Машинное обучение

Лекция 2. Линейные модели. Градиентный спуск

(12.02.2022)

Даниил Литвинов
Лаврентий Данилов

Общие сведения

План

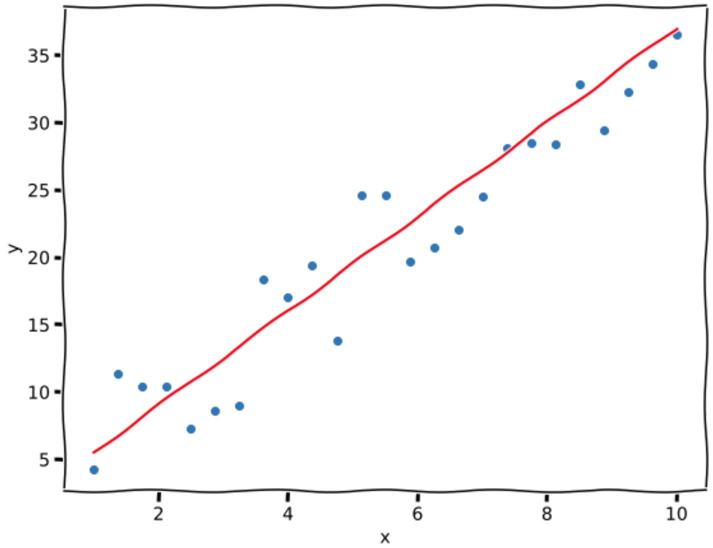
1. Линейная модель регрессии
2. Как линейные модели обучаются?
3. Линейная модель классификации

Что это такое?

x — баллы за экзамен по английскому 1

y — баллы за экзамен по английскому 2

x	y
1	5
3	11
9	35
10	33



Что это такое?

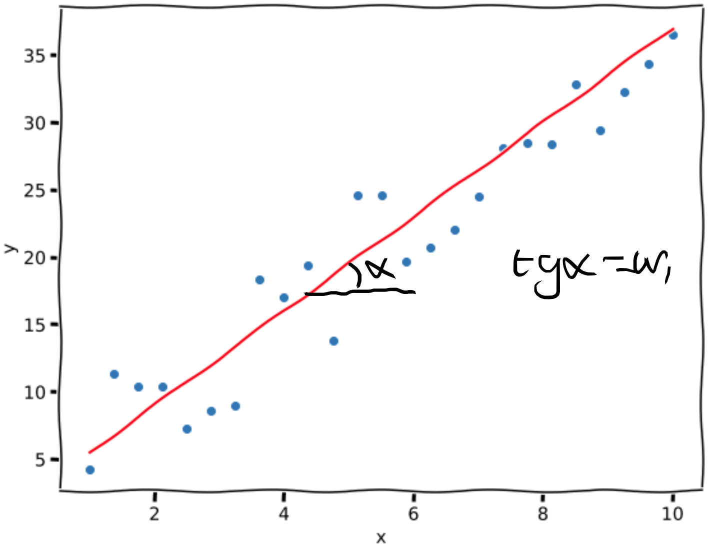
свободный член
(intercept)

случайная ошибка
(error)

$$y = w_0 + w_1 x + \epsilon$$

целевая переменная
(aka target)

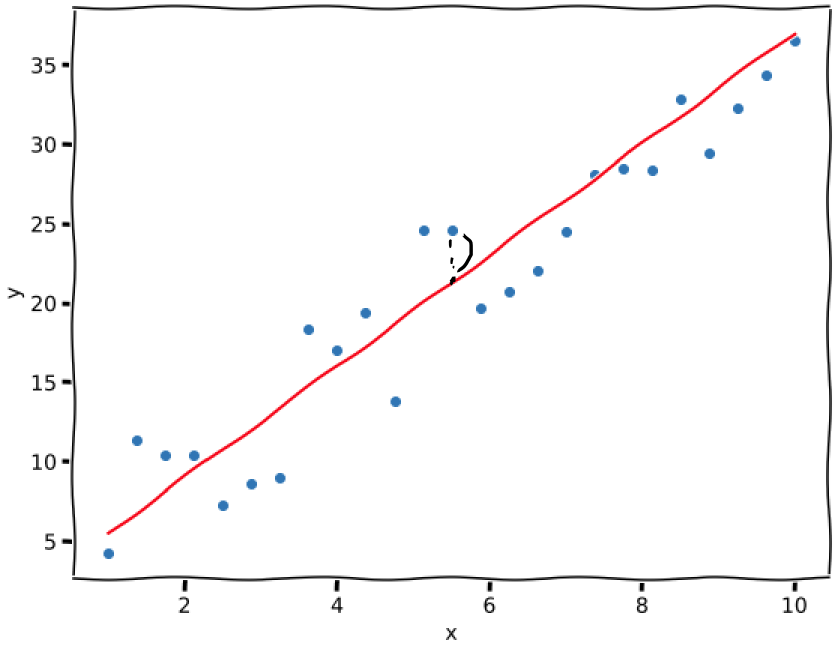
независимая переменная
(slope, predictor)



