

Регрессия

Виктор Китов
v.v.kitov@yandex.ru

МГУ им.Ломоносова, ф-т ВМиК, кафедра ММП.

I семестр 2015 г.

Усреднение в среднеквадратичном смысле

$$\sum_{n=1}^N (y_n - \mu)^2 \rightarrow \min_{\mu}$$

Условие стационарности:

$$2 \sum_{n=1}^N (y_i - \mu) = 0$$

$$\sum_{n=1}^N y_i - N\mu = 0$$

$$\mu = \frac{1}{N} \sum_{n=1}^N y_i$$

Усреднение по модулям отклонений

$$\sum_{n=1}^N |y_n - \mu| \rightarrow \min_{\mu}$$

Условие стационарности:

$$\sum_{n=1}^N \text{sign}(y_n - \mu) = 0$$

откуда видно, что производная равна нулю, когда μ меньше и больше одинакового количества y_i , что достигается, например, когда $\mu = \text{median}\{y_1, y_2, \dots, y_N\}$

Робастные оценки для ряда z_1, z_2, \dots, z_N :

среднее: медиана

$$\text{median}_i z_i$$

разброс: median absolute deviation

$$\text{median}_i \{|z_i - \text{median}_i z_i|\}$$

Оптимизация квадрата мат. ожидания

Теорема 1

Пусть $x, y \sim P(x, y)$ и $\mathbb{E}[y|x]$ существует. Тогда

$$\arg \min_{f(x)} \mathbb{E} \left\{ (f(x) - y)^2 \middle| x \right\} = \mathbb{E}[y|x]$$

$$\begin{aligned} \mathbb{E} \left\{ (f(x) - y)^2 \middle| x \right\} &= \mathbb{E} \left\{ (f(x) - \mathbb{E}[y|x] + \mathbb{E}[y|x] - y)^2 \middle| x \right\} \\ &= \mathbb{E} \left\{ (f(x) - \mathbb{E}[y|x])^2 \middle| x \right\} + \mathbb{E} \left\{ (\mathbb{E}[y|x] - y)^2 \middle| x \right\} \\ &\quad + 2\mathbb{E} \left\{ (f(x) - \mathbb{E}[y|x]) (\mathbb{E}[y|x] - y) \middle| x \right\} = \\ &= (f(x) - \mathbb{E}[y|x])^2 + \mathbb{E} \left\{ (\mathbb{E}[y|x] - y)^2 \middle| x \right\} \end{aligned} \quad (1)$$

Оптимизация квадрата мат. ожидания

Мы использовали

$$\begin{aligned}\mathbb{E} \{ (f(x) - \mathbb{E}[y|x]) (\mathbb{E}[y|x] - y) | x \} = \\ (f(x) - \mathbb{E}[y|x]) \mathbb{E} \{ \mathbb{E}[y|x] - y | x \} \equiv 0\end{aligned}$$

Минимум в (1) достигается при $f(x) = \mathbb{E}[y|x]$.

$\mathbb{E} \left\{ (\mathbb{E}[y|x] - y)^2 \middle| x \right\}$ определяет уровень естественного неуменьшаемого шума в данных.

Оптимизация модуля отклонения

Теорема 2

Пусть $x, y \sim P(x, y)$. Тогда

$$\arg \min_{f(x)} \mathbb{E} \{ |f(x) - y| \mid x \} = \text{median}[y|x]$$

$$\begin{aligned} \mathbb{E} \{ |\mu - y| \mid x \} &= \int_{-\infty}^{+\infty} |y - \mu| p(y|x) dy = \\ &= \underbrace{\int_{\mu}^{+\infty} (y - \mu) p(y|x) dy}_{I(\mu)} + \underbrace{\int_{-\infty}^{\mu} (\mu - y) p(y|x) dy}_{J(\mu)} \end{aligned}$$

Оптимизация модуля отклонения

Используя формулу дифференцирования по параметру

функции $F(\mu) = \int_{\alpha(\mu)}^{\beta(\mu)} f(y, \mu) dy$:

$$F'(\mu) = \int_{\alpha(\mu)}^{\beta(\mu)} f'_\mu(y, \mu) dy + \beta'(\mu)f(\beta(\mu), \mu) - \alpha'(\mu)f(\alpha(\mu), \mu)$$

получим:

$$I'(\mu) = \int_{\mu}^{+\infty} -p(y|x) dy - (\mu - \mu)p(\mu|x) = -P(y \geq \mu|x)$$

$$J'(\mu) = \int_{-\infty}^{\mu} p(y|x) dy + (\mu - \mu)p(\mu|x) = P(y \leq \mu|x)$$

