vdots \mathbf{b}_2 \vdots \ldots \vdots \mathbf{b}_r$, где $\vdots$ — знак присоединения векторов. Обозначим $\Phi = \{f_i\}$ — множество всех суперпозиций, индуктивно порожденное элементами множества G.

Требуется выбрать такую модель f_i, которая доставляет максимум заданного функционала $p(\mathbf{w}|D)$. Этот функционал определяет целевую функцию $S(\mathbf{w})$, которая используется при вычислениях.

4.3. Алгоритм Левенберга-Марквардта

Алгоритм Левенберга-Марквардта предназначен для оптимизации параметров нелинейных регрессионных моделей. Предполагается, что в качестве критерия оптимизации используется модели на обучающей выборке. Алгоритм заключается в последовательном приближении заданных начальных значений параметров к искомому локальному оптимуму.

Алгоритм отличается от метода сопряженных градиентов тем, что использует матрицу Якоби модели, а не indexградиент вектора пара-

метров. От алгоритма Гаусса-Ньютона этот алгоритм отличается тем, что использует параметр регуляризации.

Задана регрессионная выборка — множество пар $D = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ свободной переменной $\mathbf{x} \in \mathbb{R}^M$ и зависимой переменной $y \in \mathbb{R}$. Задана регрессионная модель — функция $f(\mathbf{w}, \mathbf{x}_n)$ непрерывно дифференцируемая в области $W \times X$.

Требуется найти такое значение вектора параметров $\mathbf{w}$, которое бы доставляло локальный минимум функции ошибки

$$E_D = \sum_{n=1}^{N} \big(y_n - f(\mathbf{w}, \mathbf{x}_n)\big)^2 . (*)$$

Перед началом работы алгоритма задается начальный вектор параметров $\mathbf{w}$. На каждом шаге итерации этот вектор заменяется на вектор $\mathbf{w} + \Delta\mathbf{w}$. Для оценки приращения $\Delta\mathbf{w}$ используется линейное приближение функции

$$f(\mathbf{w} + \Delta\mathbf{w}, \mathbf{x}) \approx f(\mathbf{w}, \mathbf{x}) + J\Delta\mathbf{w},$$

где J — якобиан функции $f(\mathbf{w}, \mathbf{x}_n)$ в точке $\mathbf{w}$. $(N \times R)$-матрицу J наглядно можно представить в виде

$$J = \begin{bmatrix} \frac{\partial f(\mathbf{w},\mathbf{x}_1)}{\partial w_1} & \cdots & \frac{\partial f(\mathbf{w},\mathbf{x}_1)}{\partial w_R} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{w},\mathbf{x}_N)}{\partial w_1} & \cdots & \frac{\partial f(\mathbf{w},\mathbf{x}_N)}{\partial w_R} \end{bmatrix}.$$

Здесь вектор параметров $\mathbf{w} = [w_1, \ldots, w_R]^T$.

Приращение $\Delta\mathbf{w}$ в точке $\mathbf{w}$, доставляющий минимум E_D равно нулю. Поэтому для нахождения последующего значения приращения $\Delta\mathbf{w}$ приравняем нулю вектор частных производных E_D по $\mathbf{w}$. Для этого $(*)$ представим в виде

$$E_D = \|\mathbf{y} - \mathbf{f}(\mathbf{w} + \Delta\mathbf{w})\|^2,$$

где $\mathbf{y} = [y_1, \ldots, y_N]^T$ и $\mathbf{f}(\mathbf{w} + \Delta\mathbf{w}) = [f(\mathbf{w} + \Delta\mathbf{w}, \mathbf{x}_1), \ldots, f(\mathbf{w} + \Delta\mathbf{w}, \mathbf{x}_N))]^T$. Преобразовывая это выражение

$$\|\mathbf{y}-\mathbf{f}(\mathbf{w}+\Delta\mathbf{w})\|^2 = \big(\mathbf{y}-\mathbf{f}(\mathbf{w}+\Delta\mathbf{w})\big)^T\big(\mathbf{y}-\mathbf{f}(\mathbf{w}+\Delta\mathbf{w})\big) = \mathbf{f}^T(\mathbf{w}+\Delta\mathbf{w})\mathbf{f}(\mathbf{w})-2\mathbf{y}^T\mathbf{f}(\mathbf{w}+\Delta\mathbf{w})+\mathbf{y}^T\mathbf{y}$$

и дифференцируя (ср. дифференцирование вектора невязки в разделе «Метод наименьших квадратов»), получим

$$\frac{\partial E_D}{\partial \mathbf{w}} = (J^T J)\Delta\mathbf{w} - J^T\big(\mathbf{y} - \mathbf{f}(\mathbf{w})\big) = 0.$$

Таким образом, чтобы найти значение $\Delta\mathbf{w}$ нужно решить систему линейных уравнений

$$\Delta\mathbf{w} = (J^T J)^{-1} J^T\big(\mathbf{y} - \mathbf{f}(\mathbf{w})\big).$$

Так как число обусловленности матрицы $J^T J$ есть квадрат числа обусловленности матрицы J (см. соотв. раздел в разделе «Сингулярное разложение»), то матрица $J^T J$ может оказаться существенно вырожденной. Поэтому Марквардт предложил ввести параметр регуляризации $\lambda \geq 0$,

$$\Delta\mathbf{w} = (J^T J + \lambda I)^{-1} J^T\big(\mathbf{y} - \mathbf{f}(\mathbf{w})\big),$$

где I — единичная матрица. Этот параметр назначается на каждой итерации алгоритма. Если значение ошибки E_D убывает быстро, малое значение λ сводит этот алгоритм к алгоритму Гаусса-Ньютона.

