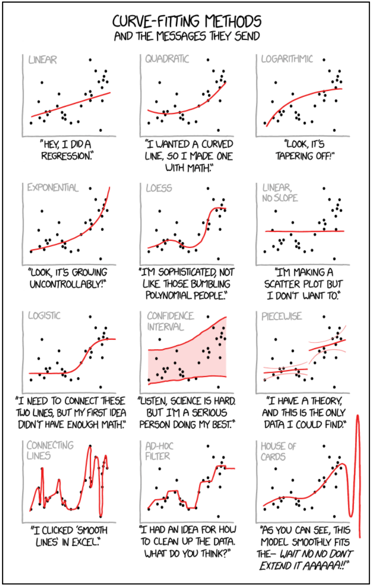
Методы восстановление регрессии

МФТИ 2020

Москва

Различные типы линейных моделей



Математическое описание

В случае когда значения прецедентов принадлежат прямой $Y = \mathbb{R}$. Задан тренеровочный набор обучающих обьектов $X^l = \{x_i, y_i\}$, $i = 1, l$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}$, так что $y_i = y^*(x_i)$.

Парметризация модели $a(x) = f(x; \alpha)$. Для ее решения, необходимо найти вектор параметров α.

Метод наименьших квадратов

Метод наименьших квадратов

$$Q(\alpha, X^l) = \sum_{i=1}^{l} (f_i(x, \alpha) - y_i)^2$$

В случае линейной модели, имеем задачу квадратичного программирования:

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}^p} Q(\alpha, X^l).$$

В общем случае квадраты признаков суммируются с весами:

$$Q(\alpha, X^l) = \sum_{i=1}^{l} w_i (f_i(x, \alpha) - y_i)^2,$$

так что можно отбрасывать лишние прецеденты, подчеркивать из значимость, а также нормировать.

Почему норма квадратичная?

Метод максимума правдоподобия

Пусть дана задача с некоррелированным Гаусовым шумом

$$y_i = f(x_i, \alpha) + \varepsilon_i, \;\; i = 1..l, \; \varepsilon_i \sim N(0, \sigma_i^2),$$

$$L(\varepsilon_1, .., \varepsilon_l | \alpha) = \prod_{i=1}^{l} \frac{1}{\sigma_i \sqrt{(2\pi)}} \exp\left(-\frac{1}{2\sigma_i^2}\varepsilon_i\right) \to \max \alpha.$$

$$-\ln L(\varepsilon_1, .., \varepsilon_l | \alpha) = \sum_{i=1}^{l} (f_i(x, \alpha) - y_i)^2 \to \min_{\alpha}.$$

Вывод: постановки метода наименьших квадратов и метод максимума правдоподобия совпадают.

Теорема Гаусса-Маркова

Пусть измерения имеют ошибки с нормальным распределением

$$y_i = f(x_i, \alpha) + \varepsilon_i,\ \ i = 1..l.$$

Теорема

1. *несмещенное среднее* $E[\varepsilon_i] = 0$, $\forall i$,
2. *с ограниченной вариацией* $Var[\varepsilon_i] < \infty$,
3. *независимые переменные* $Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$,

тогда указанная система имеет решение $\alpha = (F^T F)^{-1} F^T Y$.

Ядерное сглаживание

Рассмотрим простейшку модель $g(x, \alpha) = \alpha$, или для каждой точки пространства решим построим регрессионую модель $a(x) = g(x, \alpha)$, вычислим α для произвольного x:

$$Q(\alpha, X^l) = \sum_i w(x)(\alpha - y_i)^2 \to \min,$$

Определим веса как функции расстояния от точек в пространстве признаков $w_i(x) = K\left(\frac{\rho(x,x_i)}{h}\right)$.

$$a_h(x; X^l) = \frac{\sum y_i w_i(x)}{\sum w_i(x)} = \frac{\sum y_i K\left(\frac{\rho(x,x_i)}{h}\right)}{\sum K\left(\frac{\rho(x,x_i)}{h}\right)}.$$

Выбор параметров алгоритма

- **Выбор ядра** не влияет на точность, но связан со степенью гладкости функции $a_h(x)$,
- **ширина окна** компромисс сглаживания и точности, необходимо найти оптимальное h^*,
- **локальное сгущение** возникает, когда обьекты выборки распределены неравномерно, в этом случе рекомендуется брать $w_i(x) = K\left(\frac{\rho(x, x_i)}{h(x)}\right)$ ядра с переменной шириной.

