

Лекция 2: Линейные модели

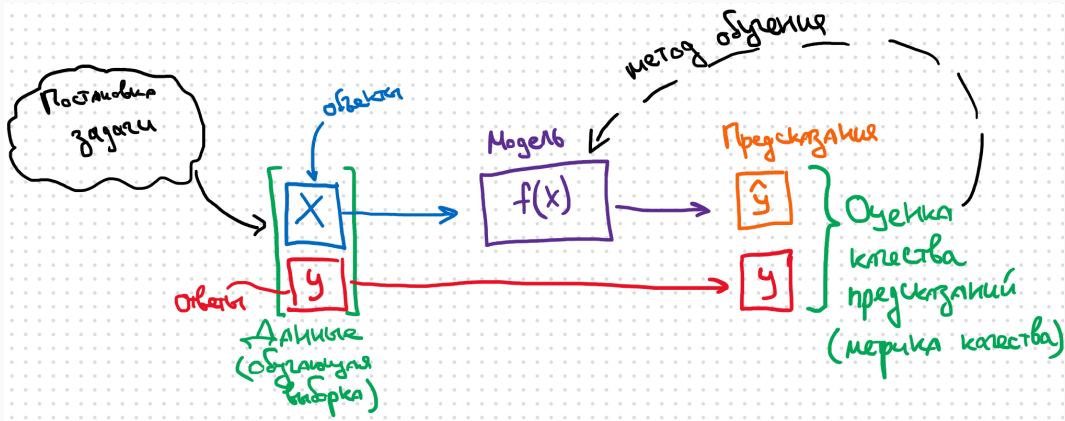
Линейная регрессия

13 февраля 2021

- Линейная регрессия
- Линейные классификаторы

Сегодня будем разбираться с линейной регрессией.

Как формализовывать задачу?



- Постановка задачи
- Данные
- Модель (семейство алгоритмов)
- Оценка качества (функционал ошибки)
- Метод обучения

- Постановка задачи
- Данные
- Модель (семейство алгоритмов)
- Оценка качества (функционал ошибки)
- Метод обучения

Постановка задачи

Перенесемся в начало XIX века.



- Вы физик-инженер, у вас на столе лежит кусок неизвестного гибкого материала, который вы хотите использовать в своем устройстве

Постановка задачи

Перенесемся в начало XIX века.



- Вы физик-инженер, у вас на столе лежит кусок неизвестного гибкого материала, который вы хотите использовать в своем устройстве
- Однако, вы не знаете его коэффициента растяжения, и не можете рассчитать возникающие силы

Постановка задачи

Перенесемся в начало XIX века.



- Вы физик-инженер, у вас на столе лежит кусок неизвестного гибкого материала, который вы хотите использовать в своем устройстве
- Однако, вы не знаете его коэффициента растяжения, и не можете рассчитать возникающие силы
- Вы хотите построить зависимость силы упругости от растяжения для этого материала (и узнать коэффициент растяжения)

- Постановка задачи: регрессия, предсказать F по x
- Данные
- Модель (семейство алгоритмов)
- Оценка качества (функционал ошибки)
- Метод обучения

- Постановка задачи: регрессия, предсказать F по x
- **Данные**
- Модель (семейство алгоритмов)
- Оценка качества (функционал ошибки)
- Метод обучения

Вы проводите некоторое количество измерений и заносите все в табличку подобного вида:

| Растяжение (м) | Сила упругости (Н) |
|----------------|--------------------|
| 0.182654 | 0.328200 |
| 0.482319 | 0.870494 |
| 0.177778 | 0.326217 |
| 0.742819 | 1.334474 |
| 0.117160 | 0.212199 |
| ... | ... |
| 0.074164 | 0.142423 |

- Постановка задачи: регрессия, предсказать F по x
- **Данные: одномерные (единственный признак x)**
- Модель (семейство алгоритмов)
- Оценка качества (функционал ошибки)
- Метод обучения

