

Лекция 4. Задачи выбора модели

Д. П. Ветров¹ Д. А. Кропотов²

¹МГУ, ВМиК, каф. ММП

²ВЦ РАН

Спецкурс «Байесовские методы машинного обучения»

План лекции

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

1 Ликбез

2 Постановка задачи выбора модели

Общий характер проблемы выбора модели

Примеры задач выбора модели

3 Общие методы выбора модели

Кросс-валидация

Теория Вапника-Червоненкиса

Принцип минимальной длины описания

Информационные критерии

Оптимальное кодирование

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

- Рассматривается задача кодирования алфавита \mathcal{A} словами из алфавита \mathcal{B} , как правило содержащего меньше символов
- Пусть каждый символ $a \in \mathcal{A}$ встречается в текстах с вероятностью $p(a)$. Обозначим $l(a)$ длину его кодирования в \mathcal{B}
- Задача построить схему кодирования, обеспечивающую минимальную среднюю длину кодированных сообщений, т.е.

$$\mathbb{E}_{\mathcal{A}} l(a) = \sum_{a \in \mathcal{A}} p(a) l(a) \rightarrow \min$$

Теорема Шеннона

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

- Теорема Шеннона. Если кодируемые элементы a могут встречаться с разными вероятностями $p(a)$, то существует оптимальное кодирование данного множества элементов и длина описания элемента a равна $l(a) = -\log_B p(a)$
- Основание логарифма B — это мощность кодирующего алфавита. Если алфавит \mathcal{B} состоит из двух символов (например, ноль и единица), то логарифм двоичный, а длина описания измеряется в битах
Если логарифм натуральный, то длина описания измеряется в натах, хотя довольно трудно представить себе алфавит из 2.7182... символов :)
- Теорема Шеннона согласуется со здравым смыслом: чем чаще встречается символ, тем короче должна быть длина его описания, и наоборот
Лень Морзе дорого обошлась (и обходится) человечеству из-за того, что пропускаемая способность каналов связи оказалась ниже оптимальной

Проблема выбора модели I

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общий характер
проблемы
выбора модели

Примеры задач
выбора модели

Общие методы
выбора модели

- Рассмотрим произвольную задачу машинного обучения
- Дана обучающая выборка (X, T) и некоторое параметрически заданное решающее правило, по которому мы получаем оценку на скрытую компоненту t для произвольного объекта с вектором признаков \mathbf{x}

$$\hat{t} = f(\mathbf{x}, \mathbf{w})$$

- Настройка параметров (весов) решающего правила \mathbf{w} производится путем минимизации некоторого функционала качества, связанного с ошибкой на обучающей выборке

Проблема выбора модели II

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общий характер
проблемы
выбора модели

Примеры задач
выбора модели

Общие методы
выбора модели

- Нетрудно придумать алгоритм, блестяще работающий на обучающей выборке (например, запомнить для нее правильные ответы). Вот только на новых объектах такой алгоритм, скорее всего, будет работать плохо
- Возникает идея ограничить множество допустимых решающих правил, чтобы в процессе обучения мы не могли бы получить «плохие» решения
- Ценой неизбежно становится ухудшение работы алгоритма на обучающей выборке
- Процесс выбора модели в машинном обучении — это поиск компромисса между точностью решающего правила на обучении и его «надежностью» на произвольных объектах генеральной совокупности

Определение модели

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общий характер
проблемы
выбора модели

Примеры задач
выбора модели

Общие методы
выбора модели

- Пусть $f(., \mathbf{w})$ — решающее правило, полученное в результате настройки весов $\mathbf{w} \in \Omega$ в ходе обучения
- Модель — это совокупность всех решающих правил, которые получаются путем присваивания весам всех возможных допустимых значений $\{f(., \mathbf{w}) | \mathbf{w} \in \Omega\}$
- Модель определяется множеством допустимых весов Ω , и структурой решающего правила $f(., .)$
- Заметим, что ограничения на множество допустимых весов могут быть нечеткими и задаваться, например, с помощью функции, штрафующей те или иные значения \mathbf{w} на этапе обучения