Алгоритм останавливается, в том случае, если приращение $\Delta\mathbf{w}$ в последующей итерации меньше заданного значения, либо если параметры $\mathbf{w}$ доставляют ошибку E_D меньшую заданной величины. Значение вектора $\mathbf{w}$ на последней итерации считается искомым.

Недостаток алгоритма — значительное увеличение параметра λ при плохой скорости аппроксимации. При этом обращение матрицы $J^T J + \lambda I$ становиться бессмысленным. Этот недостаток можно устранить, используя диагональ матрицы Гессе $J^T J$ в качестве регуляризующего слагаемого: $\Delta\mathbf{w} = \big(J^T J + \lambda \mathrm{diag}(J^T J)\big)^{-1} J^T\big(\mathbf{y} - \mathbf{f}(\mathbf{w})\big)$.

5. Сравнение и выбор моделей

Связанный Байесовский вывод — метод сравнения регрессионных моделей основанный на анализе их пространства параметров. Этот метод использует классический дважды: для вычисления апостериорной вероятности параметров модели и для вычисления апостериорной вероятности самой модели. Связанность заключается в том, что

оба вывода используют общий сомножитель, называемый *достоверностью модели*. Неотъемлемой частью этого метода является анализ пространства параметров модели и зависимости целевой функции от значений параметров. Результатом такого анализа является возможность оценить насколько важны отдельные параметры модели для аппроксимации данных. Связанный Байесовский вывод используется как в задачах регрессии, так и в задачах классификации.

5.1. Сравнение моделей

При сравнении моделей используется правило бритвы Оккама в следующей формулировке: «Совместный Байесовский вывод автоматически количественно выполняет правило Оккама». Бритва Оккама[2] — принцип предпочтения простых моделей (теорий, гипотез) сложным. Если несколько моделей одинаково хорошо описывают наблюдения, принцип Оккама рекомнедует выбор простейшей модели.

Теорема Байеса говорит о том, что наиболее вероятными будут те модели, которое наиболее точно предсказывают появление некоторых данных [3]. Эта вероятность определяется нормализованной функцией

[2]Уильям Оккам — английский философ-схоласт, логик и церковно-политический писатель (ок. 1285–1349), автор принципа: «не умножай сущности без необходимости». Этот принцип поддерживается двумя соображениями. Во-первых, эстетическим: «При описании результатов экспериментов у теории с красивой математикой больше шансов на успех, чем у безобразной» — Поль Дирак. Во-вторых, применение бритвы Оккама уже имело большой успех при решении практических задач.

[3]Формула Байеса, кратко. Условная вероятность $P(D|H)$ есть вероятность события D при условии наступления события H. Из всех элементов множества Ω элементарных событий, принадлежащих H входят в D лишь те события, которые принадлежат пересечению H и D. Эти элементы определяют $P(D|H)$. Но, если бы H было нормировано, то $P(D|H)$ равнялось бы $P(HD)$. Поэтому, чтобы условная вероятность отвечала условиям нормировки, используют нормирующий множитель:

$$P(D|H) = \frac{P(HD)}{P(H)}.$$

Согласно формуле умножения вероятностей, числитель этой дроби равен

$$P(HD) = P(D|H)P(H).$$

Разбиение множества Ω на полную группу несовместимых событий $H_1, \ldots, H_n$ позволяет любое событие D записать в виде

$$D = DH_1 + \cdots + DH_n,$$

распределения на пространстве данных D. Вероятность $P(D|H_i)$ появления данных D при фиксированной модели H_i называется правдоподобием модели H_i.

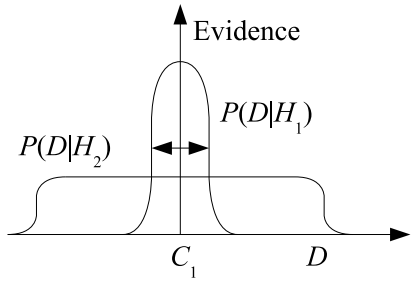

Рис. 6. Правдоподобие двух моделей. Рисунок дает интуитивное представление о том, почему более сложные модели являются менее правдоподобными. Ось абсцисс является предполагаемым пространством данных D.