- Постановка задачи: регрессия, предсказать F по x
- Данные: одномерные (единственный признак x)
- **Модель (семейство алгоритмов)**
- Оценка качества (функционал ошибки)
- Метод обучения

Благодаря Роберту Гуку вы знаете, что сила упругости зависит от растяжения линейно:

$$F = kx$$

Благодаря Роберту Гуку вы знаете, что сила упругости зависит от растяжения линейно:

$$F = kx$$

Общий вид линейной функции одной переменной:

$$y(x) = w_1x + w_0$$

Благодаря Роберту Гуку вы знаете, что сила упругости зависит от растяжения линейно:

$$F = kx$$

Общий вид линейной функции одной переменной:

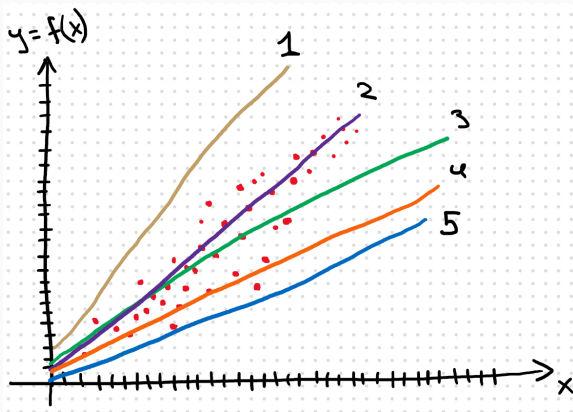
$$y(x) = w_1x + w_0$$

Таким образом, ваша задача среди всех таких функций (семейства) найти ту конкретную, которая лучше всего вам подходит (определить параметры)

- Постановка задачи: регрессия, предсказать F по x
- Данные: одномерные (единственный признак x)
- **Модель: линейная функция одного аргумента**
- Оценка качества (функционал ошибки)
- Метод обучения

Что значит “лучше всего вам подходит”?

Формализация задачи



Какая из функций лучше?

- Постановка задачи: регрессия, предсказать F по x
- Данные: одномерные (единственный признак x)
- Модель: линейная функция одного аргумента
- **Оценка качества (функционал ошибки)**
- Метод обучения

Функционал ошибки — функция $\mathcal{L}(y, \hat{y})$, значение которой показывает, насколько хороши предсказания модели относительно правильных ответов.

Функционал ошибки — функция $\mathcal{L}(y, \hat{y})$, значение которой показывает, насколько хороши предсказания модели относительно правильных ответов.

Наиболее популярные для задачи регрессии:

$$MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Ещё есть MSLE, MAPE, SMAPE, и т.д.

Мы в ближайшее время будем использовать MSE, поскольку она имеет хорошие математические свойства.

- Постановка задачи: регрессия, предсказать F по x
- Данные: одномерные (единственный признак x)
- Модель: линейная функция одного аргумента
- **Функционал ошибки: MSE**
- Метод обучения

- Постановка задачи: регрессия, предсказать F по x
- Данные: одномерные (единственный признак x)
- Модель: линейная функция одного аргумента
- Функционал ошибки: MSE
- **Метод обучения**

Обучение линейной регрессии: подбор таких параметров w_0, w_1 , чтобы результирующая функция давала наименьшее значение MSE, то есть

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min_{w_0, w_1}$$

Обучение линейной регрессии: подбор таких параметров w_0, w_1 , чтобы результирующая функция давала наименьшее значение MSE, то есть

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min_{w_0, w_1}$$

$$\frac{1}{n} \sum_{i=1}^n (y_i - w_1 x_i - w_0)^2 \rightarrow \min_{w_0, w_1}$$

Как это сделать?

