Регрессия

Виктор Китов
v.v.kitov@yandex.ru

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

I семестр 2015 г.

Усреднение в среднеквадратичном смысле

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

Условие стационарности:

$$\begin{aligned} 2\sum_{n=1}^{N}(y_i-\mu) &= 0 \\ \sum_{n=1}^{N} y_i - N\mu &= 0 \\ \mu &= \frac{1}{N}\sum_{n=1}^{N} y_i \end{aligned}$$

Усреднение по модулям отклонений

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

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 $\mu = \text{median}\{y_1, y_2, ...y_N\}$

Робастные оценки для ряда $z_1, z_2, ...z_N$:
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*

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

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

Оптимизация квадрата мат. ожидания

Мы использовали

$$\mathbb{E}\left\{\left(f(x)-\mathbb{E}[y|x]\right)\left(\mathbb{E}[y|x]-y\right)\middle|\, x\right\}=$$
$$\left(f(x)-\mathbb{E}[y|x]\right)\mathbb{E}\left\{\mathbb{E}[y|x]-y\middle|\, x\right\}\equiv 0$$

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

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

Оптимизация модуля отклонения

Теорема 2

Пусть $x, y \sim P(x, y)$. *Тогда*

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

$$\mathbb{E}\{|\mu - y| \,|\, x\} = \int_{-\infty}^{+\infty} |y - \mu| \, p(y|x) dy =$$

$$\underbrace{\int_{\mu}^{+\infty} (y - \mu) \, p(y|x) dy}_{I(\mu)} + \underbrace{\int_{-\infty}^{\mu} (\mu - y) \, p(y|x) dy}_{J(\mu)}$$

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:

$$\begin{aligned} I'(\mu) &= \int_{\mu}^{+\infty} -p(y|x)dy - (\mu-\mu)p(\mu|x) = -P(y \geq \mu|x) \\ J'(\mu) &= \int_{-\infty}^{\mu} p(y|x)dy + (\mu-\mu)p(\mu|x) = P(y \leq \mu|x) \end{aligned}$$

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 $X \in \mathbb{R}^{NxD}, \{X\}_{ij}$ где j-й признак i-го объекта, $Y \in \mathbb{R}^n$, $\{Y\}_i$ - целевое значение для i-го объекта.
- Метод наименьших квадратов:

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

Решение

Условие стационарности:

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

В векторном виде:

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

откуда

$$\widehat{\beta} = (X^TX)^{-1}X^TY$$

Это глобальный минимум, поскольку целевой критерий - выпуклый.

- Геометрическая интерпретация линейной регрессии, оцениваемой МНК.

Ограничение решения

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

Обобщение через нелинейные признаки

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

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

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

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

Типичные варианты преобразований

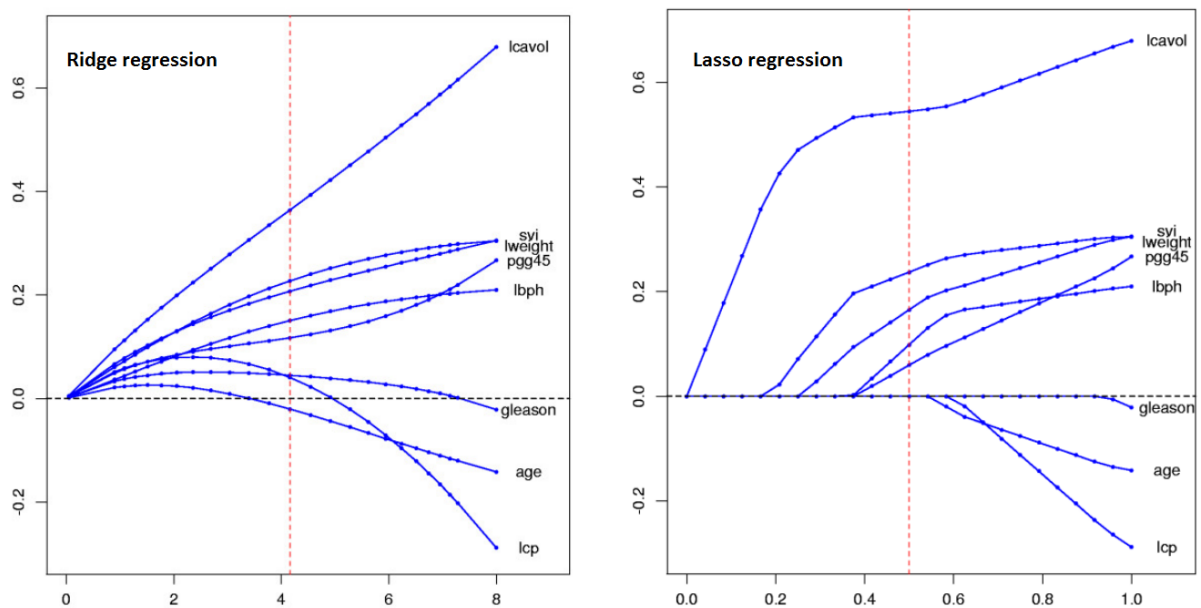
$\phi_k(x)$	**комментарии**
$\exp\left\{-\frac{\|x-\mu\|^2}{s^2}\right\}$	близость к точке признакового пространства
$x^i x^j$	взаимодействие признаков
$\ln x_k$	выравнивание распределения величины с тяжелыми хвостами
$F^{-1}(x_k)$	приведение нетипичного распределения к равномерному