Условие стационарности приобретает вид:

$$P(y \leq \mu|x) = P(y \geq \mu|x)$$

откуда следует, что $\mu = \text{median}\{y|x\}$

Линейная регрессия

- Линейная модель $f(x, \beta) = \langle x, \beta \rangle = \sum_{i=1}^D \beta_i x^i$
- Определим $X \in \mathbb{R}^{N \times D}$, $\{X\}_{ij}$ где j -й признак i -го объекта, $Y \in \mathbb{R}^n$, $\{Y\}_i$ - целевое значение для i -го объекта.
- Метод наименьших квадратов:

$$\sum_{n=1}^N (f(x, \beta) - y_n)^2 = \sum_{n=1}^N \left(\sum_{d=1}^D \beta_d x_n^d - y_n \right)^2 \rightarrow \min_{\beta}$$

Решение

Условие стационарности:

$$2 \sum_{n=1}^N \left(\sum_{d=1}^D \beta_d x_n^d - y_n \right) x_n^d = 0, \quad d = 1, 2, \dots, D.$$

В векторном виде:

$$2X^T(X\beta - Y) = 0$$

откуда

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Это глобальный минимум, поскольку целевой критерий - выпуклый.

- Геометрическая интерпретация МНК.

Ограничение решения

- Ограничение: матрица $X^T X$ должна быть невырожденной
 - возникает, когда один из признаков линейно выражается через другие
 - решается отбором признаков, преобразованием признаков (напр. PCA) или регуляризацией.
 - пример: константный признак $c = [1, 1, \dots, 1]^T$ и one-hot-encoding e_1, e_2, \dots, e_K , поскольку $\sum_k e_k \equiv c$

Анализ линейной регрессии

Преимущества:

- единственный оптимум (для невырожденной матрицы)
- аналитическое решение
- интерпретируемость алгоритма и решения

Недостатки:

- слишком простое модельное предположение (может не выполняться)
- необходимо следить, чтобы $X^T X$ была невырождена (и хорошо обусловлена)

Обобщение через нелинейные признаки

Нелинейности по x в линейной регрессии можно добиться, если применить нелинейные преобразования к признакам:

$$x \rightarrow [\phi_0(x), \phi_1(x), \phi_2(x), \dots, \phi_M(x)]$$

$$f(x) = \langle \phi(x), \beta \rangle = \sum_{m=0}^M \beta_m \phi_m(x)$$

По сути, модель остается линейной по w , поэтому все преимущества линейной регрессии сохраняются.

Типичные варианты преобразований

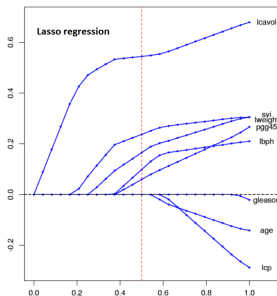
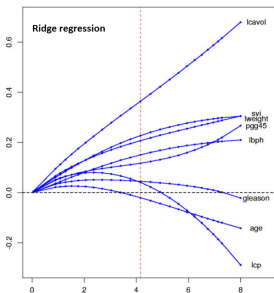
$\phi_k(x)$	комментарии
$\exp \left\{ -\frac{\ x-\mu\ ^2}{s^2} \right\}$	близость к точке признакового пространства
$x^i x^j$	взаимодействие признаков
$\ln x_k$	выравнивание распределения величины с тяжелыми хвостами
$F^{-1}(x_k)$	приведение нетипичного распределения к равномерному

Регуляризация

- Варианты критерия $Q(\beta)$ с регуляризацией:

$$\begin{aligned} & \|X\beta - Y\|^2 + \lambda\|\beta\|_1 && \text{Lasso} \\ & \|X\beta - Y\|^2 + \lambda\|\beta\|_2 && \text{Ridge} \\ & \|X\beta - Y\|^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2 && \text{Elastic net} \end{aligned}$$

- Зависимость коэффициентов β от $\frac{1}{\lambda}$:



Вероятностная интерпретация

Обозначим $X = \{x_1, x_2, \dots, x_N\}$ - признаковые описания обучающей выборки.

Если данные описываются следующей моделью:

$$\left\{ \begin{array}{l} y_i = f_{\theta}(x_i) + \varepsilon_i \\ \varepsilon_i - \text{независимы и одинаково распределены} \\ \varepsilon_i - \text{независимы от } x_i \\ \varepsilon_i \sim F(0, \sigma^2) \end{array} \right. |$$

Вероятностная интерпретация метода наименьших квадратов

$$F = N(0, \sigma^2)$$

Правдоподобие обучающей выборки:

$$p(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N | X) = \prod_{n=1}^N p(\varepsilon_i | X) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{n=1}^N e^{-\frac{(f_\theta(x_i) - y_i)^2}{2\sigma^2}} \rightarrow \max_{\theta}$$

Максимизация логарифма правдоподобия:

$$\text{const} - \sum_{n=1}^N \frac{1}{2\sigma^2} (f_\theta(x_i) - y_i)^2 \rightarrow \max_{\theta}$$