Задача выбора модели

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общий характер
проблемы
выбора модели

Примеры задач
выбора модели

Общие методы
выбора модели

- Выбрать модель — значит определить множество Ω , и определить структуру (схему, по которой настраиваемые веса будут образовывать композицию с входными данными) решающего правила $f(.,.)$
- Поскольку рассмотреть всевозможные множества, и структуры невозможно, их обычно ограничивают некоторым параметрическим семейством, зависящим от **структурных параметров** или параметров модели
- Под задачей выбора модели будем понимать проблему автоматической настройки всех структурных параметров для данного алгоритма машинного обучения

Почему сложно выбрать модель

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общий характер
проблемы
выбора модели

Примеры задач
выбора модели

Общие методы
выбора модели

- Настройку структурных параметров **нельзя проводить, минимизируя ошибку на обучении**, т.к. эти параметры по смыслу как раз должны ограничивать допустимые решающие правила, чтобы избежать эффекта переобучения
- Таким образом, последовательность действий такая: сначала из некоторых соображений выбирается модель $\{f(\cdot, \cdot), \Omega\}$, а потом в рамках выбранной модели ищется наилучшее решающее правило, путем оптимизации функционала качества, связанного с ошибкой на обучении, по параметрам \mathbf{w}
- Ошибки в выборе модели могут приводить к недообучению или переобучению
- Но тогда непонятно, как ввести качество модели...

Философский смысл проблемы выбора модели

Лекция 4.
Задачи выбора модели

Ветров,
Кропотов

Ликбез

Постановка задачи выбора модели

Общий характер проблемы выбора модели

Примеры задач выбора модели

Общие методы выбора модели

- Данная проблема проявляется в различных областях математики в частности, и человеческой деятельности вообще
- Не будет преувеличением сказать, что она носит общенаучный и даже философский характер
- Это проблема выбора средств решения задачи: по воробьям можно стрельнуть из рогатки, а можно пальнуть из пушки; длину комнаты можно измерить в локтях, а можно с помощью штангенциркуля и т.п.
- Более сложные средства, как правило, приводят к более точным результатам, вот только не всегда нам нужна такая точность. А за сложные средства приходится платить...

Задача классификации методом опорных векторов

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общий характер
проблемы
выбора модели

Примеры задач
выбора модели

Общие методы
выбора модели

- Решающее правило имеет вид

$$f(\mathbf{x}, \mathbf{w}) = \text{sign} \left(\sum_{i=1}^n w_i K(\mathbf{x}, \mathbf{x}_i) + b \right)$$

- Оптимизационная задача для поиска весов

$$\sum_{i=1}^n w_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n t_i t_j w_i w_j K(\mathbf{x}_i, \mathbf{x}_j) \rightarrow \max$$

$$\sum_{i=1}^n t_i w_i = 0 \quad 0 \leq w_i \leq C$$

- Коэффициент регуляризации C и ядровая функция $K(\mathbf{x}', \mathbf{x}'')$ определяют модель метода опорных векторов

