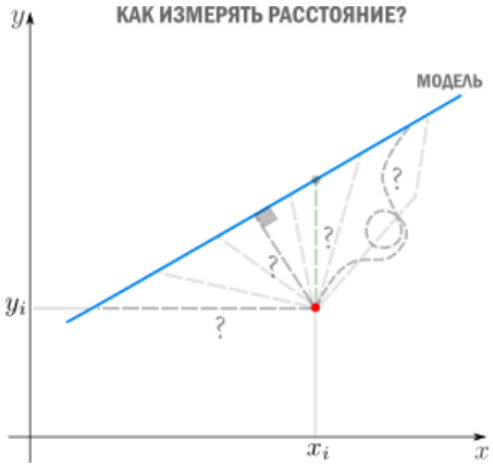
Метод наименьших квадратов

Начнём с простейшего двумерного случая. Пусть нам даны точки на плоскости $\{(x_1, y_1), \cdots, (x_N, y_N)\}$ и мы ищем такую аффинную функцию

$$f(x) = a + b \cdot x,$$



чтобы ее график ближе всего находился к точкам. Таким образом, наш базис состоит из константной функции и линейной $(1, x)$.

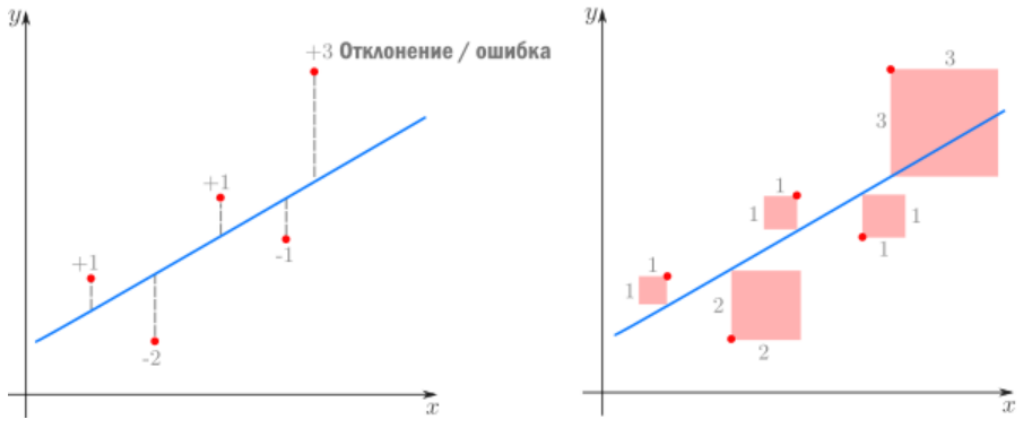
Как видно из иллюстрации, расстояние от точки до прямой можно понимать по-разному, например геометрически — это длина перпендикуляра. Однако в контексте нашей задачи нам нужно функциональное расстояние, а не геометрическое. Нас интересует разница между экспериментальным значением и предсказанием модели для каждого x_i, поэтому измерять нужно вдоль оси y.

Первое, что приходит в голову, в качестве функции потерь попробовать выражение, зависящее от абсолютных значений разниц $|f(x_i) - y_i|$. Простейший вариант — сумма модулей отклонений $\sum_i |f(x_i) - y_i|$ приводит к Least Absolute Distance (LAD) регрессии.

Впрочем, более популярная функция потерь — сумма квадратов отклонений регрессанта от модели. В англоязычной литературе она носит название Sum of Squared Errors (SSE)

$$\mathrm{SSE}(a, b) = \mathrm{SS}_{res[iduals]} = \sum_{i=1}^{N} \text{отклонение}_i{}^2 = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} (y_i - a - b \cdot x_i)^2,$$

Метод наименьших квадратов (по англ. OLS) — линейная регрессия с $\mathrm{SSE}(a, b)$ в качестве функции потерь.



Такой выбор прежде всего удобен: производная квадратичной функции — линейная функция, а линейные уравнения легко решаются.