Матанализ: в точке экстремума функции производная обращается в ноль

$$\begin{cases} \frac{d}{dw_0} \frac{1}{n} \sum_{i=1}^n (y_i - w_1 x_i - w_0)^2 = 0 \\ \frac{d}{dw_1} \frac{1}{n} \sum_{i=1}^n (y_i - w_1 x_i - w_0)^2 = 0 \end{cases}$$

$$\frac{d}{dw_0} \frac{1}{n} \sum_{i=1}^n (y_i - w_1 x_i - w_0)^2 = \frac{d}{dw_0} \sum_{i=1}^n (y_i - w_1 x_i - w_0)^2 =$$

$$\begin{aligned}\frac{d}{dw_0} \frac{1}{n} \sum_{i=1}^n (y_i - w_1 x_i - w_0)^2 &= \frac{d}{dw_0} \sum_{i=1}^n (y_i - w_1 x_i - w_0)^2 = \\ &= \sum_{i=1}^n \frac{d}{dw_0} (y_i - w_1 x_i - w_0)^2 = \sum_{i=1}^n 2(y_i - w_1 x_i - w_0)(-1) = 0\end{aligned}$$

$$\begin{aligned}\frac{d}{dw_0} \frac{1}{n} \sum_{i=1}^n (y_i - w_1 x_i - w_0)^2 &= \frac{d}{dw_0} \sum_{i=1}^n (y_i - w_1 x_i - w_0)^2 = \\&= \sum_{i=1}^n \frac{d}{dw_0} (y_i - w_1 x_i - w_0)^2 = \sum_{i=1}^n 2(y_i - w_1 x_i - w_0)(-1) = 0 \\&\Rightarrow \sum_{i=1}^n (y_i - w_1 x_i - w_0) = \sum_{i=1}^n y_i - w_1 \sum_{i=1}^n x_i - n w_0 = 0\end{aligned}$$

$$\begin{aligned}\frac{d}{dw_0} \frac{1}{n} \sum_{i=1}^n (y_i - w_1 x_i - w_0)^2 &= \frac{d}{dw_0} \sum_{i=1}^n (y_i - w_1 x_i - w_0)^2 = \\&= \sum_{i=1}^n \frac{d}{dw_0} (y_i - w_1 x_i - w_0)^2 = \sum_{i=1}^n 2(y_i - w_1 x_i - w_0)(-1) = 0 \\&\Rightarrow \sum_{i=1}^n (y_i - w_1 x_i - w_0) = \sum_{i=1}^n y_i - w_1 \sum_{i=1}^n x_i - n w_0 = 0 \\&\Rightarrow n w_0 = \sum_{i=1}^n y_i - w_1 \sum_{i=1}^n x_i\end{aligned}$$

$$\frac{d}{dw_1} \frac{1}{n} \sum_{i=1}^n (y_i - w_1 x_i - w_0)^2 =$$

$$\begin{aligned} & \frac{d}{dw_1} \frac{1}{n} \sum_{i=1}^n (y_i - w_1 x_i - w_0)^2 = \\ &= \sum_{i=1}^n \frac{d}{dw_1} (y_i - w_1 x_i - w_0)^2 = \sum_{i=1}^n 2(y_i - w_1 x_i - w_0)(-x_i) = 0 \end{aligned}$$

$$\begin{aligned} & \frac{d}{dw_1} \frac{1}{n} \sum_{i=1}^n (y_i - w_1 x_i - w_0)^2 = \\ &= \sum_{i=1}^n \frac{d}{dw_1} (y_i - w_1 x_i - w_0)^2 = \sum_{i=1}^n 2(y_i - w_1 x_i - w_0)(-x_i) = 0 \\ &\Rightarrow \sum_{i=1}^n x_i (y_i - w_1 x_i - w_0) = 0 \end{aligned}$$

$$\begin{aligned} & \frac{d}{dw_1} \frac{1}{n} \sum_{i=1}^n (y_i - w_1 x_i - w_0)^2 = \\ &= \sum_{i=1}^n \frac{d}{dw_1} (y_i - w_1 x_i - w_0)^2 = \sum_{i=1}^n 2(y_i - w_1 x_i - w_0)(-x_i) = 0 \\ &\Rightarrow \sum_{i=1}^n x_i (y_i - w_1 x_i - w_0) = 0 \\ &\Rightarrow w_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i - w_0 \sum_{i=1}^n x_i \end{aligned}$$

$$\begin{cases} nw_0 = \sum_{i=1}^n y_i - w_1 \sum_{i=1}^n x_i \\ w_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i - w_0 \sum_{i=1}^n x_i \end{cases}$$