что эквивалентно:

$$\sum_{n=1}^N (f_\theta(x_i) - y_i)^2 \rightarrow \min_{\theta}$$

Вероятностная интерпретация метода наименьших квадратов

$$F = Laplace(0, 2b^2)$$

Правдоподобие обучающей выборки:

$$p(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N | X) = \prod_{n=1}^N p(\varepsilon_n | X) = \frac{1}{(2b)^N} \prod_{n=1}^N e^{-\frac{|f_\theta(x_i) - y_i|}{b}} \rightarrow \max_{\theta}$$

Максимизация логарифма правдоподобия:

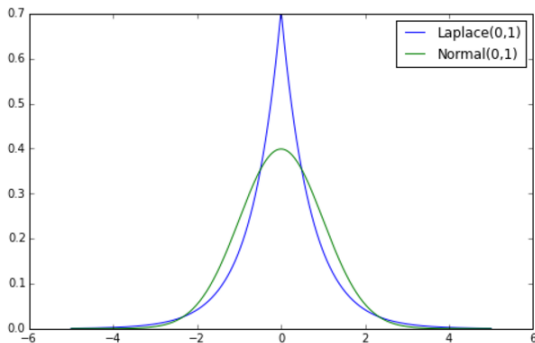
$$\text{const} - \sum_{n=1}^N \frac{1}{b} |f_\theta(x_i) - y_i| \rightarrow \max_{\theta}$$

что эквивалентно:

$$\sum_{n=1}^N |f_\theta(x_i) - y_i| \rightarrow \min_{\theta}$$

Распределение Лапласа и нормальное распределение

Laplace($\mu, 2b^2$)	$p(\varepsilon) = \frac{1}{2b} e^{-\frac{ \varepsilon-\mu }{b}}$
Normal(μ, σ^2)	$p(\varepsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



Линейная монотонная регрессия

- Можно экспертно налагать ограничения на коэффициенты, например неотрицательность:

$$\begin{cases} Q(\beta) = \|X\beta - Y\|^2 \rightarrow \min_{\alpha} \\ \beta_n \geq 0, \quad j = 1, 2, \dots, N \end{cases}$$

- Пример: усреднение индивидуальных прогнозов композицией алгоритмов
- $\beta_i = 0$ означает, что i -й компонент не добавляет точности прогнозирования.

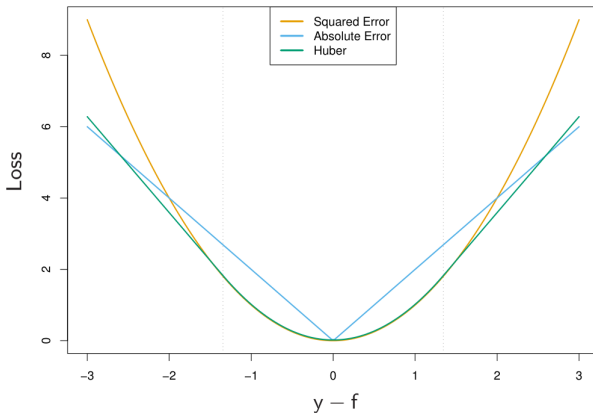
Модификации

- Взвешенный учет наблюдений

$$\sum_{n=1}^N w_n (x_n^T \beta - y_n)$$

- Веса могут быть:
 - увеличены для ошибочных объектов (оптимизируем алгоритм под исправление ошибок)
 - уменьшены для ошибочных объектов (считаем их выбросами, сбивающими модель)
- В вероятностной модели различные веса обозначают различные дисперсии.

Неквадратичные функции потерь



Нелинейная регрессия

- $f(x, \alpha)$ может быть нелинейной функцией:

$$Q(\alpha, X_{training}) = \sum_{i=1}^N (f(x_i, \alpha) - y_i)^2$$

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{R}^D} Q(\alpha, X_{training})$$

- Условие для нахождения α :

$$\frac{\partial Q}{\partial \alpha}(\alpha, X_{training}) = 2 \sum_{i=1}^N (f(x_i, \alpha) - y_i) \frac{\partial f}{\partial \alpha}(x_i, \alpha) = 0$$

- Исправление мультиколлинеарности, регуляризация, взвешенный учет наблюдений применимы и здесь.