Линейная модель регрессии

Пусть кажому обьекту соответствует его признаковое описание:

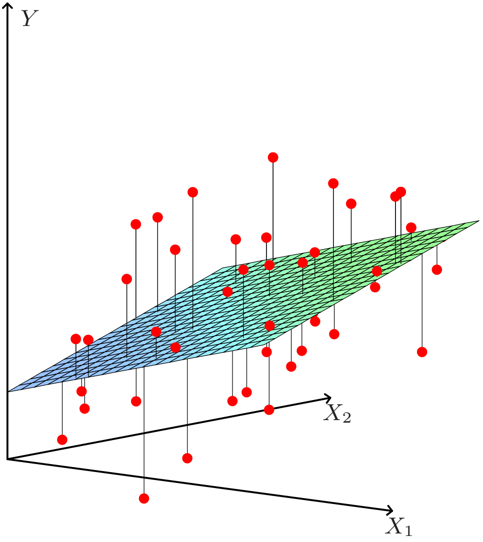
$$f_1(x), f_2(x), .., f_n(x).$$

Так что модель линейна по отношению к коэффициентам

$$f(\alpha, x) = \sum_{j=1}^{n} \alpha_j f_j(x),$$

$$Q(\alpha, X^l) = \sum_{i=1}^{l} \sum_{j=1}^{n} (\alpha_j f_j(x_i) - y_i)^2 = ||F\bar{\alpha} - y||^2.$$

Гиперплоскость в двумерном пространстве



Частное решение. Проекция решения.

Продифференцируем квадратичный функционал (**пусть матрица F полного ранга**)

$$\frac{\partial Q}{\partial \alpha} = 2F^T(F\alpha - y) = 0,$$

получаем уравнение Эйлера

$$F^T F\alpha = F^T y.$$

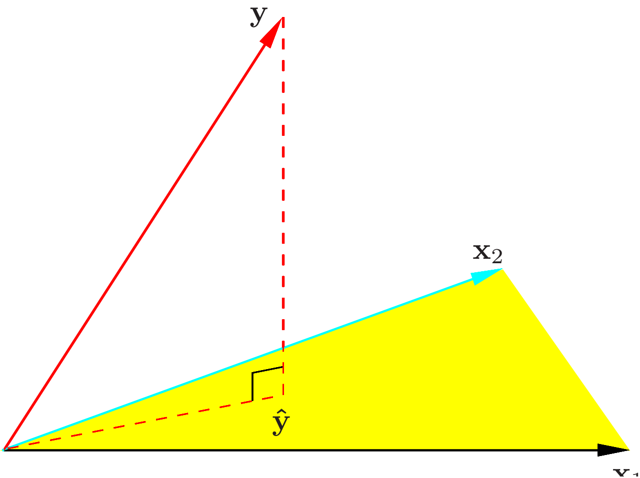
Таким образом получаем решение системы через **псевдообратную** матрицу F^+:

$$\alpha^* = (F^T F)^{-1} F y = F^+ y.$$

Как быть в случае линейно зависимости??

Иллюстрация проектирования

Метод наименьших квадратов позволяет получить проекцию на линейную оболочку столбцов матрицы F.

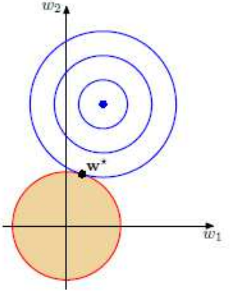


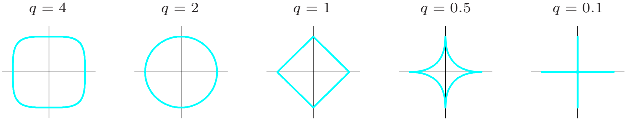
Решение в различных нормах

1. $MSE(a, X) = \frac{1}{l}\sum_i (a(x_i) - y_i)^2$ среднеквадратичное отклонение,
2. $R^2(a, X) = 1 - \frac{\sum_i (a(x_i) - y_i)^2}{\sum_i (y_i - \bar{y})^2}$, $\bar{y} = \frac{1}{l}\sum_i y_i$ - коэффициент детерминации измеряет долю дисперсии обьясненную моделью, в общей дисперсии целевой переменной.
3. $MAE(a, X) = \frac{1}{l}\sum_i (a(x_i) - y_i)^2$, среднее абсолютное отклонение.