Простая модель H_1 описывает ограниченное множество данных, что показано на рисунке функцией плотности распределения $P(D|H_1)$. Более сложная модель H_2, имеющая, например, большее количество параметров, описывает (иначе говоря, приближает с некоторой точностью, не хуже заданной) большее множество данных. Это, согласно нормированию функции плотности распределения,

---

откуда

$$P(D) = P(D|H_1)P(H_1) + \cdots + P(D|H_n)P(H_n) \quad (3)$$

Пусть $P(H), P(D) > 0$. Тогда из

$$P(HD) = P(H|D)P(D) = P(D|H)P(H)$$

вытекает

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)},$$

что после учета (3) при водит к формуле Байеса

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{\sum\limits_{j=1}^{n} P(D|H_j)P(H_j)}.$$

означает, что в некоторой области C_1 простая модель H_1 будет более вероятной при условии, что обе модели имеют одинаковую априорную вероятность.

Найдем правдоподобие двух альтернативных моделей H_1 и H_2, описывающих данные D. По теореме Байеса мы связываем правдоподобие $\mathrm{P}(H_1|D)$ модели H_1 при фиксированных данных, вероятность $P(D|H_1)$ получения данных с этой моделью и априорное правдоподобие $P(H_1)$ модели H_1. Так как значение нормирующего множителя $P(D) = \sum\limits_{j=1}^{n} P(D|H_j)P(H_j)$ для обоих моделей одинаково, то отношение правдоподобия моделей H_1 и H_2 имеет вид

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(H_1)P(D|H_1)}{P(H_2)P(D|H_2)}. \quad (1)$$

Отношение $\frac{P(H_1)}{P(H_2)}$ в правой части указывает на то, насколько велико априорное предпочтение модели $P(H_1)$ модели $P(H_2)$. Отношение $\frac{P(D|H_1)}{P(D|H_2)}$ указывает насколько модель H_1 соответствует наблюдаемым данным лучше, чем модель H_2.

Выражение (1) вводит правило Оккама следующим образом. Во-первых, возможно задать отношение $\frac{P(H_1)}{P(H_2)}$ так, чтобы оно отражало сложность моделей на основании некоторой дополнительной информации. Во-вторых, независимо от предыдущего способа задания критерия отбора моделей, это отношение автоматически выполняет правило Оккама. Действительно, если H_2 — более сложная модель, ее плотность распределения $P(D|H_2)$ имеет меньшие значения, при том условии, что ее дисперсия больше. Если невязки, доставляемые обеими моделями равны, простая модель H_1 будет более вероятна, чем сложная модель H_2. Таким образом, независимо от априорных предпочтений, вводится правило Оккама, согласно которому при равных априорных предпочтениях и равном соответствии предполагаемых моделей измеряемым данным, простая модель более вероятна, чем сложная.

5.2. Пример вычисления правдоподобия моделей

Рассмотрим последовательность $-1, 3, 7, 11$. Требуется предсказать следующие два числа и найти закономерность последовательности.

Первый вариант: $15, 19$. Закономерность H_a есть последующее число есть предыдущее плюс 4, иначе $x_{i+1} = x_i + 4$. Второй вариант: $-19.9, 1043.8$. Закономерность H_c есть $x_{i+1} = -x_i^(3)/11 + 9/11x_i^(2) + 23/11$.

С одной стороны, возможно непосредственно назначить априорные вероятности для обеих моделей так, чтобы штрафовать более сложную модель. С другой стороны, возможно вычислить их правдоподобие, $P(D|H_a)$ и $P(D|H_c)$ чтобы определить, насколько хорошо обе функции описывают данные.

Модель H_a зависит от двух параметров: добавляемого числа n и от первого числа в последовательности. Пусть каждый из этих параметров принадлежит множеству $\{-50, \ldots, 50\} \supset \mathbb{Z}$. Так как только пара значений $(n = 4, x_1 = -1)$ доставляют функцию, соответствующую данным $D = \{-1, 3, 7, 11\}$, то вероятность появления данных D при заданной модели H_a равна

$$P(D|H_a) = \frac{1}{101}\frac{1}{101} = 0.00010.$$

Для того, чтобы вычислить $P(D|H_c)$, требуется вычислить вероятность параметров c, d, e в кубическом многочлене H_c.

Эти параметры представлены с виде рациональных чисел (в противном случае обе эти модели были бы несравнимы). Пусть числители параметров, так же как и в предыдущем случае, принимают значения из множества $\{-50, \ldots, 50\}$ а знаменатели — из множества $\{1, \ldots, 50\}$. При вычислении вероятности принимается во внимание, что несколько способов представить дробь на заданных множествах. Например, $c = -1/11 = -2/22 = -3/33 = -4/44$. Вероятность $P(D|H_c)$ равна

$$P(D|H_c) = \left(\frac{1}{101}\right)\left(\frac{4}{101}\frac{1}{50}\right)\left(\frac{4}{101}\frac{1}{50}\right)\left(\frac{2}{101}\frac{1}{50}\right) = 0.0000000000025 =$$

$$= 2.5 \times 10^{-12}.$$

Отношение правдоподобия двух моделей (а значит и их апостериорных вероятностей при условии равенства априорных предпочтений), $P(D|H_a) = 0.00010$ и $P(D|H_a) = 2.5{\times}10^{-12}$ составляет одну сорокамиллионную.