Результаты классификации SVM с различными структурными параметрами

Лекция 4.
Задачи выбора модели

Ветров,
Кропотов

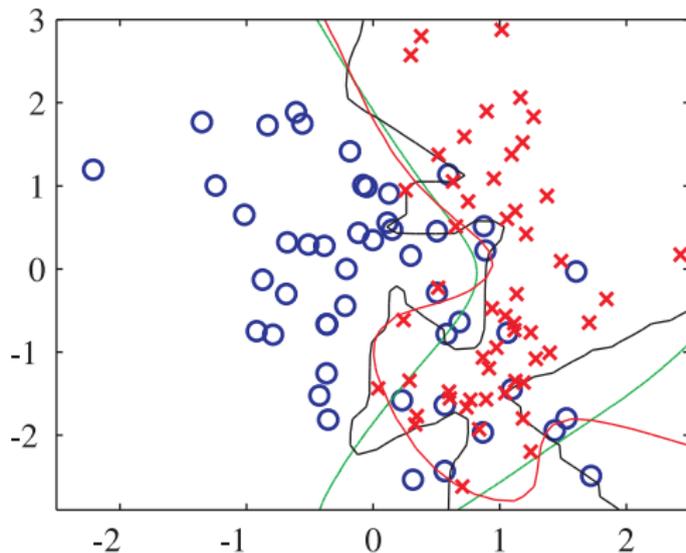
Ликбез

Постановка
задачи выбора модели

Общий характер
проблемы
выбора модели

Примеры задач
выбора модели

Общие методы
выбора модели



Задача регрессии

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общий характер
проблемы
выбора модели

Примеры задач
выбора модели

Общие методы
выбора модели

- Обобщенная линейная регрессия

$$f(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^m w_j \phi_j(\mathbf{x})$$

- Веса регрессии вычисляются по следующей формуле

$$\mathbf{w} = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T \mathbf{t}$$

- Параметр регуляризации $\lambda \geq 0$, система базисных функций $\{\phi_j(\mathbf{x})\}_{j=1}^m$ и их количество m определяют модель

Результаты восстановления регрессии с различными базисными функциями

Лекция 4.
Задачи выбора модели

Ветров,
Кропотов

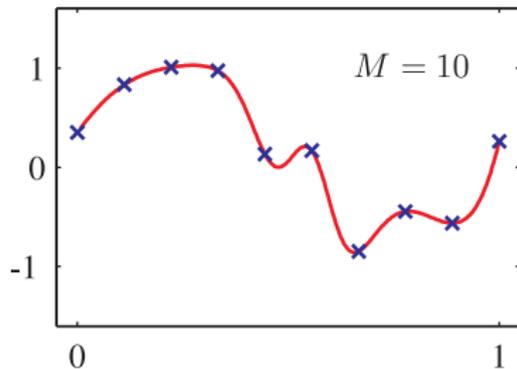
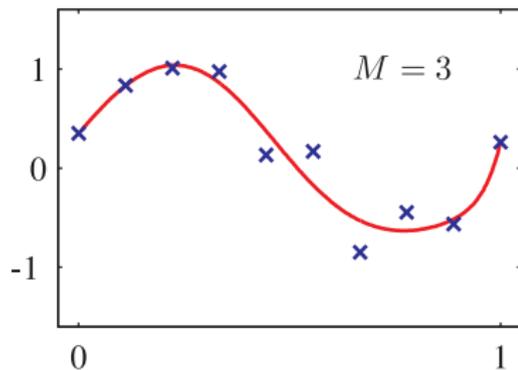
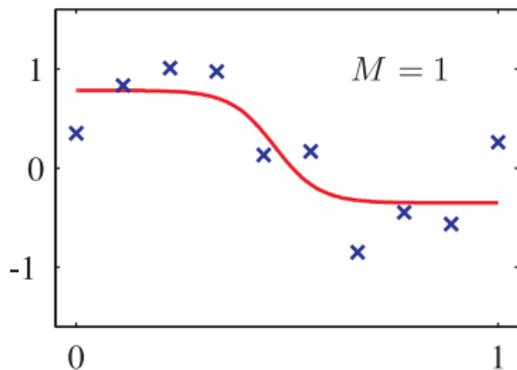
Ликбез

