ОБРАБОТКА И МОДЕЛИРОВАНИЕ ДАННЫХ В MS EXCEL

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Екатерина Золотарева

ТЕМА 3. ПРОГНОЗИРОВАНИЕ

Обучение с учителем

Что нужно сделать?

Предсказать значение переменной Y (метки) на основе имеющихся данных о переменных X_1, X_2, X_3…X_n (признаках) **= восстановить зависимость**

Как?

Минимизировать расхождение между предсказанным и истинным значением переменной Y.

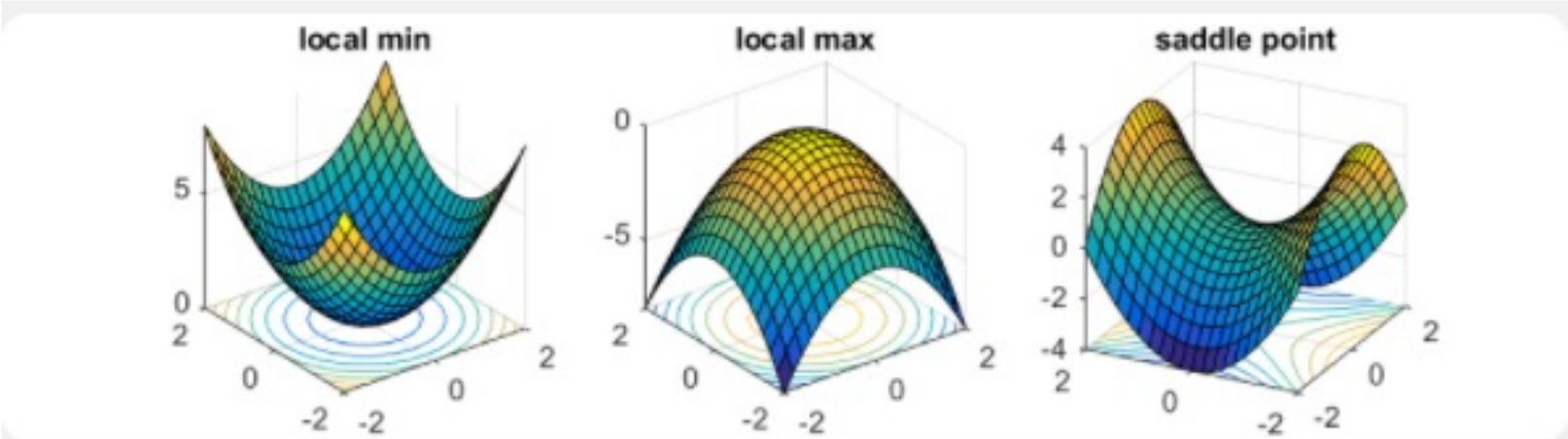
- **Функция потерь** (cost function) - целевая функция, которая оценивает это расхождение
- **Обучение модели** (training) - подбор оптимальных параметров модели за счет минимизации функции потерь

Обучение= оптимизация

Обучение модели (training) - подбор оптимальных параметров модели за счет минимизации функции потерь

Оптимизация – нахождение критических точек, среди которых нам нужны точки минимума

Различные типы критических точек



local min – локальный минимум local max – локальный максимум saddle point – седловая точка

Общая схема оптимизации

1. Записываем функцию f(x), которую хотим оптимизировать
2. Берем производную f’(x)
3. Приравниваем производную к нулю f’(x)=0
4. Решаем уравнение, находим корни x*.
5. Среди x* находим такие, которые соответствуют минимуму