5.3. Два уровня Байесовского вывода

При создании моделей различают два уровня Байесовского вывода. На *первом уровне* предполагается, что рассматриваемая модель адекватна. Производится настройка параметров модели по данным. В результате получаются наиболее правдоподобные значения параметров и значения ошибок моделей при этих параметрах. Эта процедура повторяется для каждой модели. Задача, решаемая на *втором уровне вывода* — сравнение моделей. Результатом является упорядоченное множество моделей.

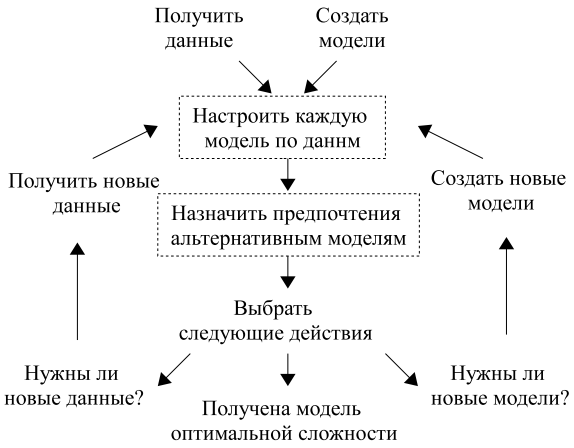

Рис. 7. Использование Байесовского вывода при создании моделей. Первый и второй уровень вывода обведены линией.

Каждая модель H_i имеет вектор параметров $\mathbf{w}$. Задача первого уровня — получить оценку параметров $\mathbf{w}$ модели при полученных данных D. Согласно теореме Байеса, апостериорная вероятность параметров $\mathbf{w}$ равна

$$P(\mathbf{w}|D, H_i) = \frac{P(D|\mathbf{w}, H_i)P(\mathbf{w}|H_i)}{P(D|H_i)}, \quad (2)$$

Нормирующая константа $P(D|H_i)$ обычно не принимается во внимание на первом уровне вывода. Однако она становится весьма важ-

ной на втором уровне вывода. Эта константа называется в англоязычной литературе «evidence» то есть «достоверность модели».

При отыскании параметров на практике обычно применяют оптимизационные методы типа метода сопряженных градиентов, чтобы получить наиболее вероятные параметры $\mathbf{w}_{MP}$. (Различают наиболее вероятные параметры $\mathbf{w}_{MP}$ выводятся на первом уровне как аргумент функции вероятности и наиболее правдоподобные параметры $\mathbf{w}_{ML}$, которые отыскиваются как аргумент функции наибольшего правдоподобия.)

Ошибка (иногда называемая прогностической способностью) модели H оценивается с помощью функции апостериорного распределения параметров модели. Для оценки используется приближение рядом Тейлора логарифма апостериорного распределения функции $P(\mathbf{w}|D, H_i)$

$$P(\mathbf{w}|D, H_i) \approx P(\mathbf{w}_{MP}|D, H_i) \exp(-\frac{1}{2}\Delta\mathbf{w}^T A \Delta\mathbf{w})),$$

где $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{MP}$, и отыскивается значение гессиана при значении весов максимального правдоподобия $\mathbf{w}_{\text{MP}}$ в окресности $\mathbf{w}_{MP}$:

$$A = -\nabla^(2) \ln P(\mathbf{w}|D, H_i)|_{\mathbf{w}_{MP}}.$$

Таким образом, функция апостериорного распределения параметров модели H_i может быть локально приближена с помощью матрицы A^{-1}, которая является матрицей ковариации в окресности значения ее параметров $\mathbf{w}_{MP}$.

На *втором уровне* байесовского вывода требуется определить, какая модель наиболее адекватно описывает данные. Апостериорная вероятность i-й модели задана как

$$P(H_i|D) \propto P(D|H_i)P(H_i). \quad (2)$$

Следует отметить, что сомножитель $P(D|H_i)$, включающий данные D, есть достоверность модели H_i, которая была названа ранее, в выражении (2), нормирующей константой. Достоверность модели может быть получено интегрированием функции правдоподобия по всему пространству параметров модели

$$P(D|H_i) = \int P(D|\mathbf{w}, H_i)P(\mathbf{w}|H_i)d\mathbf{w}.$$

Второй сомножитель $P(H_i)$ — априорная вероятность над пространством моделей, определяет, насколько адекватной (соответствующий английский термин — plausible) является модель до того, как появились данные. Основной проблемой Байесовского вывода является отсутствие объективных методов назначения априорной вероятности $P(H_i)$. Путь априорные вероятности $P(H_i)$ всех моделей равны. Тогда модели ранжируются по значениям достоверности $P(D|H_i)$.

