Введение в анализ данных (машинное обучение)

Очень краткое введение

Анализ данных

- Область математики и информатики, занимающаяся построением и исследованием наиболее общих математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных (в широком смысле) данных; процесс исследования, преобразования и моделирования данных с целью извлечения полезной информации и принятия решений.
- Анализ данных имеет множество аспектов и подходов, охватывает разные методы в различных областях науки и деятельности.
- Синоним **DATA MINING**.

Основные задачи анализа данных

- Классификация (распознавание образов)

По экспериментальным данным (X1,Y1,…,Xn,Yn) требуется предсказать значение Y для нового X.

Xi – вектор признаков, Yi – признак (метка) класса, например, 0 или 1 при распознавании двух классов

Основные задачи анализа данных

- Восстановление зависимостей (построение регрессионных моделей)

По экспериментальным данным (X1,Y1,…,Xn,Yn) требуется предсказать значение Y для нового X.

Xi – вектор факторов (ковариат, предикторов, факторов), Yi – значение зависимости – непрерывная величина

Формализация основных задач анализа данных

- Классификация

Построить функцию

g(X,X1,Y1,…,Xn,Yn)

такую, чтобы ошибка классификации на новых данных была минимальна.

Если распределение *P(X,Y)* известно, то решением будет

$$g(X) = \arg\max_k P(Y = k \mid X)$$

$$= \arg\max_k \frac{P(X \mid Y = k)P(Y = k)}{P(X \mid Y = 0)P(Y = 0) + P(X \mid Y = 1)P(Y = 1)}$$

Формализация основных задач анализа данных

- Восстановление зависимостей (построение регрессионных моделей)

Построить функцию

g(X,X1,Y1,…,Xn,Yn)

такую, чтобы ошибка предсказания на новых данных была минимальна.

Если распределение *P(X,Y)* известно, то при квадратичной функции потерь решением будет функция условного математического ожидания (регрессии)

$$g(X) = E(Y \mid X)$$

Если *P(X,Y)* неизвестна

- по выборке *X1,Y1,…,Xn,Yn* оценить *P(X,Y)*
- по выборке *X1,Y1,…,Xn,Yn* оценить ошибку классификации (ошибку предсказания) при использовании функции *g(X,X1,Y1,…,Xn,Yn)* и минимизировать эту оценку (минимизация эмпирического риска)

Мы будем обсуждать только задачу восстановления зависимостей

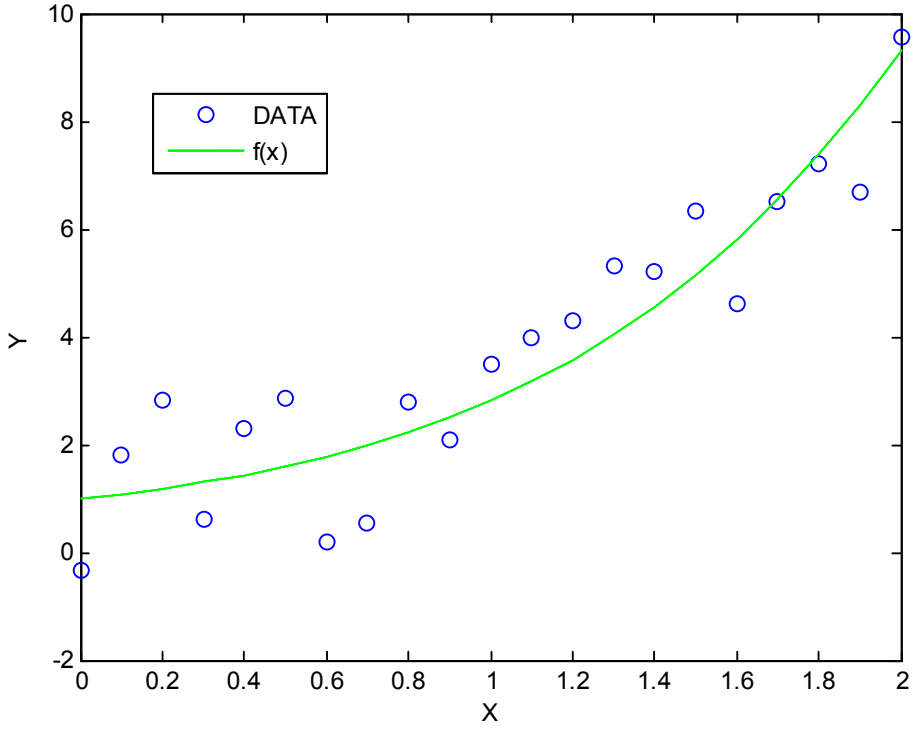
Часть 1

Практика

1. Практическое рассмотрение задачи. МНК

- Имеются искажённые шумом данные

$$f(x) = 1 + 0.8x + 0.6x^2 + 0.3x^3 + 0.1x^4 + 0.01x^5$$



Как построить прямую, наилучшим образом приближающую точки?