Запишите эту систему, она нам еще понадобится на семинаре

- Постановка задачи: регрессия, предсказать F по x
- Данные: одномерные (единственный признак x)
- Модель: линейная функция одного аргумента
- Функционал ошибки: MSE
- **Метод обучения: аналитическое решение системы уравнений**

- Постановка задачи: регрессия, предсказать F по x
- Данные: одномерные (единственный признак x)
- Модель: линейная функция одного аргумента
- Функционал ошибки: MSE
- Метод обучения: аналитическое решение системы уравнений

Численно рассчитаем этот пример на семинаре, а сейчас разберем многомерную регрессию

Многомерная регрессия

Рассмотрим случай многомерной (множественной) регрессии — когда функция зависит больше чем от одного входного признака.

Каждый пример x теперь вектор в d -мерном пространстве

$$\vec{x} = (x_1, x_2, \dots, x_d)$$

Как будет выглядеть линейная модель в данном случае?

Многомерная регрессия

Рассмотрим случай многомерной (множественной) регрессии — когда функция зависит больше чем от одного входного признака.

Каждый пример x теперь вектор в d -мерном пространстве

$$\vec{x} = (x_1, x_2, \dots, x_d)$$

Как будет выглядеть линейная модель в данном случае?

$$f(\vec{x}) = f(x_1, x_2, \dots) =$$

Многомерная регрессия

Рассмотрим случай многомерной (множественной) регрессии — когда функция зависит больше чем от одного входного признака.

Каждый пример x теперь вектор в d -мерном пространстве

$$\vec{x} = (x_1, x_2, \dots, x_d)$$

Как будет выглядеть линейная модель в данном случае?

$$\begin{aligned} f(\vec{x}) &= f(x_1, x_2, \dots) = \\ &= w_0 + w_1x_1 + w_2x_2 + \dots + w_dx_d = w_0 + \sum_{j=1}^d w_jx_j \end{aligned}$$

$$W_0 + \sum_{j=1}^d W_j X_j =$$

$$w_0 + \sum_{j=1}^d w_j x_j = w_0 + \langle \vec{w}, \vec{x} \rangle =$$

$$w_0 + \sum_{j=1}^d w_j x_j = w_0 + \langle \vec{w}, \vec{x} \rangle =$$

$$= \langle \vec{w}, \vec{x} \rangle + w_0 \cdot 1 = \langle \vec{w}', \vec{x}' \rangle$$

$$\vec{x}' = (1, x_1, x_2, \dots, x_d) \quad \vec{w}' = (w_0, w_1, \dots, w_d)$$

Для всей выборки получаем

$$\vec{y} = X\vec{w}$$

$$\vec{y} = (y_1, y_2, \dots, y_N) \quad \vec{w} = (w_0, w_1, \dots, w_d)$$

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nd} \end{bmatrix}$$

$$\frac{1}{N} \|X\vec{w} - \vec{y}\|^2 \rightarrow \min_{\vec{w}}$$

$$\frac{1}{N} \|X\vec{w} - \vec{y}\|^2 \rightarrow \min_{\vec{w}}$$

$$\frac{d}{d\vec{w}} \|X\vec{w} - \vec{y}\|^2 = 0$$

$$\frac{1}{N} \|X\vec{w} - \vec{y}\|^2 \rightarrow \min_{\vec{w}}$$

$$\frac{d}{d\vec{w}} \|X\vec{w} - \vec{y}\|^2 = 0$$

$$\vec{w}^* = \arg \min_{\vec{w}} \|X\vec{w} - \vec{y}\|^2 = (X^T X)^{-1} X^T \vec{y}$$