Чрезвычайно важное предположение, которое необходимо сделать для решения задачи вычисления правдоподобия, — предположение о том, что распределение $P(\mathbf{w}|D, H_i) \propto P(D|\mathbf{w}, H_i)P(\mathbf{w}|H_i)$ имеет выраженный пик в окрестности наиболее вероятного значения $\mathbf{w}_{MP}$.

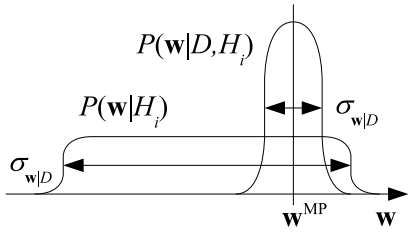

Рис. 8. Множитель Оккама — отношение $\sigma_{w|D}P(\mathbf{w}_{MP}|H_i) = \frac{\sigma_{w|D}}{\sigma_w}$

На рисунке показано, как вычисляется множитель Оккама для модели H_i с единственным параметром $\mathbf{w}$ на оси абсцисс. Сплошной линией показано априорное распределение параметра с дисперсией σ_w. Апостериорное распределение показано пунктирной линией имеет единственный максимум в точке $\mathbf{w}_{MP}$ и имеет дисперсию $\sigma_{w|D}$.

Функцию распределения вероятности параметров модели приближают гауссианой, определенной в пространстве параметров. Для этого используют метод Лапласа. Согласно этому методу, эта функция равна высоте пика подынтегрального выражения $P(D|\mathbf{w}, H_i)P(\mathbf{w}|H_i)$ умноженной на ширину пика, $\sigma_{w|D}$:

$$P(D|H_i) \approx P(D|\mathbf{w}_{MP}, H_i)P(\mathbf{w}_{MP}|H_i) \times \sigma_{w|D},$$

достоверность $\approx$ наибольшее правдоподобие $\times$ множитель Оккама.

Таким образом, достоверность модели находится с помощью оценок наибольшего правдоподобия параметров модели и множителя Окка-

ма, принимающего значения на отрезке $[0, 1]$, который штрафует модель H_i за ее параметры $\mathbf{w}$. Чем точнее была априорная оценка параметров, тем меньше штраф.

При аппроксимации Лапласа множитель Оккама может быть получен с помощью определителя ковариационной матрицы весов

$$P(D|H_i) \approx P(D|\mathbf{w}_{MP}, H_i)P(\mathbf{w}_{MP}|H_i)\det{}^{-\frac{1}{2}}(\mathbf{A}/2\pi),$$

где $\mathbf{A} = -\nabla^(2)\ln P(\mathbf{w}|D, H_i)$ — гессиан ковариационной матрицы весов, вычисленный в точке $\mathbf{w}_{MP}$. Аглоритмически, Байесовский метод сравнения моделей посредством вычисления достоверности не сложнее, чем задача настройки параметров каждой модели и оценки матрицы Гессе.

Итак, для того, чтобы отранжировать альтернативные модели H_i по предпочтению, необходимо, воспользовавшись Байесовским выводом, вычислить достоверность $P(D|H_i)$. Байесовское сравнение моделей — это расширение метода выбора моделей по методу наибольшего правдоподобия. Достоверность возможно вычислить как для параметрических, так и для непараметических моделей.

5.4. Пример интепретации множителя Оккама

Переменная $\sigma_{w|D}$ является апостериорной неопределенностью вектора параметров $\mathbf{w}$. Пусть априорное распределение $P(\mathbf{w}|H_i)$ является равномерным на некотором большом интервале σ_w и отражает множество значений, которые были возможны априори, согласно модели H_i. Тогда $P(\mathbf{w}_{MP}|H_i) = 1/\sigma_w$ и

$$\text{множитель Оккама} = \frac{\sigma_{w|D}}{\sigma_w}.$$

Множитель Оккама есть степень сжатия пространства параметров модели при появлении данных. Модель H_i может быть представлена семейство параметрических функций, из которых фиксируется одна, как только появляются данные. Множитель есть число, обратное количеству таких функций (для конечно го их числа). Логарифм множителя Оккама есть есть мера количества информации о параметрах модели, которая будет получена при появлении данных.

На рисунке показаны три модели, H_1, H_2 и H_3, которые имеют равные априорные вероятности. Каждая модель имеет один параметр $\mathbf{w}$

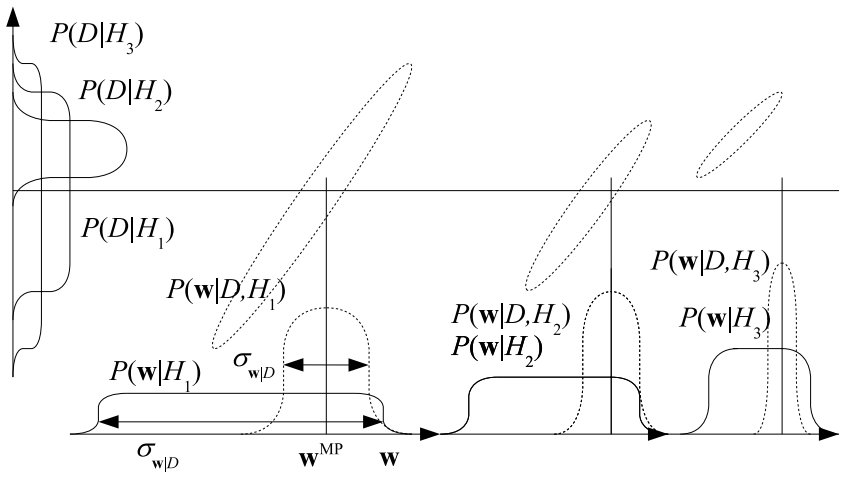

Рис. 9. Пространство данных и пространство параметров трех модели различной сложности

