Phystech@DataScience

Основы статистики и машинного обучения

План занятия

- Введение в машинное обучение: задача регрессии
  - Линейная регрессия
  - Метрики качества для задачи регрессии
  - Регуляризация

Линейная регрессия

Метод наименьших квадратов

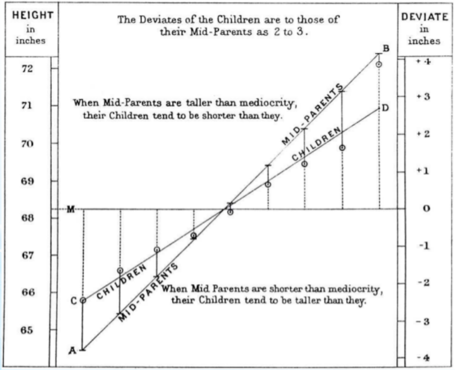
Первое упоминание регрессии

Впервые регрессия упоминается в работе Гальтона

"*Регрессия к середине в наследственности роста*", 1885 г.

x — рост родителей, y — рост детей

Установлена зависимость $y - \bar{y} \approx \frac{2}{3}(x - \bar{x})$, т.е. регрессия к середине.



Задача регрессии: интуиция

Есть объект, обладающий признаками x.

Примеры признаков: рост песика, экспрессия белка, энергия частицы.

Мы предполагаем, что есть зависимость какой-то численной характеристики объекта y от его признаков:

$$y \approx f(x)$$

Пример: зависимость уровня когнитивных способностей от параметров поражения мозга при рассеянном склерозе.

Однако мы не знаем, какова эта зависимость на самом деле.

На основании данных – набора объектов, для которых известны x и y, мы пытаемся "восстановить" зависимость:

$$y \approx \widehat{f}(x)$$

Пример

Пусть x — рост песика, а y — его вес.

Что мы знаем?

- чем крупнее песик, тем больший вес он имеет;
- песики одинакового роста могут иметь разный вес.

Выводы:

- для фиксированного роста песика x
  его вес $y = f(x)$ является случайной величиной;
- в среднем вес $f(x)$ возрастает при увеличении роста песика x.

Пример

Простая зависимость:

$$y = \theta_0 + \theta_1 x + \varepsilon,$$

x — рост песика,

y — вес песика,

θ_0, θ_1 — неизвестные параметры,

ε — случайная составляющая с нулевым средним (погрешность).

Зависимость **линейна по параметрам**, линейна по аргументу.

Пример

Более сложная зависимость:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_2^2 + \varepsilon,$$

x_1 — рост песика,

x_2 — обхват туловища песика,

y — вес песика,

$\theta_0, \theta_1, \theta_2, \theta_3$ — неизвестные параметры,

ε — случайная составляющая с нулевым средним.

Зависимость **линейна по параметрам**, квадратична по аргументам.

Модель линейной регрессии

Рассматриваем функциональную зависимость вида

$$y = y(x) = \theta_1 x_1 + \ldots + \theta_d x_d$$

$x_1, \ldots, x_d$ — признаки ,

$\theta = (\theta_1, \ldots, \theta_d)^T$ — вектор параметров.

Для оценки θ производится n испытаний вида

$$Y_i = \theta_1 x_{i1} + \ldots + \theta_d x_{id} + \varepsilon_i, \quad i = 1, \ldots, n,$$

$x_i = (x_{i1}, \ldots, x_{id})$ — признаковые описания объекта i

(обычно неслучайные),

ε_i — случайная ошибка измерений.

Модель линейной регрессии

Введем обозначения

$$Y = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \dots & & \\ x_{n1} & \dots & x_{nd} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}.$$

Матричная форма записи проведенных испытаний

$$Y = X\theta + \varepsilon.$$

$X \in \mathbb{R}^{n \times d}$ — регрессоры (или матрица плана эксперимента),

$Y \in \mathbb{R}^n$ — отклик.

Матричный вид зависимости: $y(x) = x^T\theta$.

Замечание

Зависимость $y = y(x)$ должна быть **линейна по параметрам**, но не обязана быть линейной по признакам.

Пусть $z_1, ..., z_k$ — набор "независимых" переменных.

Можно рассматривать модель

$$y(x) = \theta_1 x_1(z_1, ..., z_k) + ... + \theta_d x_d(z_1, ..., z_k),$$

где $x_j(z_1, ..., z_k)$ — некоторые функции (м.б. нелинейные).