А какая модель нам нужна?

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \to \min$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \to \min$$



Интерпретация коэффициентов

Зачем нужны линейные модели?

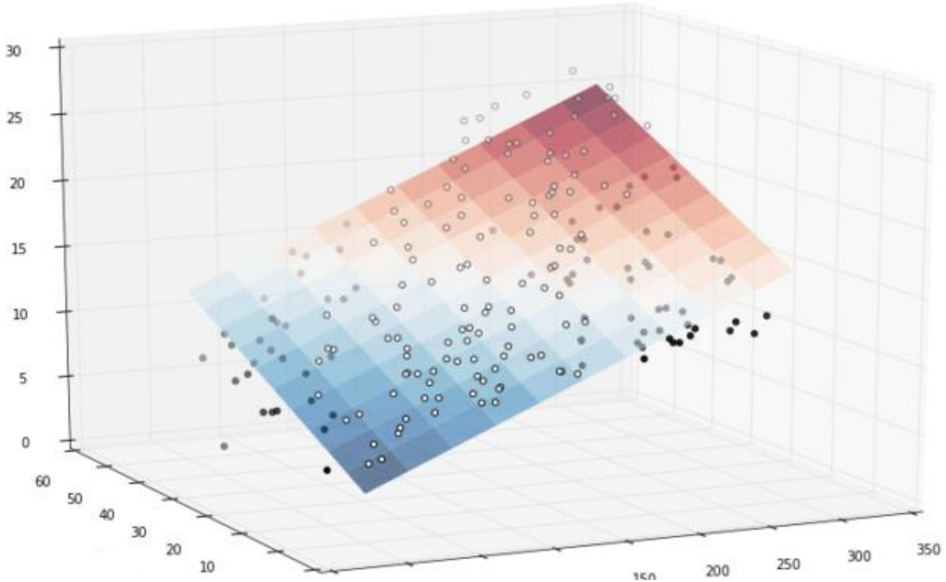
1. Предсказание интересующей нас величины
2. Оценка влияния различных факторов на нашу целевую переменную
3. Линейные модели очень легко использовать и интерпретировать
4. Линейные модели могут восстанавливать даже **нелинейные зависимости**

А если у нас много независимых переменных?

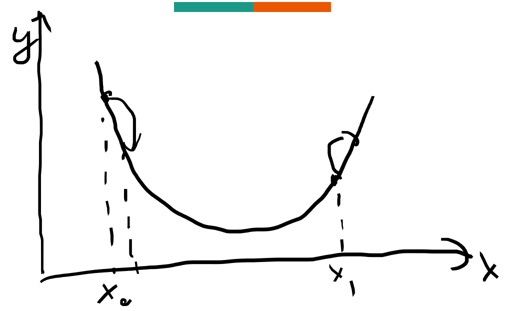
$$y = w_0 + w_1 x + w_2 z + \ldots + w_n t + \epsilon$$

площадь	число комнат	школа близко	цена квартиры
50	2	нет	5000
1000	7	да	11000
30	1	нет	3500
100	4	нет	33333