(показан на осях абсцисс), причем параметрам назначены различные априорные области определения σ_W. Модель H_3 — наиболее «гибкая», или «сложная», с наибольшей априорной областью определения. Одномерное пространство данных показано на оси ординат. Для каждой модели назначено совместное распределение вероятности $P(D, \mathbf{w}|H_i)$ для данных и для параметров. Распределение показано облаками точек — случайных значений этой функции. Число точек для каждой из трех моделей одинаково, так как моделям были назначены одинаковые априорные вероятности.

Когда приходят набор данных D (в данном примере это одна единственная точка на оси ординат), выводится апостериорное распределение параметров моделей. Апостериорное распределение $P(\mathbf{w}|D, H_i)$ показано пунктирной линией внизу. Сплошной линией показано априорное распределение параметров $P(\mathbf{w}|H_3)$.

Для набора данных D, показанных пунктирной горизонтальной линией, достоверность $P(D|H_3)$ наиболее гибкой модели H_3 имеет меньшее значение, чем достоверность модели H_2. Это происходит из-за того, что модель H_3 имеет меньшую область пересечения распреде-

ления вероятности $P(D, \mathbf{w}|H_i)$ с линиией D, чем модель H_2. В терминах распределения параметров, модель H_3 имеет меньшую достоверность, так как множитель Оккама $\sigma_{w|D}/\sigma_w$ для модели H_3 меньше, чем для модели H_2. Самая простая модель H_1 имеет самую малую достоверность, так как хуже всего приближает данные D. Модель H_3 слишком универсальна, ее множитель Оккама — штраф за универсальность модели — велик и поэтому она не является лучшей. Для полученного набора данных наиболее вероятна модель H_2.

5.5. Оценка значимости элементов моделей

Оценка значимости элементов моделей (optimal brain surgery, LeCun, 1990) — метод упрощения структуры регрессионной модели, например, нейронной сети. Основная идея прореживания заключается в том, что те элементы модели или те нейроны сети, которые оказывают малое влияние на ошибку аппроксимации, можно исключить из модели без значительного ухудшения качества аппроксимации.

Рассмотрим регрессионную модель $y_n = f(\mathbf{w}, \mathbf{x}_n)+\nu$, в которой $\mathbf{x}$ — независимая переменная, y — [[регрессионный анализ|зависимая переменная]], $\mathbf{w}$ — параметры регрессионной модели f, и ν — аддитивная [[случайная величина]]. Задана [[выборка|регрессионная выборка]] — множество пар $D = \{(\mathbf{x}_n, y_n)\}$, $n = 1, \ldots, N$. Для построения регрессии требуется найти такие параметры $\mathbf{w}^{MP}$, которые доставляли бы наименьшее значение функции ошибки E_D.

Найдем локальную аппроксимацию функции E_D в окрестности точки $\mathbf{w}^{MP}$ с помощью разложения в ряд Тейлора:

$$E_D(\mathbf{w} + \Delta\mathbf{w}) = E_D(\mathbf{w}) + \mathbf{g}^T(\mathbf{w})\Delta\mathbf{w} + \frac{1}{2}\Delta\mathbf{w}^T H \Delta\mathbf{w} + o(\|\mathbf{w}\|^3),$$

где $\mathbf{w}$ — возмущение вектора параметров $\mathbf{w}$, $\mathbf{g}$ — градиент $\frac{\partial S}{\partial \mathbf{w}}$, и $H = H(\mathbf{w})$ — матрица вторых производных (матрица Гессе) $\frac{\partial^2 S}{\partial \mathbf{w}^2}$.

Предполагается, что функция $E_D(\mathbf{w})$ достигает своего максимума при значении параметров $\mathbf{w} = \mathbf{w}^{MP}$ и ее поверхность квадратична. Таким образом, предыдущее выражение можно упростить и представить в виде

$$\Delta E_D = E_D(\mathbf{w} + \Delta\mathbf{w}) - E_D(\mathbf{w}) = \frac{1}{2}\Delta\mathbf{w}^T H \Delta\mathbf{w}.$$

Пусть исключение элемента модели есть исключение одного параметра модели, w_i. Исключенный параметр будем считать равным нулю. Это самое сильное ограничение, не позволяющее применять данный метод для регрессионных моделей произвольного вида. Исключение элемента эквивалентно выражению $\Delta w_i + w_i = 0$, иначе

$$\mathbf{e}_i^T \Delta\mathbf{w} + w_i = 0,$$

где $\mathbf{e}_i$ — вектор, i-й элемент которого равен единице, все остальные элементы равны нулю.