- Обозначим матрицу $F = [1, x]$ (21x2), вектор из двух $a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ параметров , вектор y.

Значения прямой в экспериментальных точках равны вектору *Fa*,

СКО прогноза вектора y вектором *Fa* равна

$$J(a) = \|y - Fa\|^2 = \sum_{i=1}^{21} \left(y_i - \sum_{j=1}^{2} F_{ij} a_j \right)^2$$

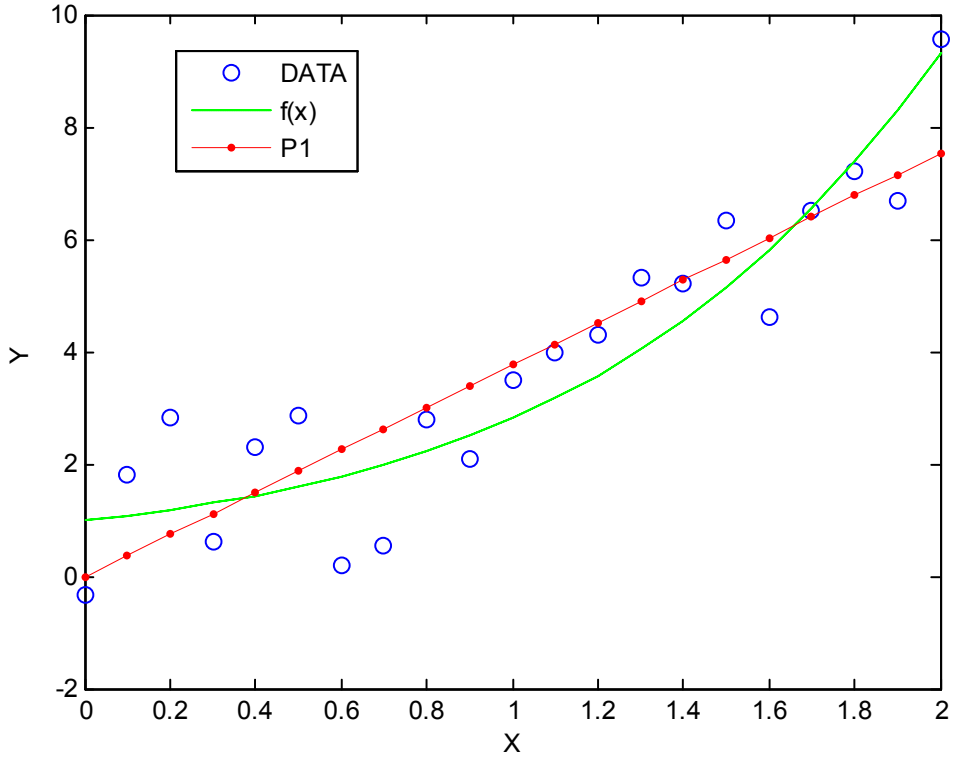
Найдём вектор $\hat{a} = \arg\min_a \|y - Fa\|^2 \qquad \hat{a} = \left(F^T F\right)^{-1} F^T y$

$$\begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = \begin{pmatrix} 21 & \sum_{i=1}^{21} x_i \\ \sum_{i=1}^{21} x_i & \sum_{i=1}^{21} x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{21} y_i \\ \sum_{i=1}^{21} x_i y_i \end{pmatrix}$$