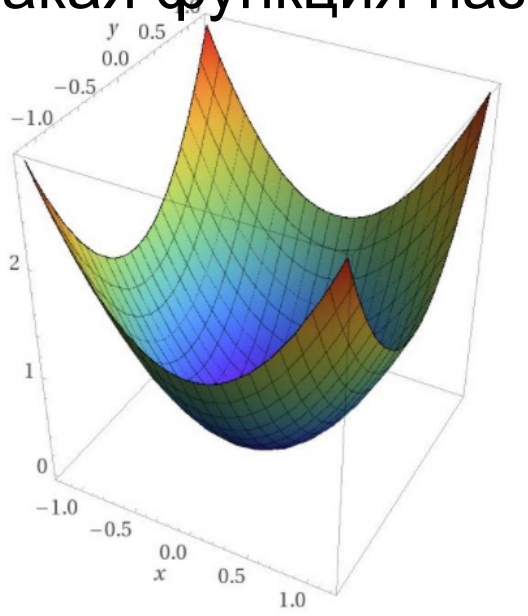
Если переменных много

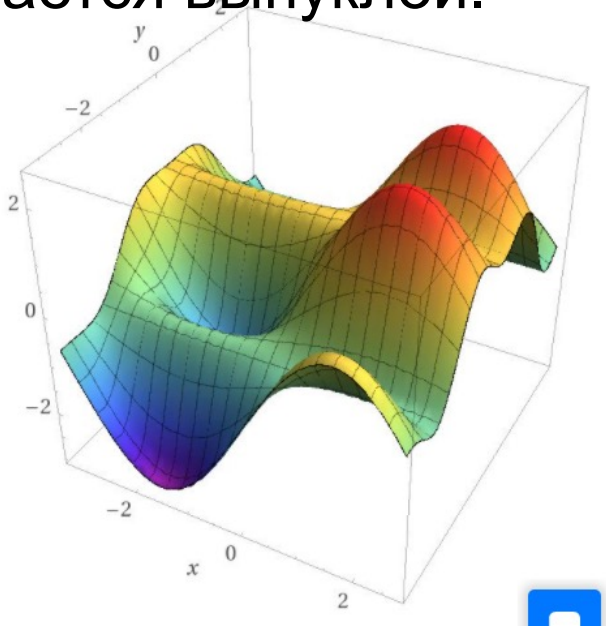
1. Записываем функцию f(x1,x2,x3…), которую хотим оптимизировать
2. Берем производные по каждой переменной f'_{x1}, f'_{x2}, f'_{x3}. Набор производных – это **градиент**
3. Приравниваем каждую из производных к нулю
4. Решаем систему уравнений, находим корни x1*, x2*, x3*.
5. Среди x* находим такие, которые соответствуют минимуму

Функция потерь

- **Функция потерь** (cost function) - целевая функция, которая оценивает расхождение между предсказанным и истинным значением переменной Y. Нужно найти ее минимум
- Удобно, когда функция имеет одну критическую точку, и она и есть минимум. Такая функция называется выпуклой.

Удобная функция (выпуклая)





Неудобная функция

Пример

Задание 12 Профильного ЕГЭ по математике

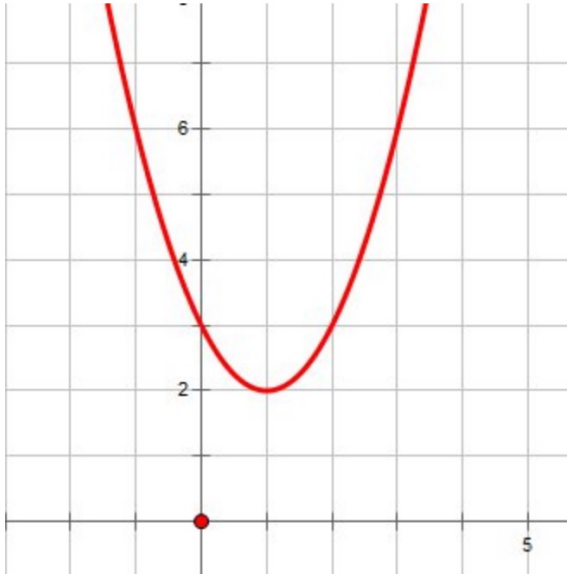
Задание 12 первой части Профильного ЕГЭ по математике — это нахождение точек максимума и минимума функции, а также наибольших и наименьших значений функции с помощью производной.