Для нахождения исключаемого элемента требуется минимизировать квадратичную форму $\Delta\mathbf{w}^T H \Delta\mathbf{w}$ относительно $\Delta\mathbf{w}$ при ограничениях $\mathbf{e}_i^T + w_i = 0$, для всех значений i. Индекс i, который доставляет минимум квадратичной форме, задает номер исключаемого элемента:

$$i = \arg\min_i(\min_{\Delta\mathbf{w}}(\Delta\mathbf{w}^T H \Delta\mathbf{w} | \mathbf{e}_i^T + w_i = 0)).$$

Задача условной минимизации решается с помощью введения Лагранжиана

$$S = \Delta\mathbf{w}^T H \Delta\mathbf{w} - \lambda(\mathbf{e}_i^T + w_i),$$

в котором λ — множитель Лагранжа. Дифференцируя Лагранжиан по приращению параметров и приравнивая его к нулю получаем (для каждого индекса i параметра w_i)

$$\Delta\mathbf{w} = -\frac{w_i}{[H^{-1}]_{ii}} H^{-1}\mathbf{e}_i.$$

Этому значению вектора приращений параметров соответствует минимальное значение Лагранжиана

$$L_i = \frac{w_i^2}{2[H^{-1}]_{ii}}.$$

Полученное выражение называется мерой выпуклости функции ошибки E_D при изменении параметра w_i.

Функция L_i зависит от квадрата параметра w_i. Это что говорит о том, что параметр с малым значением скорее всего будет удален из модели. Однако если величина $[H^{-1}]_{ii}$ достаточно мала, это означает, что данный параметр оказывает существенное влияние на качество аппроксимации модели.

Алгоритм упрощения регрессионной модели. Задана выборка D, модель $f(\mathbf{w}, \mathbf{x})$, функция ошибки E_D. Для упрощения структуры регрессионной модели выполняем следующие шаги.

1. Настраиваем модель, получаем параметры $\mathbf{w}^{MP} = \arg\min(E_D(\mathbf{w}|f, D))$.
2. Для приращения $\mathbf{w}^{MP} + \Delta\mathbf{w}$ решаем оптимизационную задачу, находим для каждого индекса i минимальное значение Лагранжиана L_i.
3. Выбираем среди L_i минимальное, отсекаем элемент модели, соответствующий i-му параметру.
4. Добавляем к вектору параметров $\mathbf{w}^{MP}$, вектор приращений $\Delta\mathbf{w}$, соответствующий отсеченому параметру.
5. Получаем упрощенную модель. Модель перенастраивать не требуется.
6. Процедуру можно повторять до тех пор, пока значение ошибки не превзойдет заранее заданное.

Литература

[1]

[2]

[3] *Айвазян, С. А.* Прикладная статистика и основы эконометрики / С. А. Айвазян. — М.: Юнити, 2001.

[4] *Айвазян, С. А.* Прикладная статистика: основы моделирования и первичная обработка данных / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. — М.: Финансы и статистика, 1983.

[5] *Айвазян, С. А.* Прикладная статистика: исследование зависимостей / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. — М.: Финансы и статистика, 1985.

[6] *Брандт, З.* / З. Брандт. — М.: Мир.

[7] *Гилязов, С. Ф.* Методы решения линейных некорректных задач / С. Ф. Гилязов. — М.: Изд-во МГУ, 1987.

[8] *Голуб, Д.* Матричные вычисления / Д. Голуб, Ч. Ван-Лоун. — М.: Мир, 1999.

[9] *Гордин, В. А.* Как это посчитать?: Обратобка метеорологической информации на компьютере: Идеи, методы, алгоритмы, задачи / В. А. Гордин. — М.: МЦМНО, 2005. — 280 с.

[10] *Демиденко, Е. З.* Оптимизация и регрессия / Е. З. Демиденко. — М.: Наука, 1989. — 296 с.

[11] *Демиденко, Е. И.* Оптимизация и регрессия / Е. И. Демиденко. — М.: Наука., 1989. — 296 с.

[12] *Деммель, Д.* Вычислительная линейная алгебра: теория и приложения / Д. Деммель. — М.: Мир, 2001. — 429 с.

[13] *Дрейпер, Н.* Прикладной регрессионный анализ / Н. Дрейпер, Г. Смит. — Издательский дом “Вильямс”, 2007.

[14] *Ивахненко, А. Г.* Индуктивный метод самоорганизации моделей сложных систем / А. Г. Ивахненко. — Киев: Наукова думка, 1981. — 296 с.

[15] *Ивахненко, А. Г.* Помехоустойчивость моделирования / А. Г. Ивахненко, В. С. Степашко. — Киев: Наукова думка, 1985. — 216 с.