Множественная линейная регрессия дает нам плоскость



Производные



$$\frac{dy}{dx} = \frac{y(x_0 + \Delta x) - y(x_0)}{\Delta x}$$

$y = f(x)$	$\frac{\mathrm{d}y}{\mathrm{d}x} = f'(x)$
k, any constant	0
x	1
x^2	$2x$
x^3	$3x^2$
x^n, any constant n	nx^{n-1}
e^x	e^x
e^{kx}	$k\mathrm{e}^{kx}$
$\ln x = \log_{\mathrm{e}} x$	$\frac{1}{x}$
$\sin x$	$\cos x$
$\sin kx$	$k \cos kx$
$\cos x$	$-\sin x$
$\cos kx$	$-k \sin kx$

Производные

$(0, 0, \frac{\pi}{2})$ - ответ

$$\varphi(x,y,z) = 2x^2 + 3y^2 - \sin z$$

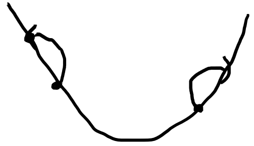
$$\frac{\partial\varphi}{\partial x} = 4x$$

$$\frac{\partial\varphi}{\partial y} = 6 \cdot y$$

$$\frac{\partial\varphi}{\partial z} = -\cos z$$

$$\begin{pmatrix} 4x \\ 6y \\ -\cos z \end{pmatrix}$$ — градиент $\varphi = \nabla_{xyz}\varphi$

Производные (градиент)



- указывает в напр. наиб. роста ф-ции
- Антиградиент указ. в напр. ~~наиб.~~ наискор. убыв.

Производные

Как оценивать коэффициенты модели?

$$\underset{n\times 3}{X}, \quad \underset{n\times 1}{y}, \quad \underset{4\times 1}{w}$$

$$y = w_0 \cdot 1 + w_1 x_1 \dots w_3 x_3$$

$$\begin{array}{cccc} 1 & x_1 & x_2 & x_3 \end{array}$$
$$\begin{bmatrix} 1 & & & \\ 1 & & & \\ 1 & & & \\ 1 & & & \end{bmatrix}$$

$$n \times 4$$

$$\underset{n\times 1}{y} = \underset{n\times 4}{X} \cdot \underset{4\times 1}{w}$$

$$= 1 \cdot w_0 + x_1 w_1 + \dots + x_3 w_3$$

$$\begin{bmatrix} 1 & x_1 & x_2 & x_3 \\ & & & \\ & & & \end{bmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} \cdot \\ \\ \\ \end{pmatrix}$$

Как оценивать коэффициенты модели?

matrix, веса

$$MSE = \frac{1}{N}\big(\underset{n\times 1}{y} - \underset{n\times 1}{Xw}\big)^2 = \frac{1}{N}\underset{n\times 1}{(y - Xw)}^T\underset{n\times 1}{(y - Xw)}$$

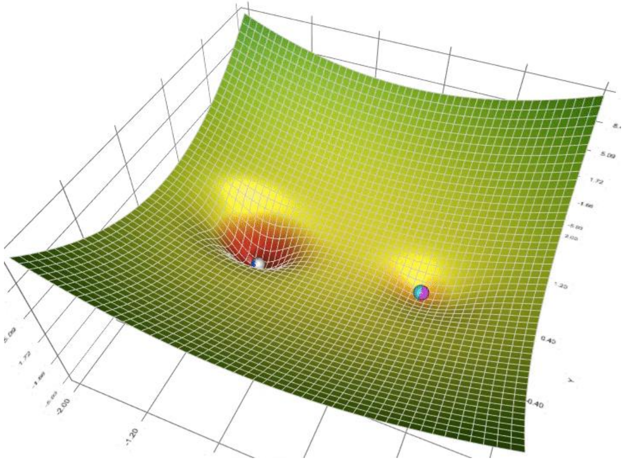
$$\underset{4\times 1}{\nabla_w MSE} = \boxed{-\frac{1}{N}\cdot 2\cdot \underset{4\times n}{X^T}\underset{n\times 1}{(y - Xw)}} = \boxed{\frac{2}{N}\cdot X^T(Xw - y)}$$