Регуляризация

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

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

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



Вероятностная интерпретация

Обозначим $X = \{x_1, x_2, ...x_N\}$ - признаковые описания обучающей выборки.
Если данные описываются следующей моделью:

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

Вероятностная интерпретация метода наименьших квадратов

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

Правдоподобие обучающей выборки:

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

Максимизация логарифма правдоподобия:

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

что эквивалентно:

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

Вероятностная интерпретация метода наименьших квадратов

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

Правдоподобие обучающей выборки:

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

Максимизация логарифма правдоподобия:

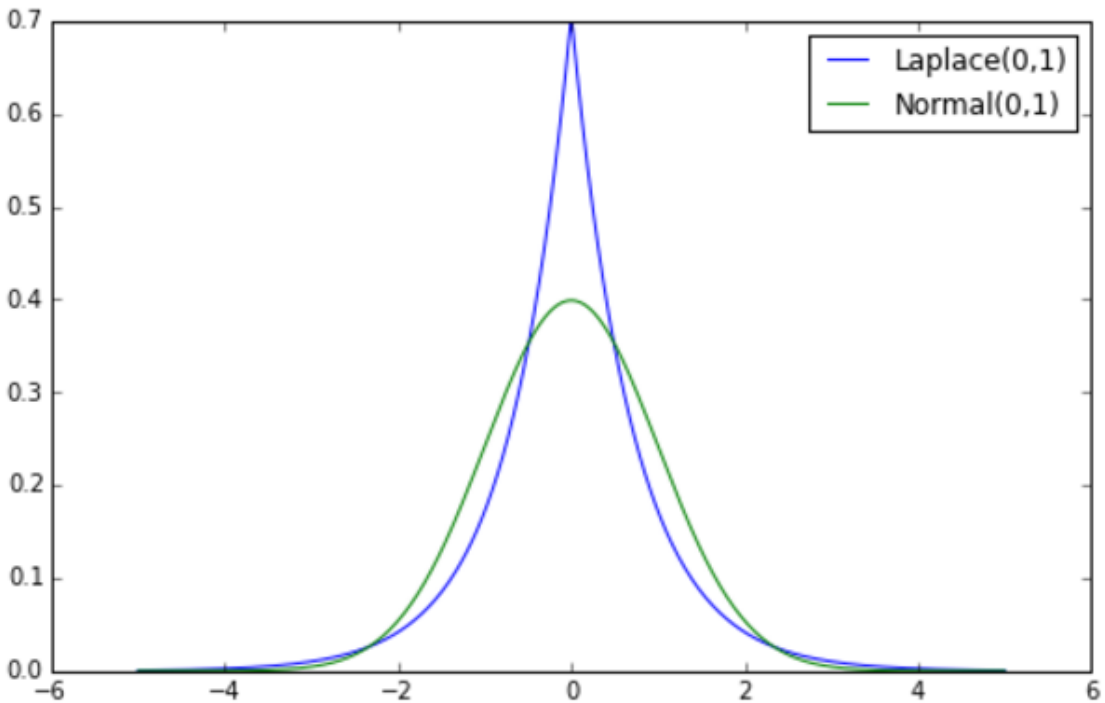
$$const - \sum_{n=1}^{N} \frac{1}{b} |f_\theta(x_i) - y_i| \to \max_\theta$$