Линейная регрессия

- Подставив данные получим

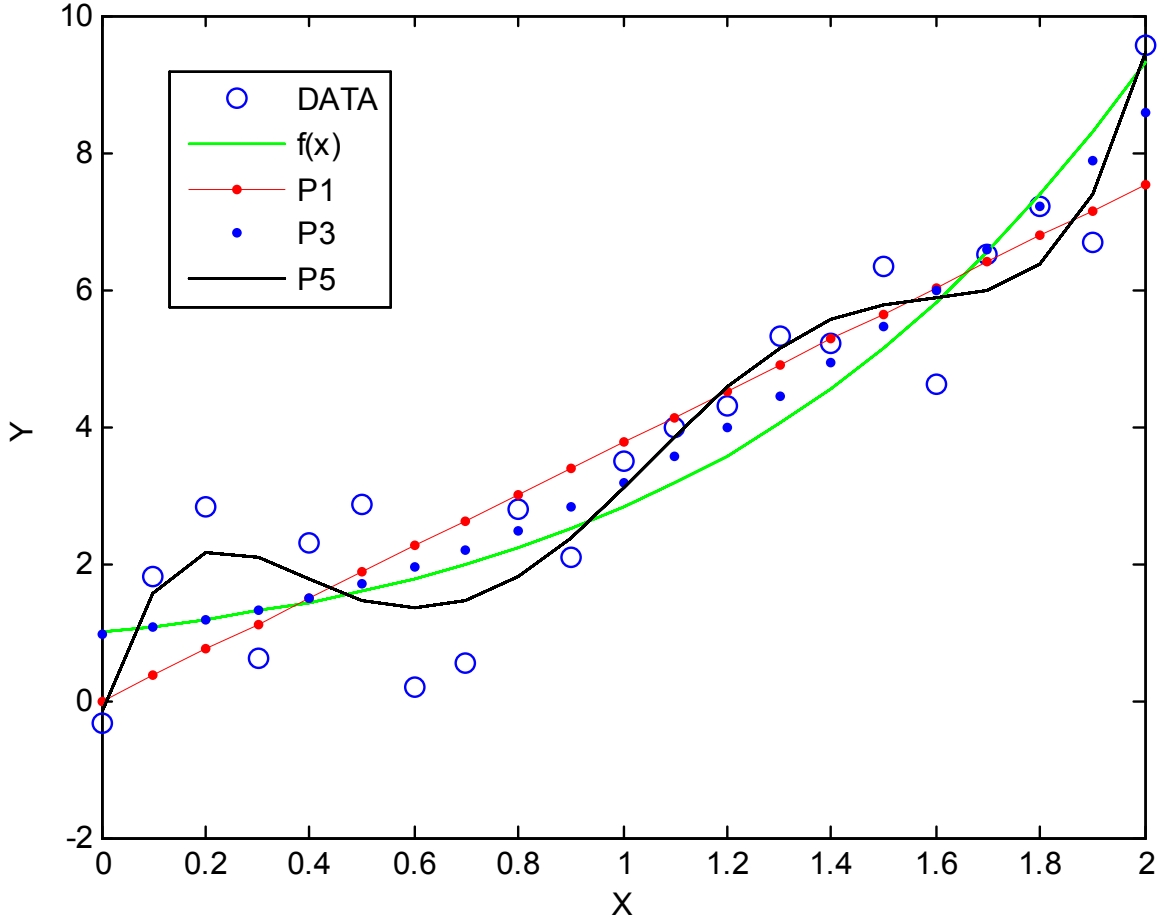
$$\hat{a}_1 = -0.013842 \qquad \hat{a}_2 = 3.7717$$



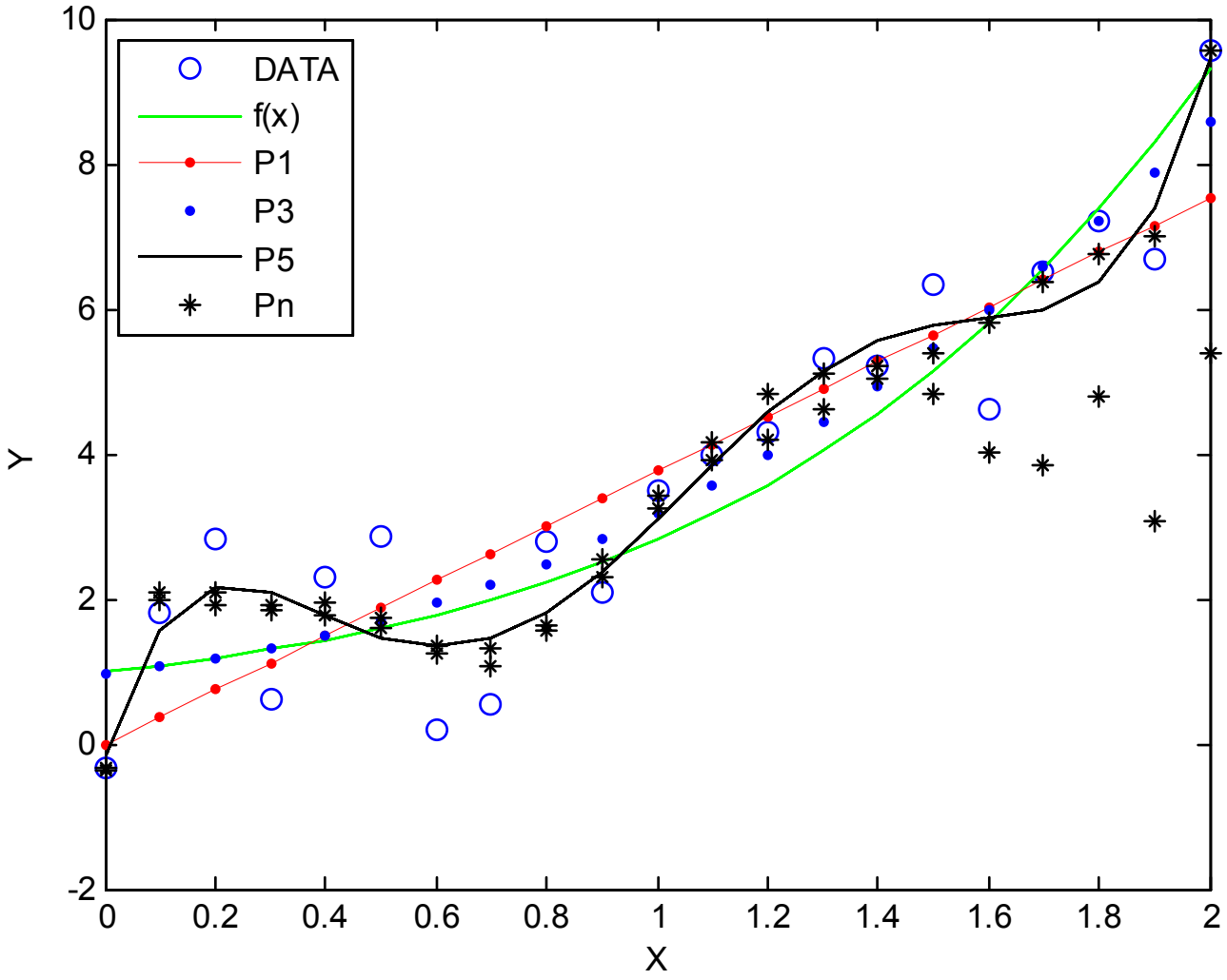
Полином пятой степени.