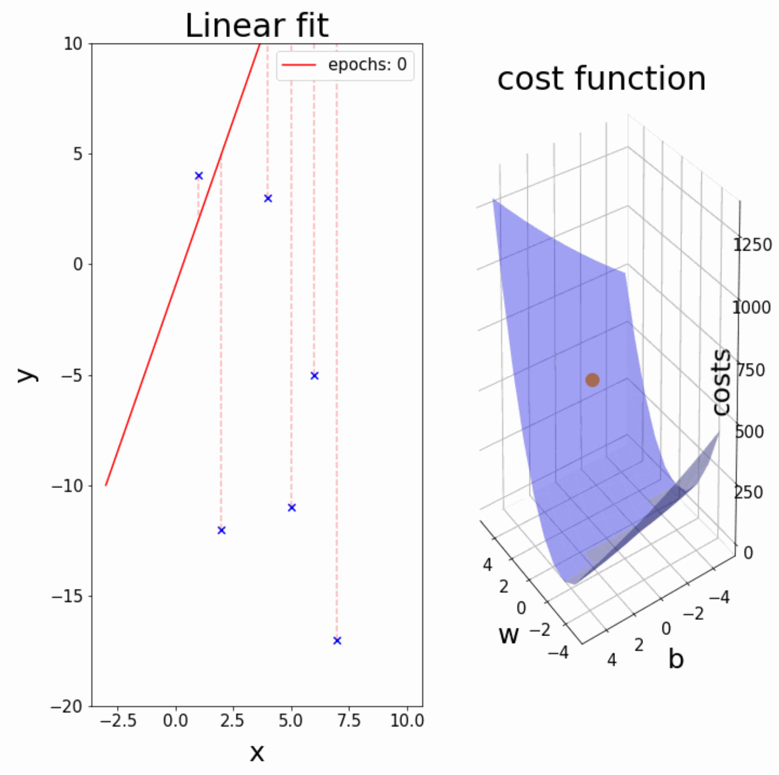
$$\underset{4\times 1}{w_{(i+1)}} = \underset{4\times 1}{w_{(i)}} - \eta\, \nabla MSE(w_i)$$

eta – learning rate

η: [0.001, 0.01]

Градиентный спуск





Формулы

$$y = w_0 + w_1 x + \epsilon$$

$$y = Xw$$

$$\frac{dLoss}{dw} = \nabla Loss = \frac{2}{n} X^T (Xw - y)$$

$$Loss = \frac{1}{n}(y - Xw)^T (y - Xw)$$

$$\boxed{w = (X^T X)^{-1} X^T y}$$

Градиентный спуск

$$Loss = \frac{1}{N}(y - Xw)^T(y - Xw) \qquad \frac{dLoss}{dw} = \nabla Loss = \frac{2}{N}X^T(Xw - y)$$

```
w = np.random.randn(m + 1)
Пока grad(Loss) != 0:
        w -= η * grad(Loss)
```

Отдых -> логистическая регрессия

Связь событий и признаков

В зависимости от предикторов события могут происходить чаще или реже – логика, совпадающая с логикой связи количественной переменной отклика с набором предикторов.

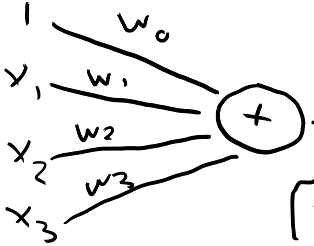
Например, по мере роста температуры воздуха летом чаще будут встречаться люди в шортах: событие “встретился человек в шортах” положительно связано с температурой воздуха.

Событие “проведение исследования” явно связана с предиктором “объем полученного финансирования”, однако эта связь может быть совсем непростой.

А что если хотим классификацию?