Постановка задачи выбора модели

Общий характер проблемы выбора модели

Примеры задач выбора модели

Общие методы выбора модели



Задача кластеризации

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

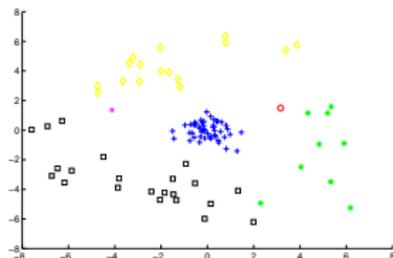
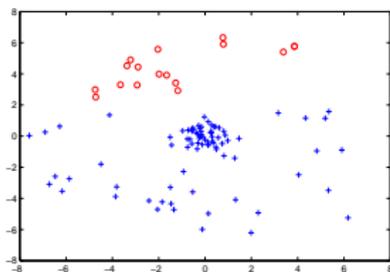
Постановка
задачи выбора
модели

Общий характер
проблемы
выбора модели

Примеры задач
выбора модели

Общие методы
выбора модели

- Большинство методов кластеризации предполагают задание пользователем количества кластеров, на которые будут разбиваться входные данные



Нейронные сети

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

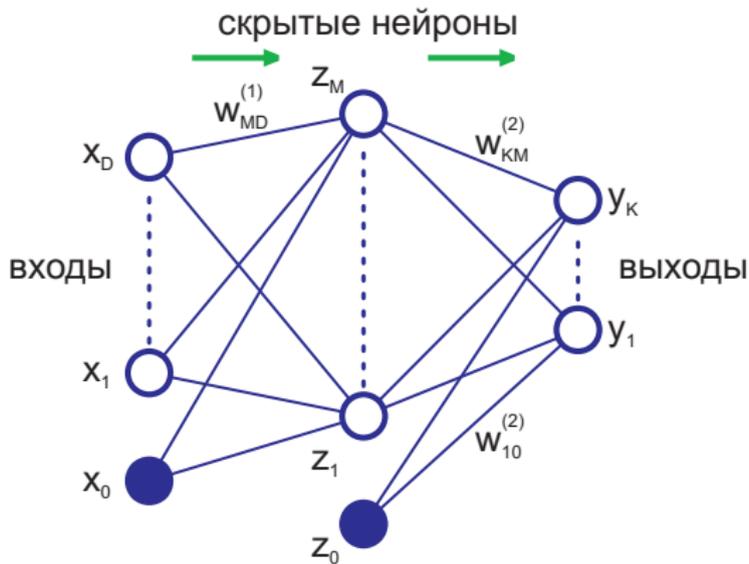
Постановка
задачи выбора
модели

Общий характер
проблемы
выбора модели

Примеры задач
выбора модели

Общие методы
выбора модели

- Выбор архитектуры нейронной сети (количество нейронов на каждом уровне) и функции активации определяют модель нейронной сети



Что такое общие методы выбора модели?

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

Кросс-
валидация

Теория Вапника-
Червоненкиса

Принцип
минимальной
длины описания

Информационные
критерии

- Под общими методами выбора модели будем понимать алгоритмы, позволяющие проводить автоматическую настройку **любых** структурных параметров для широкого множества задач машинного обучения (например, для всех задач классификации)
- На практике такой метод может использоваться для настройки лишь некоторых структурных параметров, но гипотетически он должен позволять определять любые характеристики наилучшей модели

Скольльзящий контроль

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

Кросс-
валидация

Теория Вапника-
Червоенкиса

Принцип
минимальной
длины описания

Информационные
критерии

- Процедура скользящего контроля (кросс-валидации) заключается в последовательном исключении части объектов из обучающей выборки, обучении на оставшихся объектах и распознавании исключенных объектов
- Тем самым эмулируется наличие тестовой выборки, которая не участвует в обучении, но для которой известны правильные ответы
- Структурные параметры настраиваются путем минимизации ошибки на скользящем контроле
- Процедура скользящего контроля является на сегодняшний день **самым лучшим средством настройки структурных параметров**