Вот какие типы задач могут встретиться в этом задании:

Нахождение точек максимума и минимума функций

Исследование сложных функций

Нахождение наибольших и наименьших значений функций на отрезке



https://ege-study.ru/zadanie-12-profilnogo-EGE-po-matematike

ЛИНЕЙНАЯ РЕГРЕССИЯ

Линейная регрессия

Формализация модели

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots + \theta_n x_n$, где

$\theta_0, \theta_1 \ldots \theta_n$ – коэффициенты, **параметры или веса** модели

Целевая функция → min

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

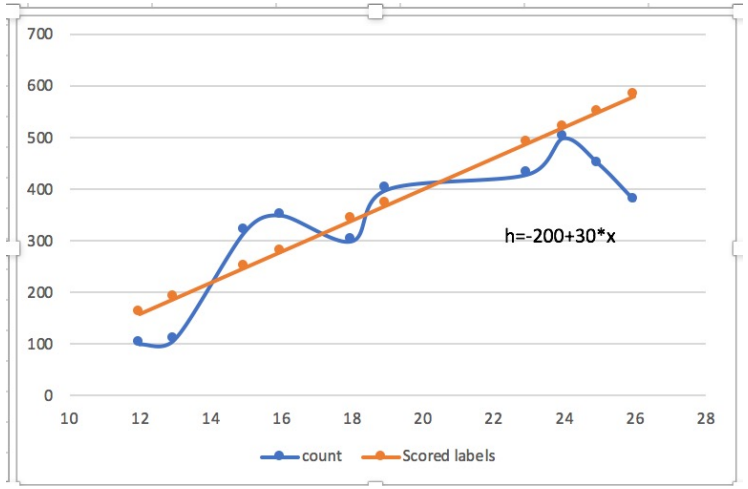
МНК в «обычной» форме

Кроме МНК могут использоваться и другие методы, например, метод максимального правдоподобия

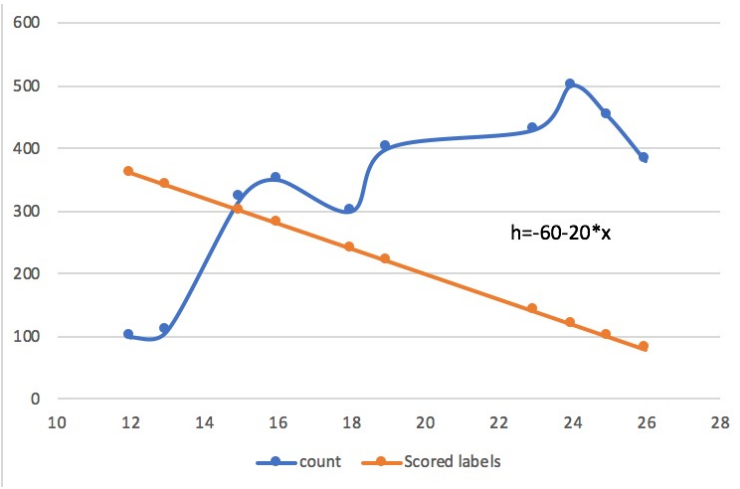
$$J(\theta) = \frac{1}{2m} (X\theta - \vec{y})^T (X\theta - \vec{y})$$, где $h_\theta(X) = X\theta$ МНК в векторной форме

Линейная регрессия. Пример

	atemp	count	Scored labels
	x	y	h=k1*x+k0
1	12	100	160
2	13	110	180
3	15	320	220
4	16	350	240
5	18	300	280
6	19	400	300
7	23	430	380
8	24	500	400
9	25	450	420
10	26	380	440