$y = \{0, 1\}$

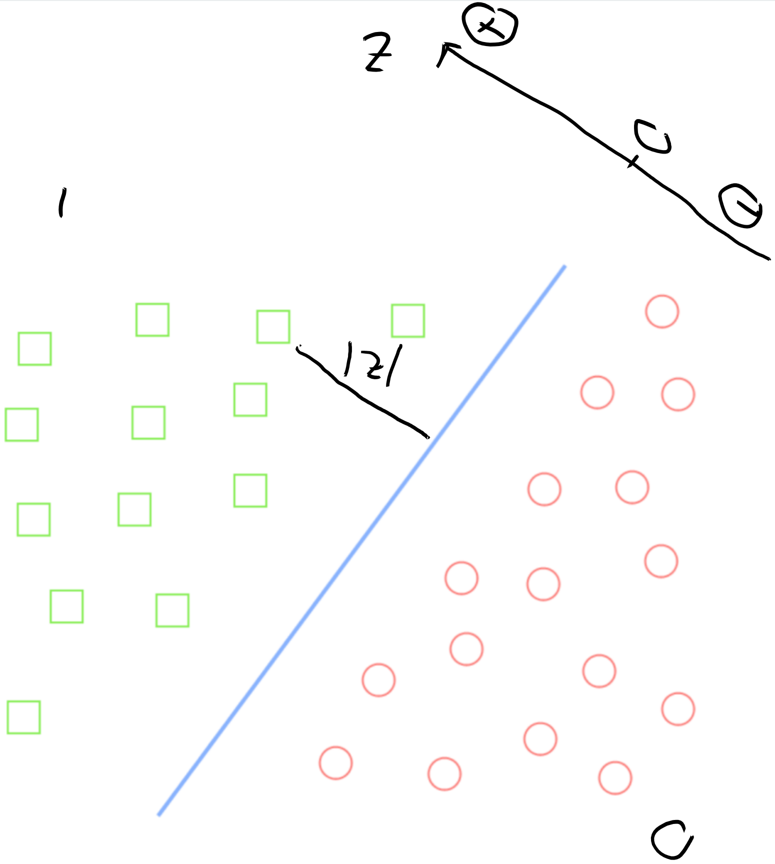
Допустим бинарная классификация



$\to z \quad (-\infty; +\infty)$

[если z больше ⊕, то y = 1

[ ] ⊖, то y = 0



[геометрия ML]

Отношение шансов

Шансы (odds) часто представляют в виде отношения шансов (odds ratio)

Если отношение шансов > 1, то вероятность наступления события выше, чем вероятность того, что оно не произойдет.

Если отношение шансов < 1, то наоборот.

Если можно оценить вероятность положительного события, то отношение шансов выглядит так:

$$odds = \frac{\pi}{1-\pi}$$

Отношение шансов варьируется от 0 до +∞.

Попробуем сами

$(-\infty; +\infty)$

① p_i – [вероятность того, что] $y_i = 1$

$1 - p_i$ – [ ] $y_i = 0$

② $odds = \frac{p_i}{1-p_i}$ $[0; +\infty)$

③. $logit = \ln\left(\frac{p_i}{1-p_i}\right)$ $(-\infty; +\infty)$

Логиты

Отношение шансов можно преобразовать в логиты(logit):

$$ln(odds) = ln(\frac{\pi}{1-\pi})$$

- Значения логитов − это трансформированные оценки вероятности события.
- Логиты варьируют от $-\infty$ до $+\infty$.
- Логиты симметричны относительно 0, т.е. $ln(1)$.
- Для построения моделей в качестве зависимой переменной удобнее брать логиты.

Считаем вероятность

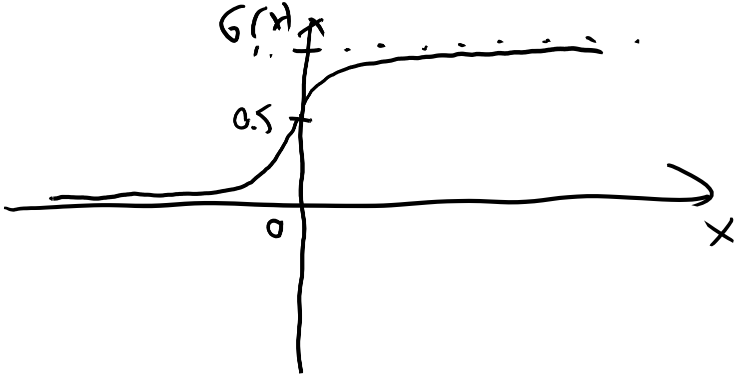
$$\ln\left(\frac{p_i}{1-p_i}\right) = Xw \quad \Big|\, \exp$$

$$\frac{p_i}{1-p_i} = e^{Xw}$$

$$p_i = e^{Xw} - p_i e^{Xw}$$

$$p_i(1 + e^{Xw}) = e^{Xw}$$

$$p_i = \frac{e^{Xw}}{1+e^{Xw}} = \frac{1}{e^{-Xw}+1} = \sigma(Xw)$$



Как такое учить? BCE Loss

$$BCE = \frac{1}{N} \sum_{i=1}^{N} -\left[y_i \cdot \ln p_i + (1 - y_i) \cdot \ln(1 - p_i) \right]$$