Схема k-fold cross validation

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

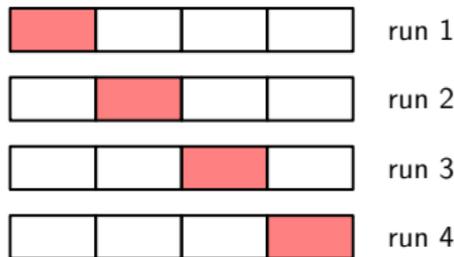
Кросс-
валидация

Теория Вапника-
Червоненкиса

Принцип
минимальной
длины описания

Информационные
критерии

- Выборка разбивается на k непересекающихся (одинаковых по объему) частей. На каждой итерации обучение проводится по $k - 1$ части, а тестирование на исключенных объектах



- На рисунке приведена процедура 4-fold cross validation
- При $k = n$ процедура называется leave-one-out
- Наилучшим режимом скользящего контроля считается 5×2 -fold cross validation.

Особенности скользящего контроля

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

Кросс-
валидация

Теория Вапника-
Червоенкиса

Принцип
минимальной
длины описания

Информационные
критерии

- Ошибка на скользящем контроле является довольно точной оценкой ошибки на генеральной совокупности (обобщающей способности)
- Проведение скользящего контроля требует значительного времени на многократное повторное обучение алгоритмов и применимо лишь для «быстрых» методов машинного обучения
- С помощью скользящего контроля можно настраивать не более двух-трех структурных параметров, т.к. настройка производится путем **полного перебора** всевозможных сочетаний параметров
- При его использовании для выбора модели ошибку на скользящем контроле **нельзя** рассматривать как оценку ошибки на генеральной совокупности, т.к. она получается заниженной
- Скользящий контроль плохо применим в задачах кластерного анализа и прогнозирования временных рядов

Идея структурной минимизации риска

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

Кросс-
валидация

Теория Вапника-
Червоненкиса

Принцип
минимальной
длины описания

Информационные
критерии

- Теория Вапника-Червоненкиса использует косвенные характеристики для оценки обобщающей способности (среднего риска)
- Ключевым понятием является т.н. емкость (размерность Вапника-Червоненкиса, VC-dimension) модели
- Идея данного подхода к выбору модели (структурной минимизации риска) заключается в следующем: **Чем более «гибкой» является модель, тем хуже ее обобщающая способность**
- В самом деле, «гибкое» решающее правило способно настроиться на малейшие шумы, содержащиеся в обучающей выборке

Понятие емкости

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

Кросс-
валидация

Теория Вапника-
Червоненкиса

Принцип
минимальной
длины описания
Информационные
критерии

- Рассмотрим задачу классификации на два класса
- (Несколько упрощая,) емкостью данной модели будем называть максимальное число объектов обучающей выборки, для которых при **любой** их разметке на классы найдется хотя бы одно решающее правило, безошибочно их классифицирующее
- По аналогии вводятся определения емкости для других задач машинного обучения
- Важный пример модели, для которой известна емкость — классификатор, строящий линейную гиперплоскость. Емкость линейного классификатора равна $h = d + 1$, где d — размерность пространства признаков
- Следствие: $n \leq d + 1$ объектов **всегда** можно безошибочно разделить гиперплоскостью

Формула Вапника

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

Кросс-
валидация

Теория Вапника-
Червоненкиса

Принцип
минимальной
длины описания

Информационные
критерии

- Очевидно, что чем больше емкость, тем хуже. Значит нужно добиваться минимально возможного количества ошибок на обучении при минимальной возможной емкости
- Ошибку на обучении (эмпирический риск) $P_{train}(\mathbf{w})$, емкость $h(\Omega)$ и ошибку на генеральной совокупности (средний риск) $P_{test}(\mathbf{w})$ связывает известная формула Вапника

$$P_{test}(\mathbf{w}) \leq P_{train}(\mathbf{w}) + \sqrt{\frac{h(\Omega)(\log(2d/h(\Omega)) + 1) - \log(\eta/4)}{n}}$$

Неравенство верно с вероятностью $1 - \eta$ для $\forall \mathbf{w} \in \Omega$

- Последовательно анализируя модели с увеличивающейся емкостью, согласно теории ВЧ, необходимо выбирать модель с наименьшей верхней оценкой тестовой ошибки