что эквивалентно:

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

Распределение Лапласа и нормальное распределение

Laplace$(\mu, 2b^2)$	$p(\varepsilon) = \frac{1}{2b} e^{-\frac{\lvert\varepsilon-\mu\rvert}{b}}$
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:

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

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

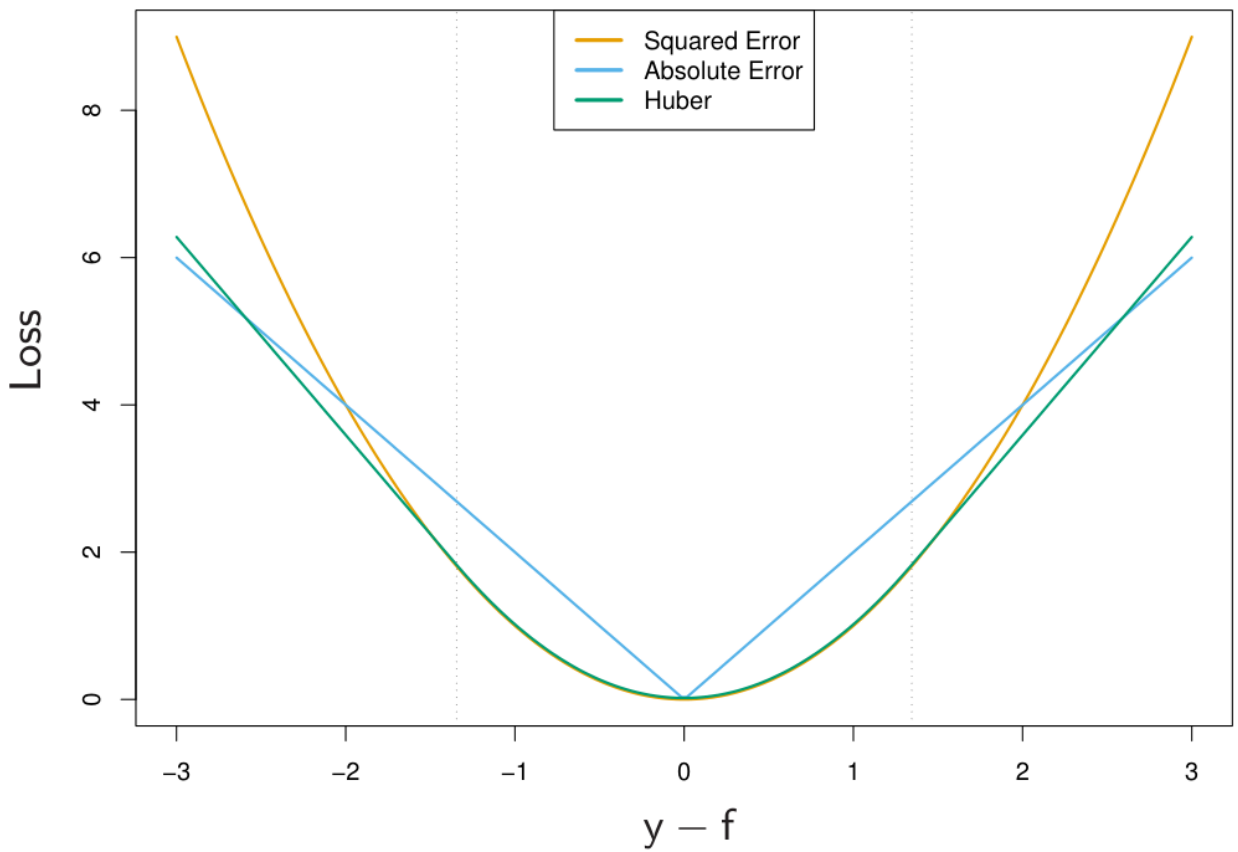
Модификации

- 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:

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

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

- 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

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

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

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 $\widehat{\alpha}(x)$:

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

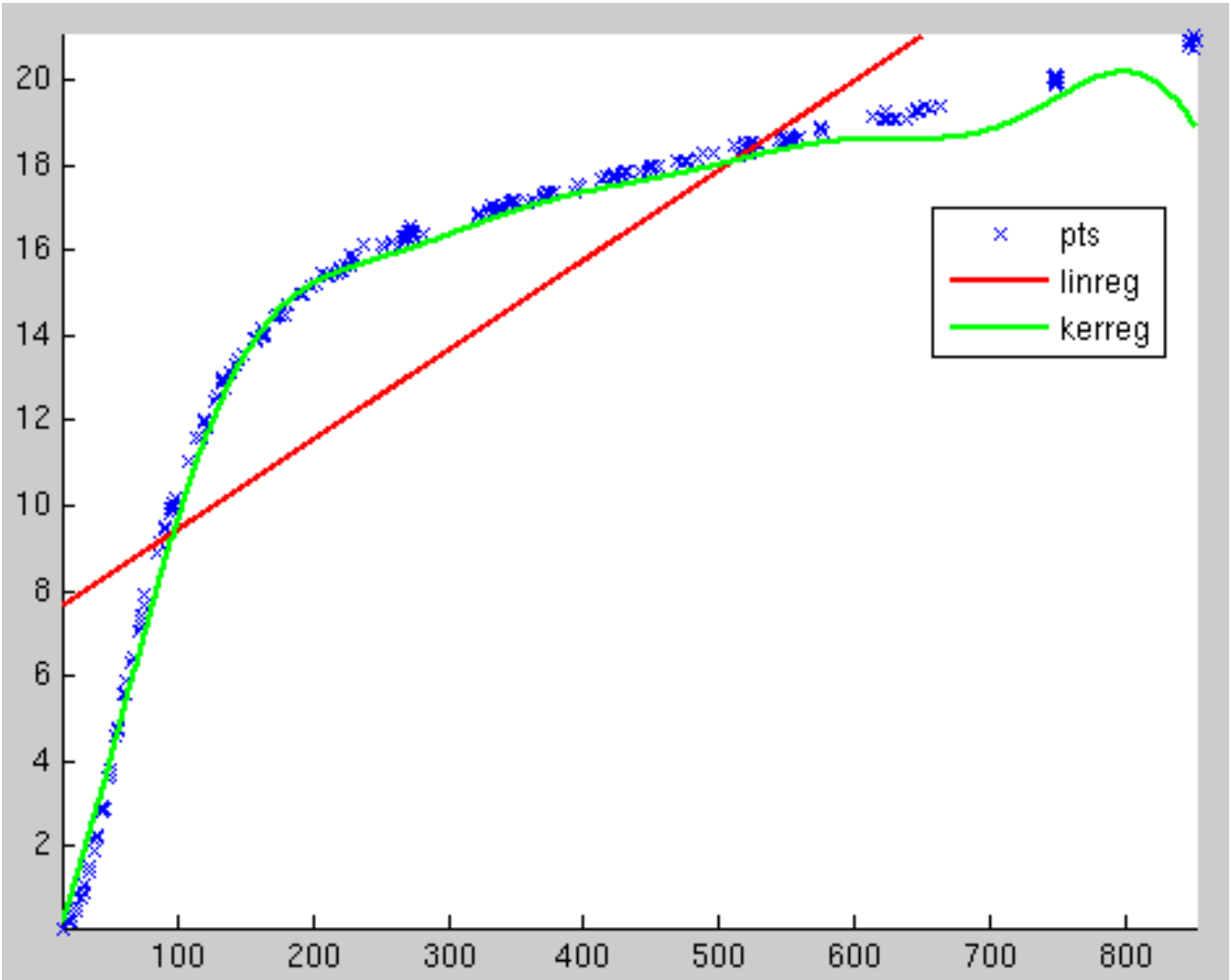
Комментарии

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

$$\begin{aligned} K_G(r) &= e^{-\frac{1}{2}r^2} - \text{гауссово ядрно} \\ K_P(r) &= (1-r^2)^2\mathbb{I}[|r|<1] - \text{квадратичное ядро} \end{aligned}$$

- 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


20
18
16
14
12
10
8
6
4
2
100
200
300
400
500
600
700
800
pts
linreg
kerreg

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 $\varepsilon_1, \varepsilon_2, ...\varepsilon_N$.
  - $D(\varepsilon_i) = K_P\left(\frac{\varepsilon_i}{6\text{med}\varepsilon_i}\right)$

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

Алгоритм

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

Ядерная линейная аппроксимация

- Локальная (в окрестности x) аппроксимация $f(u) = (u - x)^T \beta + \beta_0$
- Решить

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

- 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