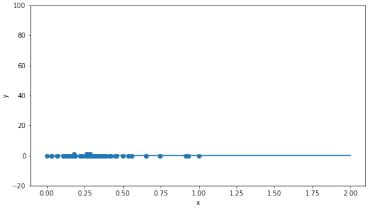
Для задачи $\sum_i (a - y_i)^2 \to \min_a$ минимум достигается на $a^*_{MSE} = \sum_i y_i$.
Для нормы MAE аналогичная задача $\sum_i |a - y_i| \to \min_a$, решение будет медиана $a^*_{MAE} = median\{y_i\}_{i=1}^l$.

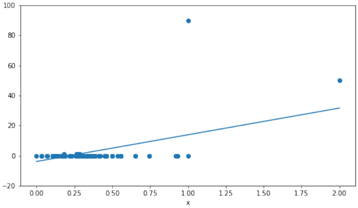
Иллюстрация различных норм



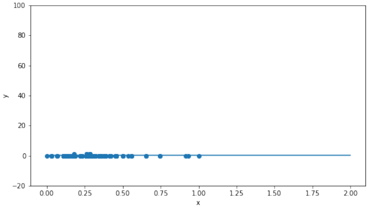


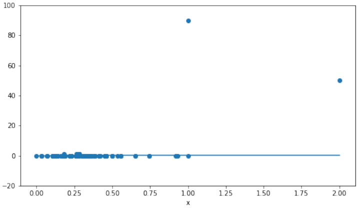
Регрессия L2, чувствительность к выбросам





Регрессия L1, чувствительность к выбросам





Число обусловленности

Количественная оценка СЛАУ с невырожденной матрицей можно связать с числом обусловленности матрицы

$$\operatorname{cond}(F) = ||F||||F^{-1}||.$$

Пусть по отношению к точной системе $Fx = y$ задана возумущенная $F_h x = y_\delta$, $||F - F_h|| \le h$, $||y - y_\delta|| \le \delta$. Возумущенная система невырождена при условии $h||F^{-1}|| < 1$, для решения возмущенной системы можно записать оценку:

$$\delta_2(x) = \frac{||x - x_\eta||}{||x||} \le \frac{\operatorname{cond}(F)(\delta_E(F) + \delta_2(y))}{1 - \delta_E(F)\operatorname{cond}_E(F)}.$$

Здесь

$$\operatorname{cond}_E(F) = ||F^{-1}||_E||F||_E$$

- евклидово число обусловленности.

Метод сингулярного разложения

Теорема

Любую матрицу F размера $m \times n$ можно представить в виде $F = VDU^T$, где – ортогональные матрицы размера $m \times m$ и $n \times n$, соответственно, а $D = diag(\rho_1, ..., \rho_M)$ - прямоугольная диагональная матрица размера $m \times n$, содержащая на диагонали неотрицательные числа $\rho_1, ..., \rho_M$, $M = \min(m, n)$, которые упорядочены по невозрастанию: $\rho_1 \geq ... \geq \rho_M \geq 0$.

Числа ρ_k называются сингулярными числами матрицы F, при этом числа ρ_k^2, являются собственными значениями матриц FF^T, столбцы U, V - собственные вектора матриц FF^T и $F^T F$. Для матриц полного ранга можно определить спектральное число обусловленности $cond_s(F) = \rho_1 \rho_M^{-1}$.