Вывод 1. При анализе данных более простые модели часто дают лучший результат, чем «истинные модели».

$$\hat{a} = \begin{pmatrix} -0.15212 & 25.063 & -89.766 & 123.47 & -69.222 & 13.698 \end{pmatrix}$$



А если увеличивать степень полинома дальше?



Чем больше параметров, тем хуже. Почему?

Матрица $F^T F$ имеет близкие к 0 собственные числа

Собственные числа для полинома степени 14

-1.1403e-011,	4.6575e-011,	6.3565e-010,
1.6026e-008,	7.2452e-007,	2.2671e-005,
0.00051246,	0.0086921,	0.11269,
1.1052,	7.8416,	36.089,
1191.9,	2.7805e+005,	4.7163e+008

Чем больше параметров, тем хуже. Что делать?

Сдвинем ось x на 1 влево. Тогда x=x-1

Собственные числа матрицы $F^T F$ отодвинулись от 0

Собственные числа для полинома степени 14

9.0518e-010,	7.5075e-009,	2.4408e-007,
1.6687e-006,	2.4992e-005,	0.00014372,
0.0013266,	0.0064589,	0.041189,
0.16901,	0.78581,	2.6676,
9.0181,	21.045,	31.29

2. Практическое рассмотрение задачи. Гребневая регрессия

А как ещё можно отодвинуть от нуля собственные числа матрицы?

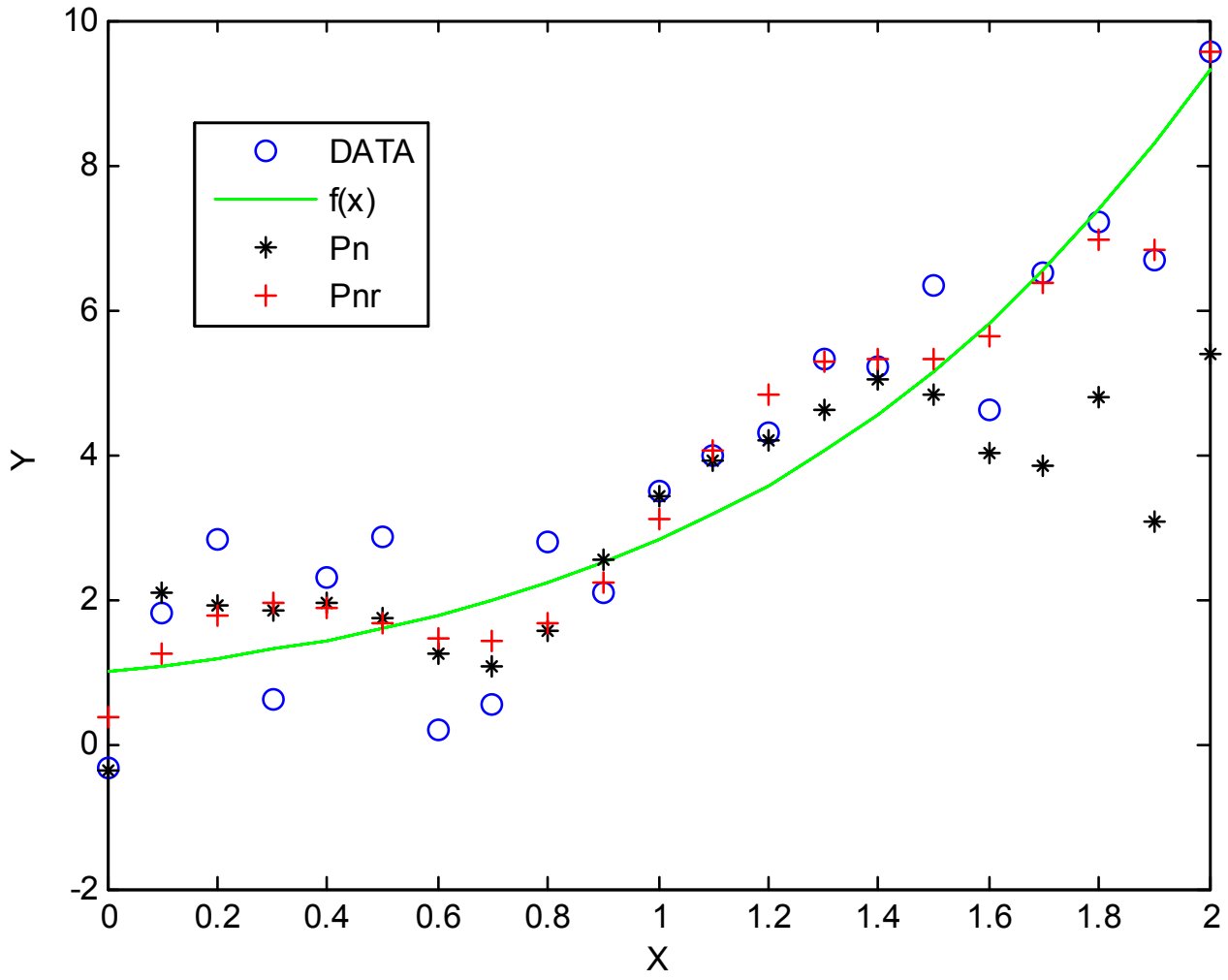
Добавим на диагональ матрицы $F^T F$ число $\lambda = 0.001$