$y_i = \{0, 1\}$

$p_i \in [0, 1]$ $\quad p_i = \sigma(x_i w)$

1) $y_i = 1; p_i = 1; BCE = 0$

2) $y_i = 1; p_i = 0; BCE = +\infty$

3) $y_i = 1; p_i = 0{,}5; BCE = 0.7$

Как такое учить? BCE Loss

$$\nabla_w BCE = \boxed{X^T \cdot (p - y)}$$

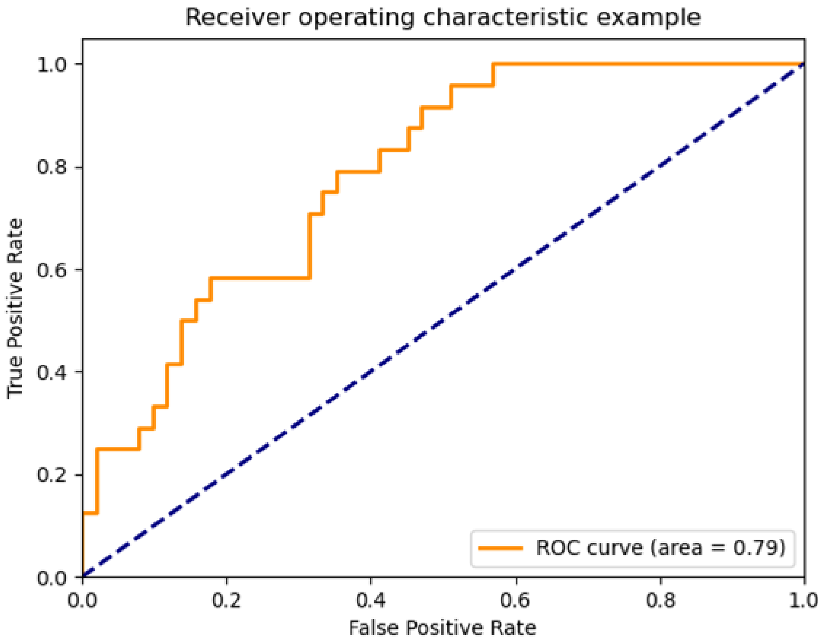
$$p = \sigma(Xw)$$

Качество классификации

Качество классификации. ROC кривая

$$TPR = \frac{TP}{TP + FN} = recall$$

$$FPR = \frac{FP}{FP + TN}$$



[рисуем свою ROC кривую]

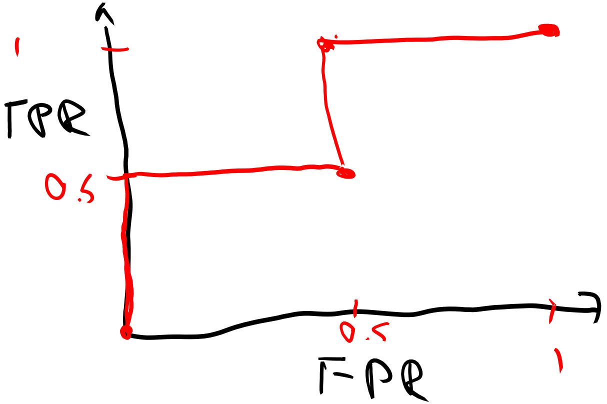
Построение ROC кривой

порог

y	p	0	0.2	0.4	0.6	0.8	1
1	0.8	1	1	1	1	1	0
0	0.6	1	1	1	1	0	0
1	0.4	1	1	1	0	0	0
0	0.2	1	1	0	0	0	0
TPR		1	1	1	0.5	0.5	0
FPR		1	1	0.5	0.5	0	0

$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{FP+TN}$$



$\boxed{\text{ROC-AUC} = 0.75}$