Примеры:

- $x(z_1, ..., z_k) = 1;$
- $x(z_1, ..., z_k) = z_1;$
- $x(z_1, ..., z_k) = \ln z_1;$
- $x(z_1, ..., z_k) = z_1^2 z_2.$

Категориальные переменные

x — id объекта (натуральное число),

y — его масса.

Предположим, что должности занумерованы следующим образом:

- $x = 1$ — черная дыра;
- $x = 2$ — нейтронная звезда;
- $x = 3$ — обычная звезда.

Если $x \in \{1, ..., k\}$, то рассматриваются **dummy-переменные**:

$x_j = I\{x = j\}, \quad j = 1, ..., k-1,$

модель $y = \theta_0 + \theta_1 x_1 + ... + \theta_{k-1} x_{k-1}$.

Трифилярный подвес

- На платформу помещается тело — диск, разрезанный по диаметру;
  - I — момент инерции тела;
  - m — масса тела;
  - h — расстояние от половинок до оси вращения;
  - I_0 — момент инерции нераздвинутого диска.
- Половинки диска постепенно раздвигаются;
- Снимается зависимость момента инерции системы I от h.

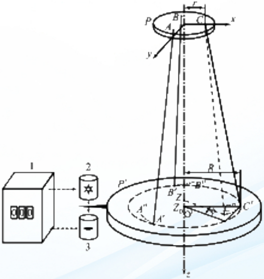


Рис. 2. Трифилярный подвес

По материалам "Модели и концепции физики: механика. Лабораторный практикум"

Пример: Момент инерции

Согласно теореме Гюйгенса-Штейнера должно выполняться:

$$I = I_0 + mh^2$$

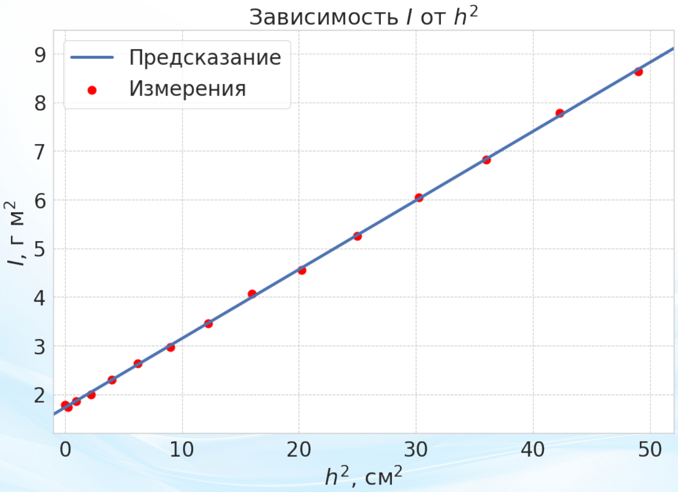
Итого, предполагается линейная зависимость момента инерции I от квадрата расстояния h^2. Мы хотим найти неизвестные m и I_0.

Наблюдения: $I_i = I_0 + mh_i^2 + \varepsilon_i$, где ε_i — погрешность.

В данном примере $x_1(h) = 1, x_2(h) = h^2$,

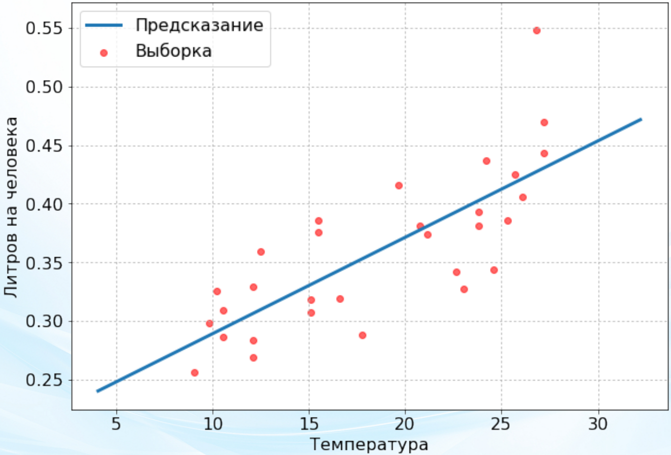
$$X = \begin{pmatrix} 1 & h_1^2 \\ \dots & \\ 1 & h_n^2 \end{pmatrix}, \; Y = \begin{pmatrix} I_1 \\ \dots \\ I_n \end{pmatrix}, \; \theta = \begin{pmatrix} I_0 \\ m \end{pmatrix}.$$

Пример: Момент инерции



Пример: Потребление мороженого

Имеет место более зашумленная зависимость.