Сингулярные числа

Сингулярные числа для операторов действующих в гильбертовых пространствах. Пусть оператор F - вполне непрерывен и не является конечномерным, то он обаладает системой сингулярных чисел $\rho_1 \geq ... \geq \rho_n \geq ... \geq 0$- собственный значения операторов $F^T F$, FF^T, причем $\lim_{n\to\infty} = 0$. Обратная задача с вполне непрерывным оператором F,

1. **умеренно некорректная**, если $\rho_n \asymp n^{-\nu}$ при $n \to \infty$,
2. **сильно некорректной**, если $\rho_n \asymp e^{-n\nu}$ при $n \to \infty$.

Решение через сингулярное разложение оператора

Псевдообратная матрица F^+, вектор МНК решения α

$$F^+ = (UDV^T UDV^T)^{-1} UDV^T = UD^{-1}V^T = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j v_j^T,$$

$$\alpha^* = F^+ y = UD^{-1}V^T y = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j (v_j^T y_j),$$

$$F\alpha^* = P_F y = (VDU^T) UD^{-1}V^T y = VV^T y = \sum_{j=1}^{n} v_j (v_j^T) y,$$

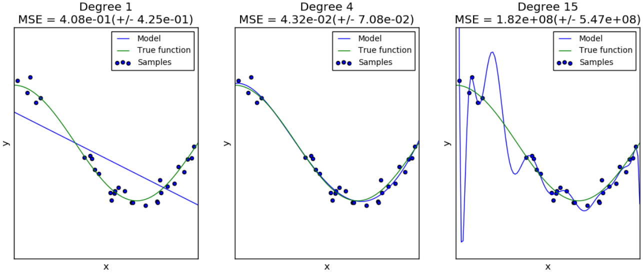
$$||\alpha^*||^2 = ||D^{-1}V^T y||^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j} (v_j^T y)^2$$

Мультиколлинеарность и переобучение

В случае когда матрица $S = F^T F$ плохо обучловленна, то

- решение неустойчиво и плохо интерпретируемо,
- $||\alpha^*||$ велико,
- переобучение
  на обучающией выборке $Q(\alpha^*, X^l) = ||F\alpha^* - y||^2$,
  на контрольной выборке $Q(\alpha^*, X^k) = ||F'\alpha^* - y'||^2$.

Переобучение



Регуляризация L2 (гребневая регрессия)

Модифицируем функционал

$$Q_\tau(\alpha) = ||F\alpha - y||^2 + \frac{1}{\sigma}||\alpha||^2,$$

где $\tau = \frac{1}{\sigma}$ - неотрицательный параметр регуляризации. Через уравнение Эйлера возможно получить следующее оптимальное решение функционала

$$\alpha^*_\tau = (F^T F + \tau I_n)^{-1} F^T y.$$

Удобно подбирать параметр через сингулярное разложение.

Решение через сингулярное разложение оператора

Псевдообратная матрица F^+, вектор МНК решения α

$$\alpha^* = U(D^2 + \tau I_n)^{-1} D V^T y = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} u_j (v_j^T y_j),$$

$$F\alpha_\tau^* = V \operatorname{diag}\left(\frac{\lambda_j}{\lambda_j + \tau}\right) V^T y = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} v_j (v_j^T) y,$$

$$||\alpha^*||^2 = ||(D^2 + \tau I_n)^{-1} D V^T y||^2 = \sum_{j=1}^{n} \frac{\lambda_j}{(\lambda_j + \tau)^2} (v_j^T y)^2.$$

Данная процедура обеспечивает устойчивость решения.

Эффективная размерность

Сокращение весов

$$||\alpha^*||^2 = \sum_{j=1}^{n} \frac{\lambda_j}{(\lambda_j + \tau)^2} (v_j^T y)^2.$$

Регуляризация сокращает эффективную размерность

$$\operatorname{tr} F^T (F^T F)^{-1} F = \operatorname{tr} (F^T F)^{-1} F^T F = \operatorname{tr} I_n = n,$$

При регуляризации

$$\operatorname{tr} F (F^T F + \tau I_n)^{-1} F^T = \operatorname{tr} \operatorname{diag} \left(\frac{\lambda_j}{\lambda_j + \tau} \right) = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} < n.$$

Регуляризация LASSO (L1)

Функционал с $L1$ сглаживанием

$$||F\alpha - y||^2 + \mu \sum_{j=1}^{n} |\alpha_j| \to \min_{\alpha},$$

данная постановка эквивалентна

$$\begin{cases} ||F\alpha - y||^2 \to \min\limits_{\alpha}, \\ \sum\limits_{j=1}^{n} |\alpha_j| \le C; \end{cases}$$

Произведем замену переменных $\alpha_j = \alpha_j^+ - \alpha_j^-$, так что $|\alpha_j| = \alpha_j^+ + \alpha_j^-$, $\alpha_j^+ \ge 0$, $\alpha_j^- \ge 0$.