и найдём коэффициенты полинома по формуле

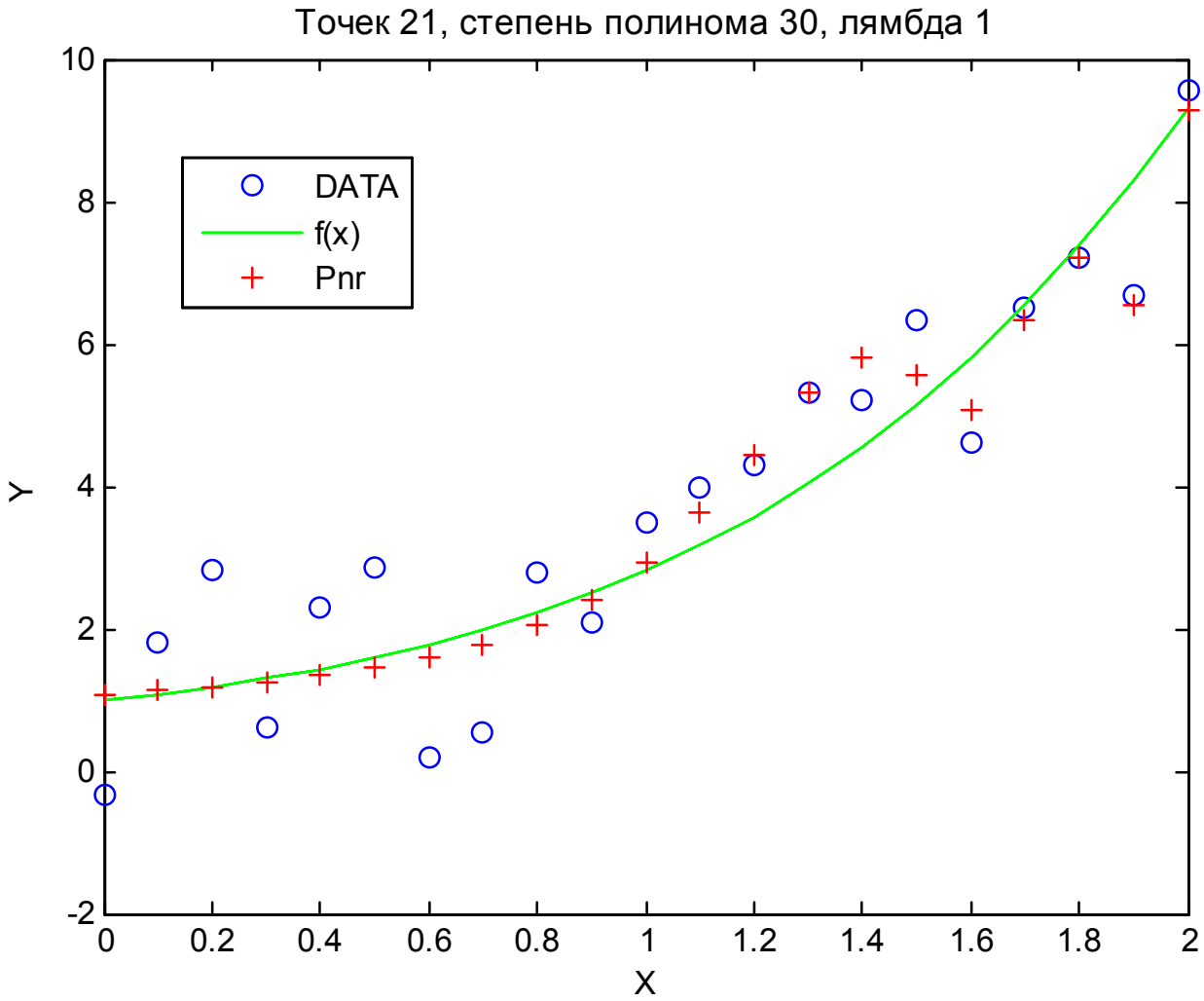
$$\hat{a}_\lambda = \left(F^T F + \lambda I\right)^{-1} F^T y$$