Математический анализ

Простейший способ найти $\operatorname{argmin}_{a,b} \operatorname{SSE}(a,b)$ — вычислить частные производные по a и b, приравнять их нулю и решить систему линейных уравнений

$$\frac{\partial}{\partial a}\operatorname{SSE}(a,b) = -2\sum_{i=1}^{N}(y_i - a - bx_i),$$
$$\frac{\partial}{\partial b}\operatorname{SSE}(a,b) = -2\sum_{i=1}^{N}(y_i - a - bx_i)x_i.$$

Значения параметров, минимизирующие функцию потерь, удовлетворяют уравнениям

$$0 = -2\sum_{i=1}^{N}(y_i - \hat{a} - \hat{b}x_i),$$
$$0 = -2\sum_{i=1}^{N}(y_i - \hat{a} - \hat{b}x_i)x_i,$$

которые легко решить

$$\hat{a} = \frac{\sum_i y_i}{N} - \hat{b}\frac{\sum_i x_i}{N},$$
$$\hat{b} = \frac{\frac{\sum_i x_i y_i}{N} - \frac{\sum_i x_i \sum_i y_i}{N^2}}{\frac{\sum_i x_i^2}{N} - \left(\frac{\sum_i x_i^2}{N}\right)^2}.$$

Мы получили громоздкие и неструктурированные выражения. Сейчас мы их облагородим и вдохнем в них смысл.

Статистика

Полученные формулы можно компактно записать с помощью статистических эстиматоров: среднего $\langle\cdot\rangle$, вариации σ. (стандартного отклонения), ковариации $\sigma(\cdot,\cdot)$ и корреляции $\rho(\cdot,\cdot)$

$$\hat{a} = \langle y\rangle - \hat{b}\langle x\rangle,$$
$$\hat{b} = \frac{\langle xy\rangle - \langle x\rangle\langle y\rangle}{\langle x^2\rangle - \langle x\rangle^2}.$$

Перепишем $\hat{b}$ как

$$\hat{b} = \frac{\sigma(x,y)}{\sigma_x^2},$$

где σ_x это нескорректированное (смещенное) стандартное выборочное отклонение, а $\sigma(x,y)$ — ковариация. Теперь вспомним, что коэффициент корреляции (коэффициент корреляции Пирсона)

$$\rho(x,y) = \frac{\sigma(x,y)}{\sigma_x\sigma_y}$$

и запишем

$$\hat{b} = \rho(x,y)\frac{\sigma_y}{\sigma_x}.$$

ЗАДАНИЕ:

1. Решить задачу регрессии для набора

x	y
0	4
1	7
2	7
3	8

Вручную (на основе математического анализа с помощью numpy) и с помощью scikit-learn.

Сравнить полученные коэффициенты.

2. Рассмотреть задачу прогнозирование цен на примере набора данных Houses.csv:

- какие признаки наиболее всего влияют, по Вашему мнению, на цену?
- возможно ли уменьшить количество признаков?
- есть ли пропуски в данных?
- можно ли обойтись одним параметром? выберите один и решите задачу.
- оставьте несколько параметров и решите задачу многомерной регрессии.

Линейная регрессия с несколькими переменными

Линейная регрессия с несколькими переменными также известна как «множественная линейная регрессия». Введем обозначения для уравнений, где мы можем иметь любое количество входных переменных:

$x^{(i)}$ - вектор-столбец всех значений признаков i-го обучающего примера;

$x_j^{(i)}$ - значение j-го признака i-го обучающего примера;

m - количество примеров в обучающей выборке;

n - количество признаков;

X - матрица признаков;

b - вектор параметров регрессии.

Заметим, что в будущем для удобства примем, что $x_0^{(i)} = 1$ для всех i. Другими словами, мы для удобства введем некий суррогатный признак, для всех наблюдений равный единице. это сильно упростит математические выкладки, особенно в матричной форме.

Теперь определим множественную форму функции гипотезы следующим образом, используя несколько признаков:

$$h_b(x) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

Используя определение матричного умножения, наша многопараметрическая функция гипотезы может быть кратко представлена в виде: *h(x) = B X*.

Функция ошибки

Для множественной регрессии функция ошибки от вектора параметров b выглядит следующим образом:

$$J(b) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_b(x^{(i)}) - y^{(i)} \right)^2$$

Или в матричной форме:

$$J(b) = \frac{1}{2m} (Xb - \vec{y})^T (Xb - \vec{y})$$