Метод наименьших квадратов

Зависимость: $y(x) = x^T\theta$, $\theta \in \mathbb{R}^d$.

Испытания: $Y = X\theta + \varepsilon$, $X \in \mathbb{R}^{n\times d}$, $Y \in \mathbb{R}^n$.

Хотим как-то оценить параметр θ на основании полученных данных.

Пусть $\widehat{\theta} = \widehat{\theta}(X, Y)$ — наша оценка θ.

Как понять, что она хорошая?

Метрика MSE:

$$MSE(\widehat{\theta}) = \left|\left|Y - X\widehat{\theta}\right|\right|^2$$

Оценка $\widehat{\theta} = \arg\min\limits_{\theta} MSE(\widehat{\theta})$ называется оценкой по методу наименьших квадратов параметра θ.

Метод наименьших квадратов

Теорема. Если матрица $X^T X$ невырождена, то $\widehat{\theta} = (X^T X)^{-1} X^T Y$.

$MSE(\theta) = ||Y - X\theta||^2 = (Y - X\theta)^T (Y - X\theta) = Y^T Y - 2Y^T X\theta + \theta^T X^T X\theta$

Берем производную по θ и приравниваем ее к нулю.

$\frac{\partial MSE(\theta)}{\partial \theta} = -2Y^T X + 2\theta^T X^T X = 0$

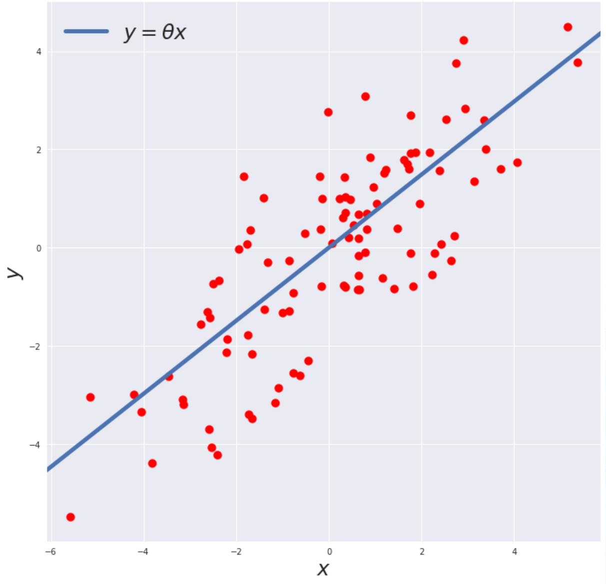
Отсюда получается утверждение теоремы. $\square$

Предсказанием отклика на новом объекте x будет величина $\widehat{y}(x) = x^T \widehat{\theta}$.

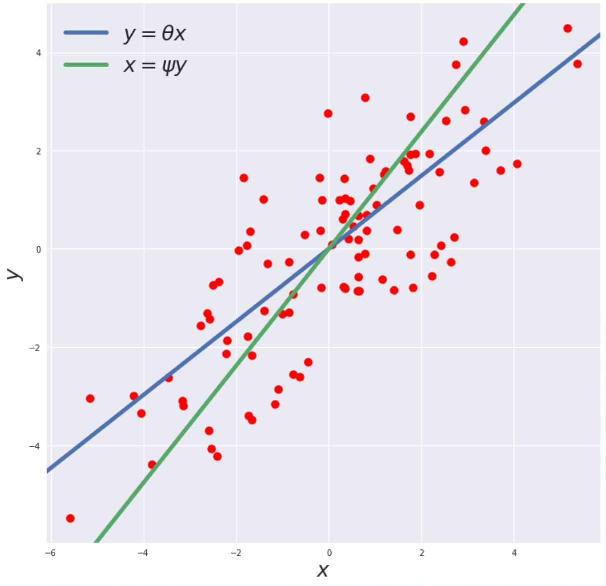
Свойства

- Если $E\varepsilon = 0$, то
$$E\hat{\theta} = \theta, E\hat{y}(x) = x\theta.$$
Оценка является несмещенной.
- Если $E\varepsilon = 0, D\varepsilon = \sigma^2 I_n$, то
$$D\theta = \sigma^2 (X^T X)^{-1}, D\hat{y}(x) = x^T \sigma^2 (X^T X)^{-1} x.$$
Дисперсия зависит от x и от матрицы $X^T X$.