Достоинства и недостатки теории ВЧ

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели
Кросс-
валидация

Теория Вапника-
Червоненкиса

Принцип
минимальной
длины описания
Информационные
критерии

- Достоинства
 - Серьезное теоретическое обоснование, связь с ошибкой на генеральной совокупности
 - Теория продолжает развиваться и в наши дни (эффективная емкость, локальная емкость, комбинаторный подход и т.д.)
- Недостатки
 - Оценки сильно завышены
 - Для большинства моделей емкость не поддается оценке
 - Многие модели с бесконечной емкостью показывают хорошие результаты на практике

Пути развития теории ВЧ (и не только ее)

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

Кросс-
валидация

Теория Вапника-
Червоненкиса

Принцип
минимальной
длины описания
Информационные
критерии

- Емкость вводится для всевозможных положений объектов выборки, в то время как реально приходится иметь дело с одной конкретной выборкой и опять-таки имеет смысл рассматривать степень адаптируемости модели под эту конкретную выборку, а не под абстрактно возможную
- В процессе обучения поиск алгоритма в модели ведется не по всем ее представителям, а лишь по конечному числу, которое и имеет смысл рассматривать при интерпретации емкости, как степени адаптируемости модели под данные
- Искомая закономерность может обладать рядом дополнительных свойств, которые сокращают объем допустимых алгоритмов модели

Предпосылка метода

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели
Кросс-
валидация
Теория Вапника-
Червонекса

Принцип
минимальной
длины описания

Информационные
критерии

- Из пункта А в пункт В передается закодированное сообщение о классификации обучающей выборки. Нужно добиться минимально возможного размера сообщения
- Стратегия 1: передаем каждый объект и его метку класса $\{(\mathbf{x}_i, t_i)\}_{i=1}^n$
- Стратегия 2: передаем длинное описание сложного алгоритма, который можно использовать для правильной классификации всей обучающей выборки $Descr(A)$
- Стратегия 3: передаем короткое описание простого алгоритма $Descr(A')$, который правильно классифицирует большинство объектов обучающей выборки, а классификацию неправильно распознанных объектов передаем отдельным списком $\{\mathbf{x}_{i_k}, t_{i_k}\}_{k=1}^p$, $p < n$

Смысл метода

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

Кросс-
валидация

Теория Вапника-
Червоенкиса

Принцип
минимальной
длины описания

Информационные
критерии

- Чем точнее на обучающей выборке алгоритм, тем он сложнее, а значит тем длиннее будет его описание...
- ... но тем меньше будет список неправильно распознанных объектов
- Принцип минимальной длины описания (minimum description length MDL, Rissanen, 1978) штрафует излишнюю алгоритмическую сложность решающего правила



Особенности подсчета длины описания

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

Кросс-
валидация

Теория Вапника-
Червоenkиса

Принцип
минимальной
длины описания

Информационные
критерии

- Существует множество подходов к оценке длины описания алгоритма вплоть до длины кода реализующей его программы
- Необходимо отметить, что кодирование должно быть эффективным, т.к. даже самый простой алгоритм можно закодировать в очень длинное сообщение
- Согласно теореме Шеннона, при оптимальном кодировании длина описания структуры пропорциональна логарифму ее вероятности, взятому с противоположным знаком

Вероятностный MDL

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

Кросс-
валидация

Теория Вапника-
Червоненкиса

Принцип
минимальной
длины описания

Информационные
критерии

- Пусть на множестве алгоритмов задано априорное распределение $p(\mathbf{w})$

$$l(\mathbf{w}) = -\log p(\mathbf{w})$$

- Длина описания данных тем меньше, чем выше вероятность данной классификации при использовании данного алгоритма, т.е. чем выше правдоподобие $p(T|X, \mathbf{w})$

$$l(T|\mathbf{w}) = -\log p(T|X, \mathbf{w})$$

- Отсюда получаем выражение, объединяющее точность на обучении и сложность алгоритма **в единое выражение**

$$l(T, \mathbf{w}) = -\log p(T|X, \mathbf{w}) - \log p(\mathbf{w})$$

$$\arg \min_{\mathbf{w}} l(T, \mathbf{w}) = \arg \max_{\mathbf{w}} p(T|X, \mathbf{w})p(\mathbf{w})$$