Работает даже при вырожденной матрице ,
когда число наблюдений меньше числа параметров!

Гребневая регрессия



Гребневая регрессия. Случай вырожденной матрицы

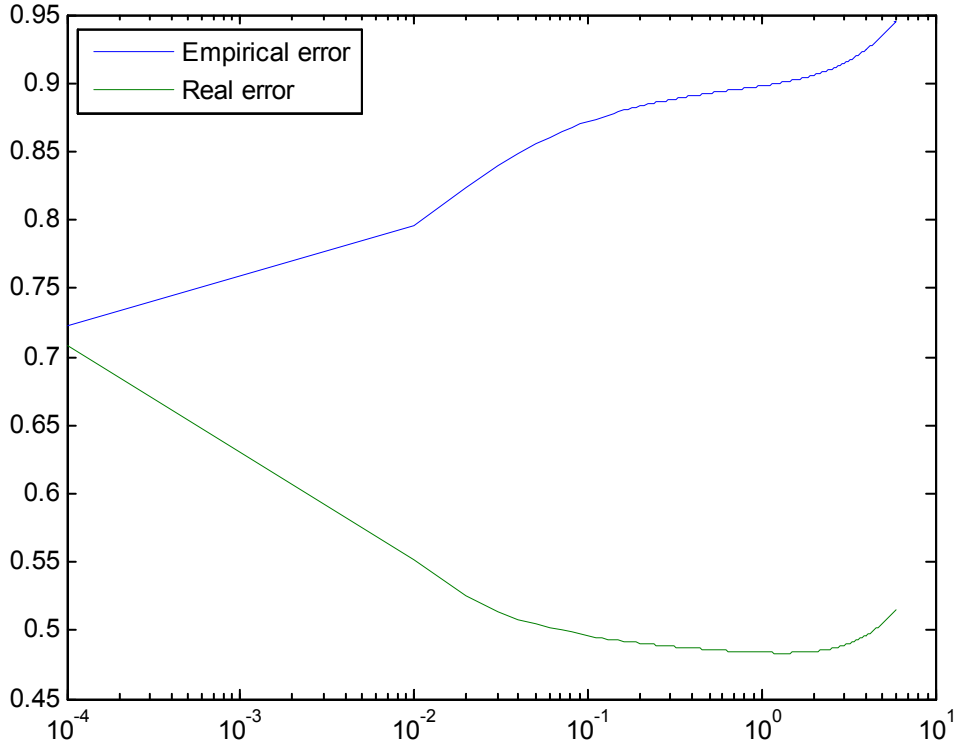


Какую величину параметра λ использовать на практике?

Поведение ошибки

Ошибка на эмпирических данных $\sqrt{\frac{1}{21}}\left\|y - F\hat{a}_\lambda\right\|$

Реальная ошибка $\sqrt{\frac{1}{100}\sum_{i=1}^{100}\left(f\left(x_i\right) - P_\lambda\left(x_i\right)\right)^2}$



Часть 2

Теория

Каковы теоретические основания метода наименьших квадратов?

Модель

$$y = Fa + \varepsilon$$

$$\varepsilon \sim N\left(0, \sigma^2 I\right)$$

Метод максимального правдоподобия

$$L(y,a)=\prod_{i=1}^{n}L(z_i,a)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{z_i^2}{2\sigma^2}\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(y_i-\sum_{j=1}^{m}F_{ij}a_j\right)^2}{2\sigma^2}\right)$$

$$=\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{m}F_{ij}a_j\right)^2}{2\sigma^2}\right)=\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{\|y-Fa\|^2}{2\sigma^2}\right)$$

$$z=y-Fa$$

Условие максимума правдоподобия при нормально распределённой помехе

$$\|y - Fa\|^2 \underset{a}{\rightarrow} \min$$

Вывод 2. Для линейной модели наблюдений при аддитивной нормально распределённой помехе метод максимального правдоподобия совпадает с методом наименьших квадратов.

4. Максимум правдоподобия для нормального распределения. Общий случай

Откажемся от независимости помехи, оставив нормальность.