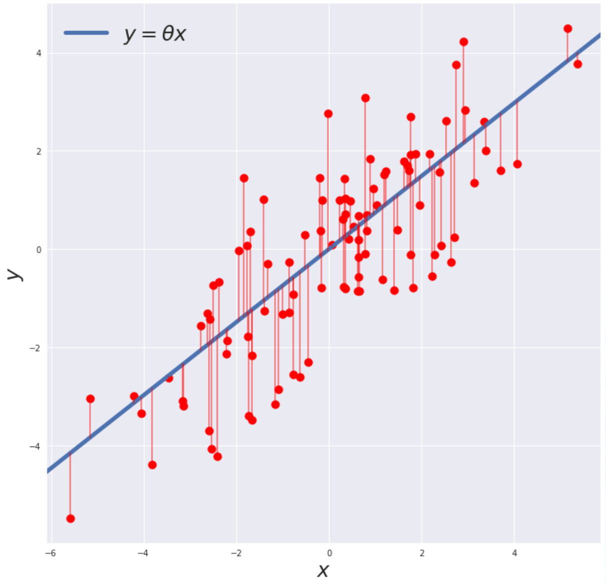
Инверсия



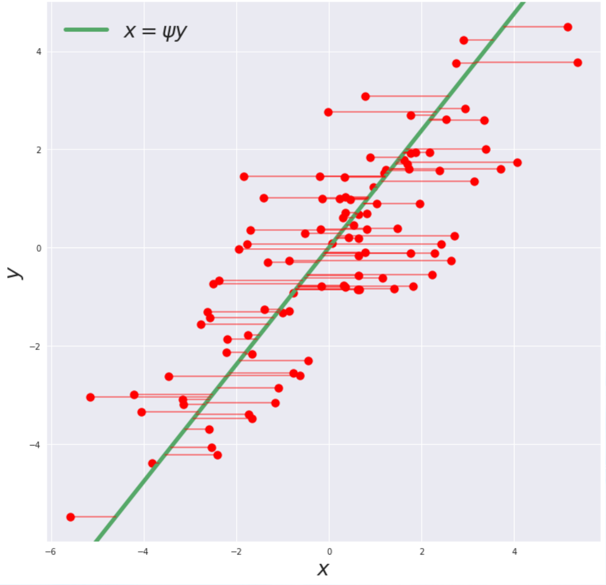
Инверсия



Инверсия



Инверсия



Реализация в sklearn

```
m = sklearn.linear_model.LinearRegression(fit_intercept=True)
```

Обучение модели:

```
m.fit(X, Y)
```

Вектор коэффициентов:

```
m.coef_
```

Свободный коэффициент:

```
m.intercept_
```

Предсказания:

```
m.predict(X)
```

Метрики качества

в задаче регрессии

Обозначения

Пусть $x_1, \dots x_n$ — признаковые описания объектов;

$Y = (Y_1, \dots Y_n)^T$ — наблюдения.

Пусть $\widehat{f}(x)$ — оцененная нами зависимость.

В случае линейной регрессии $\widehat{f}(x) = x^T\widehat{\theta}$.

Пусть $\widehat{Y}_i = \widehat{f}(x_i)$ — предсказание нашей модели на i-м объекте;

$\widehat{Y} = (\widehat{Y}_1, \dots \widehat{Y}_n)^T$.

Метрики качества в задаче регрессии

Y — реальные наблюдения, $\widehat{Y}$ — предсказания.

- MSE (Mean Squared Error):

$$MSE(Y, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

- MAE (Mean Absolute Error):

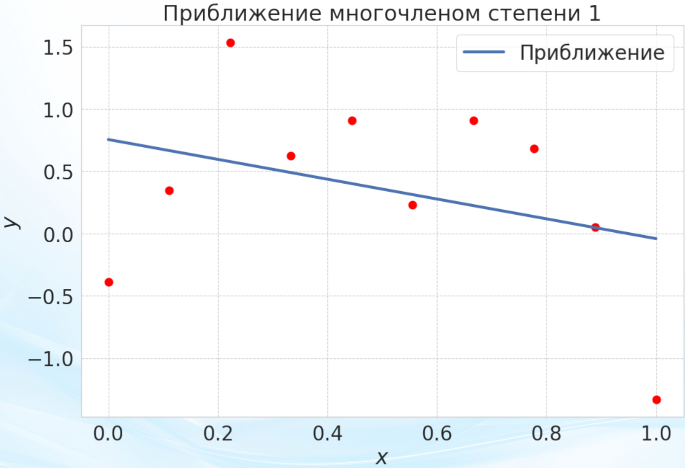
$$MAE(Y, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \widehat{Y}_i|$$