Ядерная регрессия

$$f(x, \alpha) = \alpha, \alpha \in \mathbb{R}.$$

$$Q(\alpha, X_{training}) = \sum_{i=1}^N w_i(x)(\alpha - y_i)^2 \rightarrow \min_{\alpha \in \mathbb{R}}$$

Веса зависят от близости обучающих объектов к прогнозируемому объекту:

$$w_i(x) = K \left(\frac{d(x, x_i)}{h} \right)$$

Из условия стационарности $\frac{\partial Q}{\partial \alpha} = 0$ получаем оптимальное $\hat{\alpha}(x)$:

$$f(x, \alpha) = \hat{\alpha}(x) = \frac{\sum_i y_i w_i(x)}{\sum_i w_i(x)} = \frac{\sum_i y_i K \left(\frac{d(x, x_i)}{h} \right)}{\sum_i K \left(\frac{d(x, x_i)}{h} \right)}$$

Комментарии

При некоторых условиях регулярности $g(x, \alpha) \xrightarrow{P} E[y|x]$

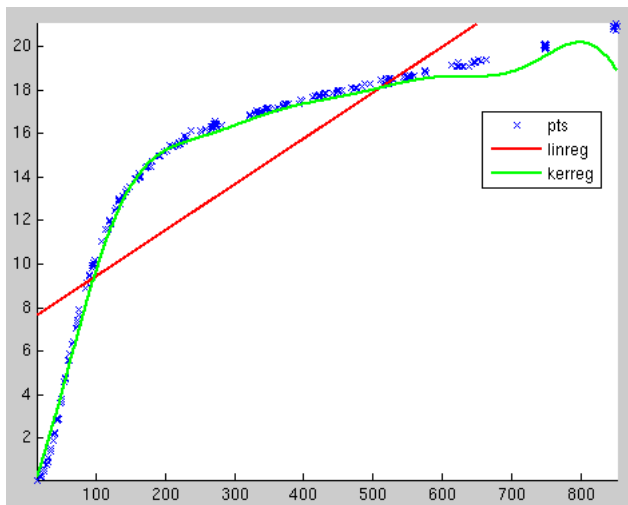
Обычно используются следующие ядра:

$$K_G(r) = e^{-\frac{1}{2}r^2} - \text{гауссово ядро}$$

$$K_P(r) = (1 - r^2)^2 \mathbb{I}[|r| < 1] - \text{квадратичное ядро}$$

- Конкретный вид ядерной функции не сильно влияет на точность
- решение с гауссовым ядром зависит от всех объектов, а с квадратичным - только от объектов $\{j : d(x, x_j) < h\}$.
- h контролирует адаптируемость модели к локальным изменениям в данных
 - можем получить переобученную/недообученную модель
 - h может быть постоянной или изменяемой (если концентрация объектов сильно меняется)
 - например $h(x)$ может быть расстоянием до K -го соседа. При $K(r) = \mathbb{I}[|r| < 1]$, получим метод K ближайших соседей.

Пример



Робастная ядерная регрессия

- Робастный алгоритм - значит алгоритм устойчив к редким и большим по величине выбросам.
- Для выбросов $\varepsilon_i = |y_i - f(x_i, \alpha)|$ велико.
- Идея - взвешивать ядра, поощряя регулярные наблюдения: $K(x, x_i) = D(\varepsilon_i)K(x, x_i)$
- Возможный выбор $D(\varepsilon)$:
 - $D(\varepsilon_i) = \mathbb{I}[\varepsilon_i \leq t]$, где t может быть выбрано как 95% квантиль ряда $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$.
 - $D(\varepsilon_i) = K_P \left(\frac{\varepsilon_i}{6 \text{med} \varepsilon_i} \right)$

$$f(x, \alpha) = \hat{\alpha}(x) = \frac{\sum_i y_i w_i(x)}{\sum_i w_i(x)} = \frac{\sum_i y_i D(\varepsilon_i) K \left(\frac{d(x, x_i)}{h} \right)}{\sum_i D(\varepsilon_i) K \left(\frac{d(x, x_i)}{h} \right)}$$

Алгоритм

- применить обычную ядерную регрессию для получения первичных прогнозов y_i
 - повторять до сходимости ε_i :
 - оценить $\varepsilon_i = y_i - \hat{\alpha}(x_i)$, $i = 1, 2, \dots, N$.
 - пересчитать $\hat{\alpha}(x_i)$ с обновленными $\varepsilon_1, \dots, \varepsilon_N$

Ядерная линейная аппроксимация

- Локальная (в окрестности x) аппроксимация

$$f(u) = (u - x)^T \beta + \beta_0$$

- Решить

$$Q(\alpha, \beta | X_{training}) = \sum_{i=1}^N w(x) ((x_i - x)^T \beta + \beta_0 - y_i)^2 \rightarrow \min_{\alpha, \beta \in \mathbb{R}}$$

- Обозначим $w_i = w_i(x)$, $d_i = x_i - x$.
- Из условий $\frac{\partial Q}{\partial \beta} = 0$ и $\frac{\partial Q}{\partial \beta_0} = 0$ получаем значения параметров.

Преимущества ядерной линейной регрессии:

- По сравнению с ядерной постоянной регрессией, ядерная линейная регрессия лучше прогнозирует:
 - локальные экстремумы
 - функцию на краях области определения