$\varepsilon \sim N(0, K)$ K - матрица ковариации

Правдоподобие

$$L(y,a) = \frac{1}{(2\pi)^{n/2}\sqrt{\det K}} \exp\left(-\frac{1}{2}(y-Fa)^T K^{-1}(y-Fa)\right)$$

Условие максимума

$$(y-Fa)^T K^{-1}(y-Fa) \underset{a}{\rightarrow} \min$$

Частный случай

$$K = diag(\sigma_i^2)$$

$$K^{-1} = diag(\sigma_i^{-2})$$

Метод взвешенных наименьших квадратов

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} F_{ij} a_j \right)^2 / \sigma_i^2 \underset{a}{\rightarrow} \min$$

5. Байесовский подход

Модель

$$y = Fa + \varepsilon$$

$$\varepsilon \sim N\left(0, \sigma^2 I\right)$$

$$a \sim N\left(a_0, K_a\right)$$

Совместное распределение векторов *y* и *a*

$$p(y,a) = p(y \mid a)p(a)$$

$$= \frac{1}{(2\pi)^{n/2}\sqrt{\det K_\varepsilon}(2\pi)^{m/2}\sqrt{\det K_a}} \exp\left(-\frac{1}{2}\left[(y-Fa)^T K_\varepsilon^{-1}(y-Fa) + (a-a_0)^T K_a^{-1}(a-a_0)\right]\right)$$

Апостериорное распределение вектора *a*

$$p(a \mid y) = \frac{p(y,a)}{p(y)} = \frac{p(y \mid a)p(a)}{\int p(y,a)da}$$

Наилучшей оценкой для вектора *a* будет апостериорное математическое ожидание

$$\int \|a-b\|^2 p(a \mid y)da \underset{b}{\rightarrow} \min$$

Апостериорное математическое ожидание

$$b_y = E(a \mid y) = \int a p(a \mid y) da$$

$$p(a \mid y) = \frac{p(y,a)}{p(y)}$$

Совместное распределение векторов *y* и *a* равно

$$p(y,a)=\frac{1}{(2\pi)^{n/2}\sqrt{\det K_\varepsilon}}\exp\left(-\frac{1}{2}(y-Fa)^T K_\varepsilon^{-1}(y-Fa)\right)\frac{1}{(2\pi)^{m/2}\sqrt{\det K_a}}\exp\left(-\frac{1}{2}(a-a_0)^T K_a^{-1}(a-a_0)\right)$$

$$=\frac{1}{(2\pi)^{(n+m)/2}\sqrt{\det K_\varepsilon}\sqrt{\det K_a}}\exp\left(-\frac{1}{2}\left[(y-Fa)^T K_\varepsilon^{-1}(y-Fa)+(a-a_0)^T K_a^{-1}(a-a_0)\right]\right)$$

Преобразуем квадратичную форму, стоящую в показателе экспоненты

$$(y-Fa)^T K_\varepsilon^{-1}(y-Fa)+(a-a_0)^T K_a^{-1}(a-a_0)=(a-b)^T B(a-b)$$

Получим снова многомерное нормальное распределение со математическим ожиданием *b*

Чтобы найти вектор *b* вычислим положение экстремума квадратичной формы по *a*

$$\underset{a}{grad}\left[(y-Fa)^T K_\varepsilon^{-1}(y-Fa)+(a-a_0)^T K_a^{-1}(a-a_0)\right]$$

$$=-2F^T K_\varepsilon^{-1} y+2F^T K_\varepsilon^{-1} Fa-2K_a^{-1} a_0+2K_a^{-1} a$$

$$\boxed{\left(F^T K_\varepsilon^{-1} F+K_a^{-1}\right)a=F^T K_\varepsilon^{-1} y+K_a^{-1} a_0}$$

Отсюда апостериорное математическое ожидание *b* равно

$$b=\left(F^T K_\varepsilon^{-1} F+K_a^{-1}\right)^{-1}\left(F^T K_\varepsilon^{-1} y+K_a^{-1} a_0\right)$$

Частный случай

$$a_0 = 0 \qquad K_a = \sigma_a^2 I \qquad K_\varepsilon = \sigma_\varepsilon^2 I$$

$$E(a \mid y) = \left(\frac{1}{\sigma_\varepsilon^2} F^T F + \frac{1}{\sigma_a^2} I \right)^{-1} \frac{1}{\sigma_\varepsilon^2} F^T y = \left(F^T F + \frac{\sigma_\varepsilon^2}{\sigma_a^2} I \right)^{-1} F^T y$$

Получили гребневую регрессию. На диагональ матрицы нормальной системы добавляем отношение дисперсии шума к дисперсии сигнала.

Получили правило выбора параметра регуляризации.

Более того, небольшой шум стабилизирует решение!