- MAPE (Mean Absolute Percentage Error):

$$MAPE(Y, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \widehat{Y}_i}{Y_i}\right| * 100\%$$

Недообучение vs Переобучение

Зависимость: $y = 5x - 6x^2$, имеется погрешность



Недообучение

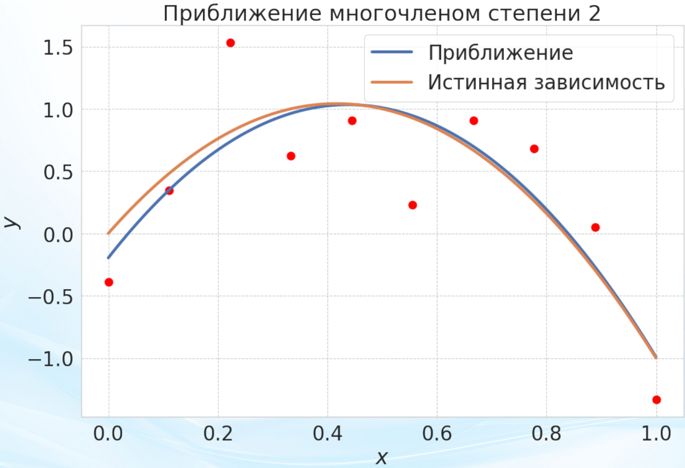
Недообучение vs Переобучение

Зависимость: $y = 5x - 6x^2$, имеется погрешность

Переобучение

Недообучение vs Переобучение

Зависимость: $y = 5x - 6x^2$, имеется погрешность



Нормально!

Тренировочная и тестовая выборки

Если все время работать с одной и той же выборкой
(это жаргон, корректно понимать "реализацией выборки")
и все больше улучшать модель, "подгонять" ее под выборку,

может возникнуть переобучение.

Предсказание на новом объекте может быть неадекватным.

Поэтому перед началом работы имеющиеся данные делят на две части:

тренировочную (обучающую) и тестовую выборки.

На тренировочной выборке происходит обучение моделей (например, оценка коэффициентов в линейной регрессии).

На тестовой выборке происходит оценка качества итоговой модели с

использванием метрик качества.

Регуляризация

Проблема: мультиколлинеарность

Мультиколлинеарность — наличие линейной зависимости между признаками.

Пример: среди признаков много признаков, связанных с размером котика. Они все зависят друг от друга и несут избыточную информацию.

Вспомним, что $D\theta = \sigma^2 (X^T X)^{-1}$.

Если признаки мультиколлинеарны, то $X^T X$ почти вырождена и дисперсия огромна.

Решение: регуляризация.

Ridge-регрессия

Задача МНК:

$$||Y - X\theta||_2 \to \min_{\theta}$$

Задача Ridge-регрессии:

$$||Y - X\theta||_2 + \lambda||\theta||_2 \to \min_{\theta}, \lambda > 0$$

Ограничиваем коэффициенты, не позволяем им "разбрасываться".

Замечание

Предварительно необходимо

- Центрировать отклик $Y := Y - \overline{Y}$ (не нужен св. член) *или не накладывать ограничение на коэффициент при константе.*
- Стандартизовать признаки — вычесть среднее, поделить на корень из дисперсии. У признаков мог быть разный масштаб!

Решение задачи

Решением задачи является

$$\hat{\theta} = (X^T X + \lambda I_d)^{-1} X^T Y$$

За счет добавки λI_d матрица стала менее вырожденной.

Свойства

- $\lambda = 0 \implies$ МНК; $\lambda = \infty \implies \hat{\theta} = 0$;
- При $\lambda > 0$ решение $\exists!$;
- Пусть $E\varepsilon = 0$. Оценка смещенная $E\hat{\theta} = (X^T X + \lambda I_d)^{-1} X^T X \theta$;
- Пусть $D\varepsilon = \sigma^2 I_n$. Дисперсия $D\hat{\theta} = \sigma^2 (X^T X + \lambda I_d)^{-1} X^T X (X^T X + \lambda I_d)^{-1}$ уменьшилась.

Lasso-регрессия

Задача МНК:

$$||Y - X\theta||_2 \to \min_\theta$$

Задача Lasso-регрессии:

$$||Y - X\theta||_2 + \lambda||\theta||_1 \to \min_\theta, \lambda > 0,$$
$$||\theta||_1 = |\theta_1| + |\theta_2| + \ldots + |\theta_d|.$$

Решается итеративными методами.

Свойства

Lasso-регрессия зануляет коэффициенты с ростом λ, может использоваться для отбора признаков.


ВСЁ!