

Регрессия

Виктор Китов
v.v.kitov@yandex.ru

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

I семестр 2015 г.

Линейная регрессия

- Линейная модель $f(x, \beta) = \sum_{j=1}^N \beta_j x_j$
- 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:

$$Q(\beta) = \|X\beta - Y\|^2 \rightarrow \min_{\alpha}$$

$$\frac{\partial Q}{\partial \beta}(\beta) = 2X^T(X\beta - Y) = 0$$

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

Ограничения использования

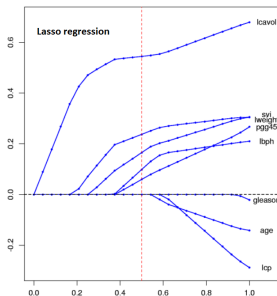
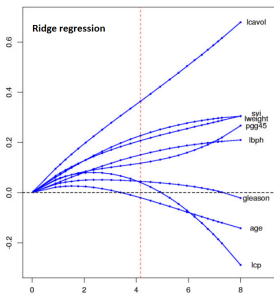
- матрица $X^T X$ должна быть невырожденной
 - возникает, когда один из признаков линейно выражается через другие
 - решается отбором признаков, преобразованием признаков (напр. PCA) или регуляризацией.
 - пример: константный признак $c = [1, 1, \dots, 1]^T$ и one-hot-encoding e_1, e_2, \dots, e_K ,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

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 $\frac{1}{\lambda}$:



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.

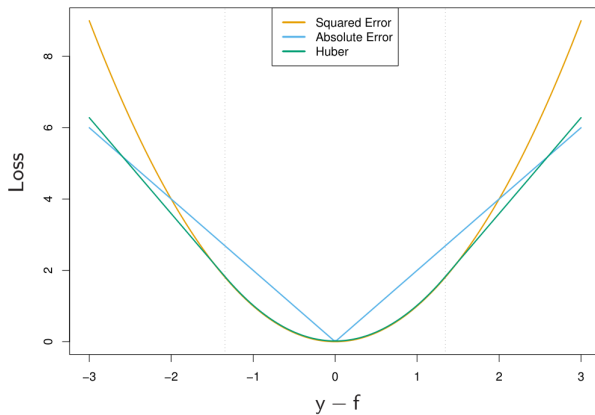
Модификации

- 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)

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

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

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

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

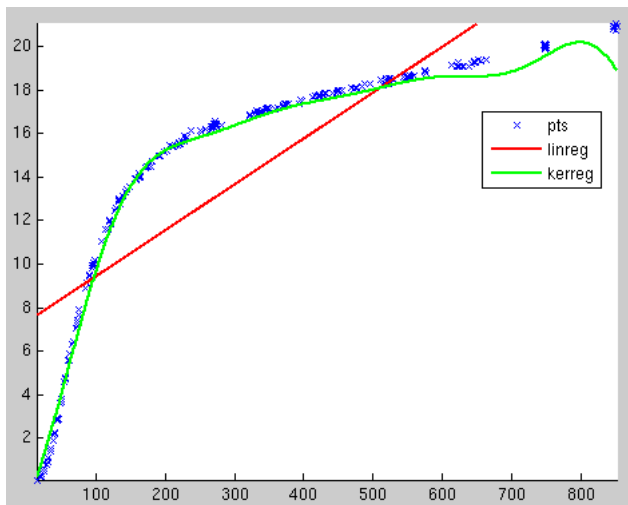
При некоторых условиях регулярности $g(x, \alpha) \xrightarrow{P} E[y|x]$
 Обычно используются следующие ядра

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

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

- Конкретный вид ядерной функции не сильно влияет на точность
- решение с гауссовым ядром зависит от всех объектов, а с квадратичным - только от объектов $\{i : d(x, x_i) < h\}$.
- 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

$$f(u) = \alpha(u - x) + \beta$$

- Решить

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

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

$$f(x) = \frac{\sum_i w_i d_i^2 \sum w_i y_i - \sum_i w_i d_i \sum_i w_i d_i y_i}{\sum_i w_i \sum_i w_i d_i^2 - (\sum_i w_i d_i)^2}$$

Преимущества ядерной линейной регрессии:

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