6. Метод "скользящего контроля" (cross-validation)

Зафиксируем параметр регуляризации. Удалим из вектора *y* i-ый элемент , а из матрицы *F* соответствующую ему строку . Коэффициенты гребневой регрессии, построенной по полученной выборке, обозначим a^i.

Величина S* характеризует качество работы алгоритма при фиксированном параметре регуляризации.

$$S^* = \frac{1}{l}\sum_{i=1}^{l}\left(y_i - f_i^T a^i\right)^2$$

Справедлива формула

$$S^* = \frac{1}{l}\sum_{j=1}^{l}\left(\frac{y_j - t_j}{1 - h_j}\right)^2$$

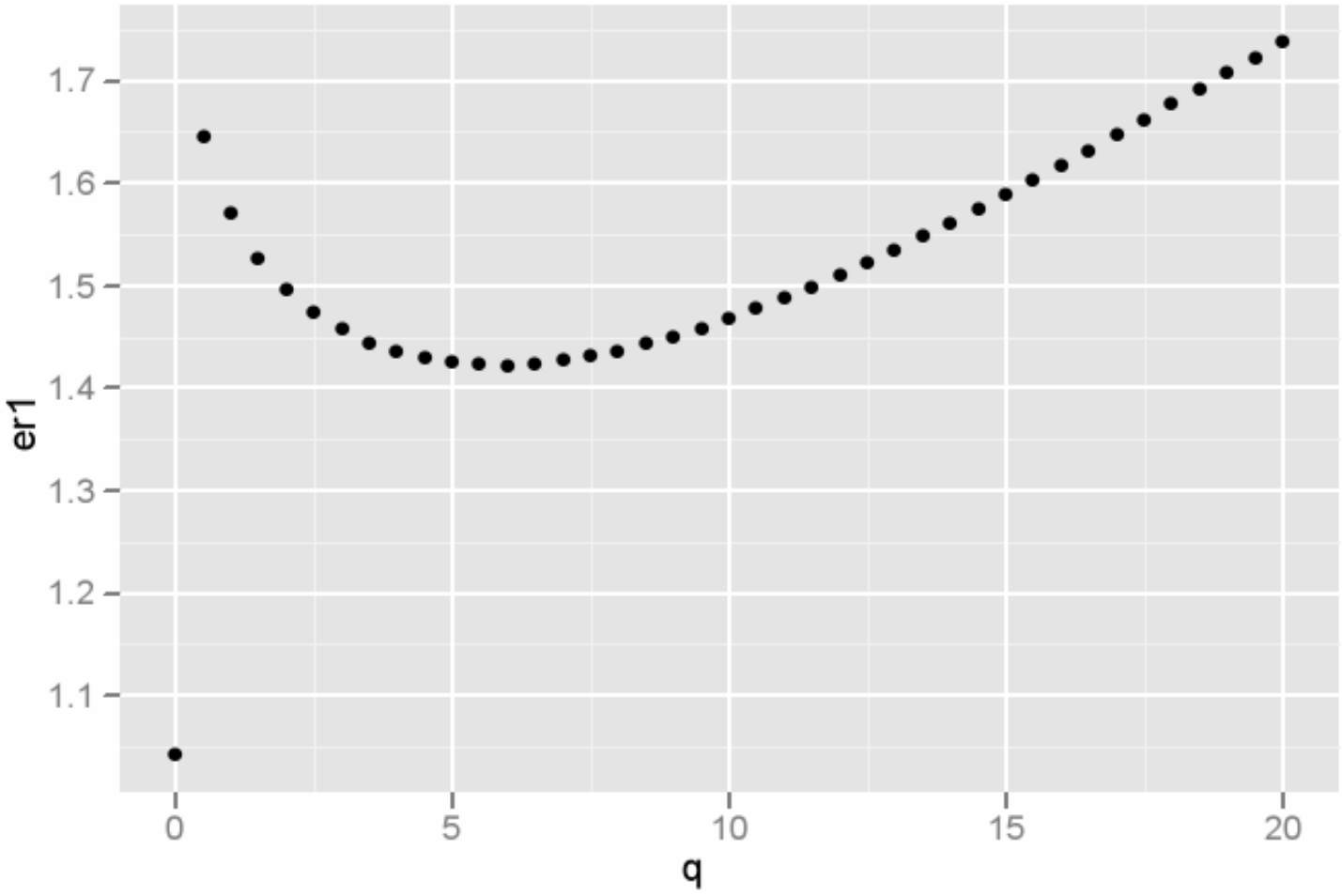
где

$$t = F\left(F^T F + \lambda I\right)^{-1} F^T y$$

h - диагональ матрицы

$$H = F\left(F^T F + \lambda I\right)^{-1} F^T$$

Подберём методом cross-validation параметр гребневой регрессии в нашей задаче построения полиномиальной регрессии пятой степени



Чёрная кривая – истинная зависимость, красная кривая – полином пятой степени, построенный при оптимальном (5.8) значении параметра регуляризации