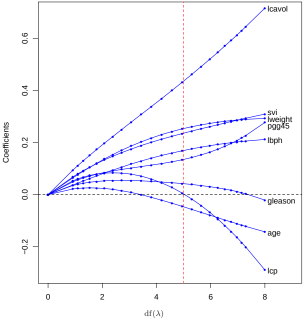
$$\sum_{j=1}^{n} (\alpha_j^+ + \alpha_j^-) \le C,$$

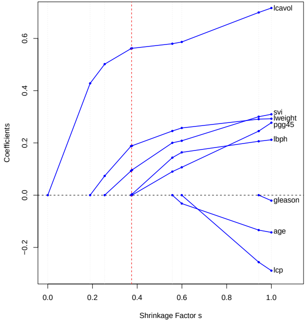
При уменьшении константы регуляризации, становится больше признаков таких что $\alpha_j^+ = \alpha_j^- = 0$.

Разряженные модели (L1 регуляризация)

- проводит отбор признаков не относящихся к задаче,
- для ускорения модель зависет от небольшого количества наиболее важных признаков,
- обьектов существенно меньше, чем признаков $N << p$, можно задать ограничение, что целевая переменная зависит от небольшого количества признаков.

Сравнение применения L1 и L2 регуляризации





Метод главных компонент

Пусть заданы матрица старых признаков $F \in \mathbb{R}^{l \times n}$ и новых $G \in \mathbb{R}^{l \times m}$ с меньших числом столбцов. При этом матрица линейных преобразовний признаков $U \in \mathbb{R}^{n \times m}$, так что

$$\hat{F} = GU^T.$$

Необходимо найти новые признаки G и матрицу преобразования U

$$\sum_{i=1}^{l} \sum_{j=1}^{n} (\hat{f}_i(x_j) - f_j(x_i))^2 = ||GU^T - F|| \to \min_{G,U}$$

Метод главных компонент, теорема

Теорема

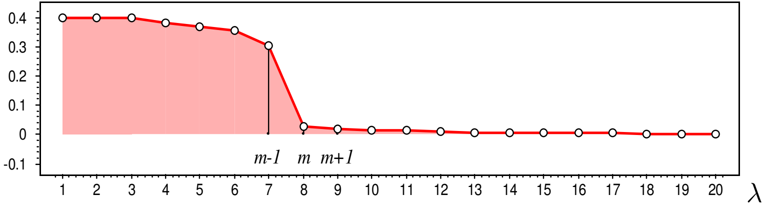
Если $m \leq rkF$, то минимум $||GU^T - F||^2$ достигается, когда столбцы U - это с.в. матрицы F^TF, соответствующие m максимальным с.з. $\lambda_1, .., \lambda_m$, а матрица $G = FU$. При этом

1. *матрица U ортонормирована $U^TU = I_m$;*
2. *матрица G ортогональна $G^TG = \Lambda = diag(\lambda_1, .., \lambda_m)$;*
3. *$U\Lambda = F^TFU$; $G\Lambda = FF^TG$;*
4. *$||GU^T - F||^2 = ||F||^2 - tr\Lambda = \sum\limits_{j=m+1}^{n} \lambda_j$.*

Эффективная размерность выборки

Эффективная размерность выборки - это наименьшее число m при котором

$$E_m = \frac{||GU^T - F||^2}{||F||^2} = \frac{\lambda_{m+1} + .. + \lambda_n}{\lambda_1 + ... + \lambda_n} \leq \varepsilon.$$



Заключение

- метод наименьших квадратов
- многомерная линейная регрессия
- боремся с мультиколлинеарностью и переобучением
- различные нормы сглаживания
- негладкая регуляризация