[16] *Ивахненко, А. Г.* Моделирование сложных систем по экспериментальным данным / А. Г. Ивахненко, Ю. П. Юрачковский. — М.: Радио и связь, 1987. — 120 с.

[17] *Каханер, Д.* Численные методы и программное обеспечение / Д. Каханер, К. Моулер, С. Нэш. — М.: Мир, 1998.

[18] *Краснощеков, П. С.* Принципы построения моделей / П. С. Краснощеков, А. А. Петров. — М.: Фазис, 2000.

[19] *Кук, Д.* Компьютерная математика / Д. Кук, Г. Бейз. — М.: Наука, 1990.

[20] *Логинов, Н. В.* Сингулярное разложение матриц / Н. В. Логинов. — М.: МГАПИ, 1996.

[21] *Миркес, Е. М.* Нейрокомпьютер. Проект стандарта / Е. М. Миркес. — Новосибирск: Наука, Сибирская издательская фирма РАН, 1999. — 337 с.

[22] Прикладная статистика: классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. — М.: Финансы и статистика, 1989.

[23] *Рао, С. Р.* Линейные статистические методы и их применения / С. Р. Рао. — М.: Наука, 1968. — 547 с.

[24] *Стренг, Г.* Линейная алгебра и ее применения / Г. Стренг. — М.: Мир, 1980. — 454 с.

[25] *Тихонов, А. Н.* Методы решения некорректных задач / А. Н. Тихонов, В. Я. Арсенин. — М.: Наука, 1986. — 284 с.

[26] *Форсайт, Д.* Численное решение систем линейных алгебраических уравнений / Д. Форсайт, К. Молер. — М.: Мир, 1969. — 167 с.

[27] *Хайкин, С.* Нейронные сети, полный курс / С. Хайкин. — М: Вильямс, 2008. — 1103 с.

[28] *Хардле, В.* Прикладная непараметрическая регрессия / В. Хардле. — М.: Мир, 1993. — 349 с.

[29] *Хорн Р.and Джонсон, Ч.* Матричный анализ / Ч. Хорн, Р.and Джонсон. — М.: Мир, 1989.

[30] *Bishop, C.* Pattern Recognition And Machine Learning / C. Bishop. — Springer, 2006.

[31] *Bishop, C. M.* Bayesian regression and classification / C. M. Bishop, M. E. Tipping.

[32] *Burnham, K.* Model Selection and Multimodel Inference / K. Burnham, D. R. Anderson. — Springer, 2002.

[33] *Grunwald, P. D.* Advances In Minimum Description Length: Theory And Applications / P. D. Grunwald, I. J. Myung. — Springer, 2005.

[34] *Hassibi, B.* Second order derivatives for network pruning: Optimal brain surgeon / B. Hassibi, D. G. Stork // Advances in Neural Information Processing Systems / Ed. by S. J. Hanson, J. D. Cowan, C. L. Giles. — Vol. 5. — Morgan Kaufmann, San Mateo, CA, 1993. — Pp. 164–171. citeseer.ist.psu.edu/hassibi93second.html.

[35] *Hastie, T.* The Elements of Statistical Learning / T. Hastie, R. Tibshirani, J. Friedman. — Springer, 2001.

[36] *Kearns, M. J.* Efficient distribution-free learning of probabilistic concepts / M. J. Kearns, R. E. Schapire // Computational Learning Theory and Natural Learning Systems, Volume I: Constraints and Prospect, edited by Stephen Jose Hanson, George A. Drastal, and Ronald L. Rivest, Bradford/MIT Press. — 1994. — Vol. 1. citeseer.ist.psu.edu/article/kearns93efficient.html.

[37] *Koza, J. R.* Genetic Programming IV: Routine Human-Competitive Machine Intelligence / J. R. Koza. — Springer, 2005.

[38] *Lehmann, E. L.* Testing Statistical Hypotheses / E. L. Lehmann, J. P. Romano. — Springer, 2005.

[39] *MacKay, D.* Information Theory, Inference, and Learning Algorithms / D. MacKay. — Cambridge University Press, 2003.

[40] *Malada, H. R.* Inductive Learning Algorithms for Complex Systems Modeling / H. R. Malada, A. G. Ivakhnenko. — CRC Press, 1994. — 368 pp.

[41] *Mueler, J. A.* Sel-organising Data Mining: An Intelligent Approach To Extract Knowledge From Data / J. A. Mueler, F. Lemke. — Berlin: Dresden, 1999. — 225 pp.

[42] *Nabney, Y. T.* Netlab: Algorithms for pattern recognition / Y. T. Nabney. — Springer, 2004.

[43] Optimal brain damage / Y. LeCun, J. Denker, S. Solla et al. // Advances in Neural Information Processing Systems II / Ed. by D. S. Touretzky. — San Mateo, CA: Morgan Kauffman, 1990. citeseer.ist.psu.edu/lecun90optimal.html.

[44] *Vetterling, W. T.* Numerical Recipies in C: The Art of Scientific Computing / W. T. Vetterling, F. B. P. — NY: Cambridge University Press, 1999.

Содержание