- Таким образом, MDL обосновывает идею максимизации регуляризованного правдоподобия

Отличительные особенности MDL

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

Кросс-
валидация

Теория Вапника-
Червоненкиса

Принцип
минимальной
длины описания

Информационные
критерии

- MDL позволяет обосновать корректность регуляризации правдоподобия
- Область применения MDL шире, чем у статистических методов обучения, т.е. MDL можно применять и там, где вводить вероятности некорректно или бессмысленно
- При использовании MDL предполагается, что чем сложнее алгоритм, тем хуже его обобщающая способность. Современные исследования (в частности, boosting) показывают, что это далеко не всегда так

Информационный критерий Акаике

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

Кросс-
валидация

Теория Вапника-
Червоненкиса

Принцип
минимальной
длины описания

Информационные
критерии

- В 1973г. Акаике установил связь между правдоподобием (ключевое понятие статистики) и дивергенцией Кульбака-Лейблера (ключевое понятие в теории информации)
- Ему удалось получить приблизительное соотношение между правдоподобием генеральной совокупности и правдоподобием обучающей выборки (т.е. данных, по которым с помощью ММП производится настройка параметров решающего правила)

$$AIC = \log p(T|X, \mathbf{w}_{ML}) - M,$$

где M — число настраиваемых параметров

- Пример использования: задача восстановления регрессии с известным гауссовским шумом в одномерном пространстве при помощи полинома степени k

$$k = \arg \min \left(\frac{\sum_{i=1}^n (t_i - f^k(\mathbf{x}_i, \mathbf{w}_{ML}))^2}{2\sigma^2} + k + 1 \right)$$

Информационный критерий Шварца

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

Кросс-
валидация

Теория Вапника-
Червоненкиса

Принцип
минимальной
длины описания

Информационные
критерии

- Критерий Шварца (часто именуемый Байесовским информационным критерием) представляет собой простейшее приближение обоснованности, широко используемой в байесовском обучении

$$BIC \approx \int p(\mathbf{t}|X, \mathbf{w})p(\mathbf{w})d\mathbf{w}$$

- Используя приближение интеграла гауссианой и сильно огрубляя, получаем

$$BIC = \log p(T|X, \mathbf{w}_{MP}) - \frac{1}{2}M \log n$$

- Пример использования: Задача восстановления регрессии с известным гауссовским шумом в одномерном пространстве при помощи полинома степени k

$$k = \arg \min \left(\frac{\sum_{i=1}^n (t_i - f^k(\mathbf{x}_i, \mathbf{w}_{MP}))^2}{2\sigma^2} + (k + 1) \log \sqrt{n} \right)$$

Особенности информационных критериев

Лекция 4.
Задачи выбора
модели

Ветров,
Кропотов

Ликбез

Постановка
задачи выбора
модели

Общие методы
выбора модели

Кросс-
валидация

Теория Вапника-
Червоникса

Принцип
минимальной
длины описания

Информационные
критерии

- Оба критерия являются (весьма грубыми) приближениями более сложных выражений, часто не поддающихся аналитическому вычислению. Значение критерия может расцениваться лишь как приблизительная характеристика обобщающей способности полученного решающего правила
- Критерии разумно использовать когда **все M настраиваемых параметров оказывают примерно одинаковое влияние** на вид решающего правила, например, входят в него линейно.
- Байесовский критерий сильнее штрафует сложные (с точки зрения дополнительных параметров) модели
- В байесовский критерий входит значение правдоподобия в точке w_{MP} , а в критерий Акаике — в точке w_{ML}