Надо найти такие значения k0 и k1, для которых суммарная ошибка будет минимальна



Линейная регрессия. Пример

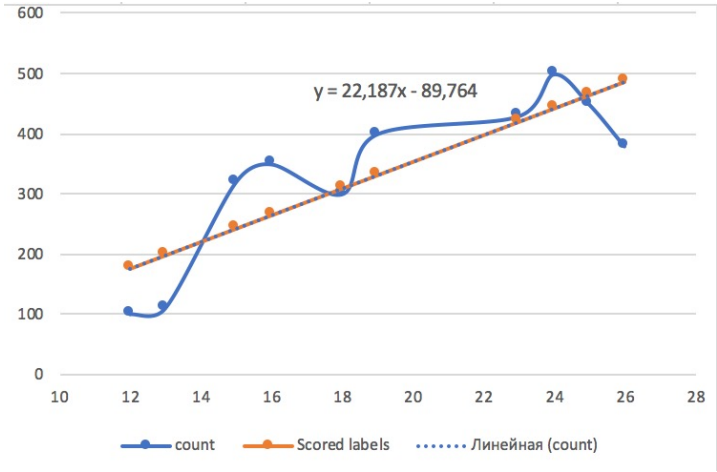
							k1	22,1865766
	atemp	count	Scored labels	Расхождение	Модуль	Квадрат	k0	-89,763613
Номер	x	y	h=k1*x+k0	y-h	|y-h|	(y-h)^2	x^2	x*y
1	12	100	176,475306	-76,47530604	76,475306	5848,47243	144	1200
2	13	110	198,6618827	-88,66188265	88,6618827	7860,92944	169	1430
3	15	320	243,0350359	76,96496412	76,9649641	5923,6057	225	4800
4	16	350	265,2216125	84,77838751	84,7783875	7187,37499	256	5600
5	18	300	309,5947657	-9,594765724	9,59476572	92,0595293	324	5400
6	19	400	331,7813423	68,21865766	68,2186577	4653,78525	361	7600
7	23	430	420,5276488	9,472351203	9,4723512	89,7254373	529	9890
8	24	500	442,7142254	57,28577459	57,2857746	3281,65997	576	12000
9	25	450	464,900802	-14,90080203	14,900802	222,033901	625	11250
10	26	380	487,0873786	-107,0873786	107,087379	11467,7067	676	9880
	191	**3340**	**3340**	**-3,41061E-13**	**593,44027**	**46627,3533**	**3885**	**69050**

Надо найти такие значения k0 и k1, для которых суммарная квадратичная ошибка (выделена красным) будет минимальна

Решение

$$\begin{cases} k_1 \sum x_i^2 + k_0 \sum x_i = \sum x_i y_i \\ k_1 \sum x_i + k_0 \cdot m = \sum y_i \end{cases}$$

	k0	k1	b
	191	3885	69050
	10	191	3340
Ответ			
		-89,763613	
		22,1865766	



Линейная регрессия: решение (для МНК)

Аналитически:

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

Повторяем до достижения критерия остановки
NB! Существует много вариаций

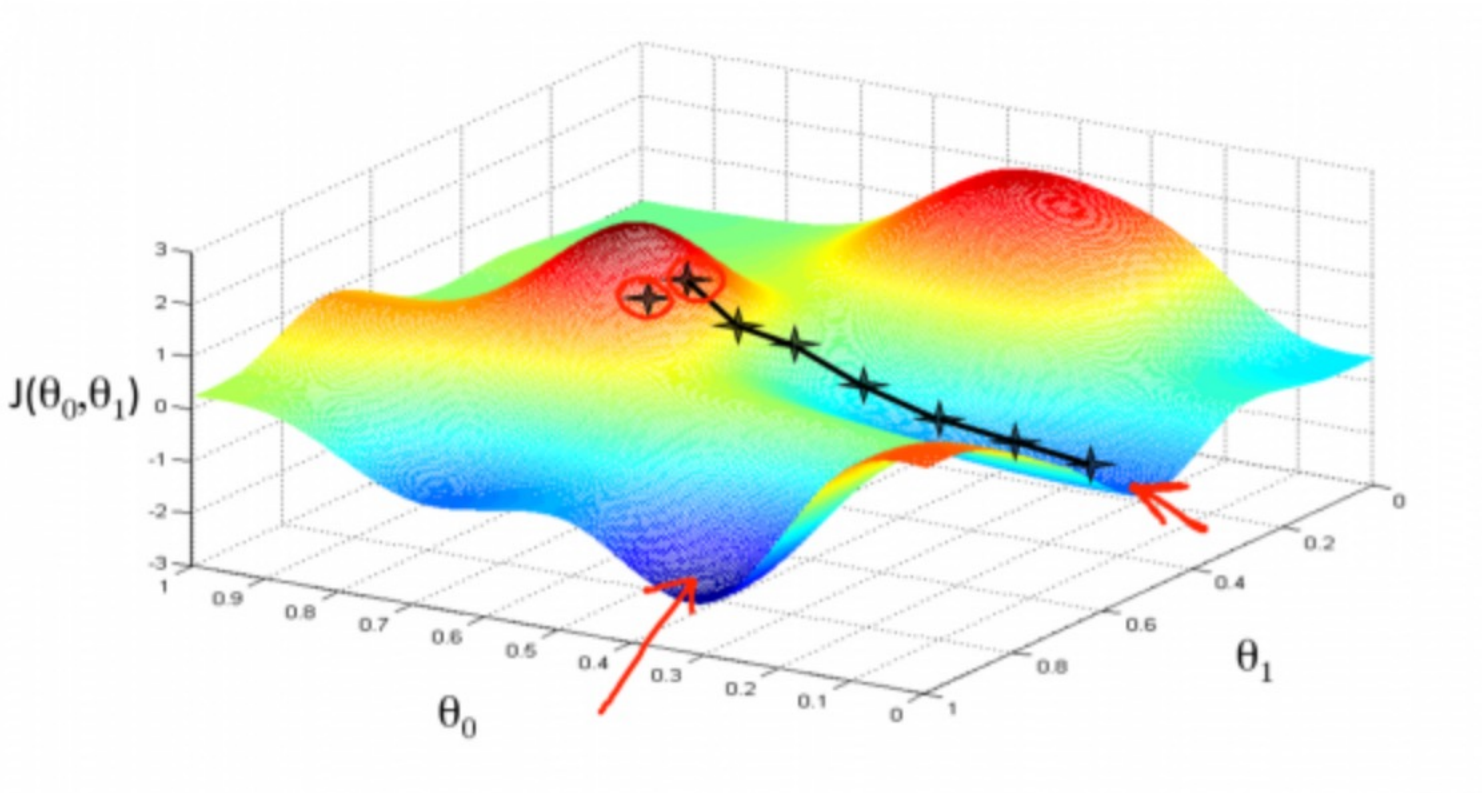
Численные методы: градиентный спуск

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$ или $$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}$$ в «обычной» форме

$$\theta := \theta - \alpha \nabla J(\theta)$$ или $$\theta := \theta - \frac{\alpha}{m} X^T (X\theta - \vec{y})$$ в векторной форме

вектор-градиент

Градиентный спуск



https://neurohive.io/ru/osnovy-data-science/gradient-descent/

Линейная регрессия: плюсы

- Оптимальна, когда зависимость близка к линейной
- Легко масштабируется
- Подходит для «разреженных» данных
- Легко интерпретируется
- Служит основой для других алгоритмов

ОЦЕНКА КАЧЕСТВА

Регрессия: оценка качества

- Средняя абсолютная ошибка (Mean absolute error)

$$MAE = \frac{\sum_{j=1}^{m} |y^{(j)} - \mathrm{h}^{(j)}|}{m}$$

- Средняя относительная ошибка (Relative absolute error)

$$RSE = \frac{\sum_{j=1}^{m} |y^{(j)} - \mathrm{h}^{(j)}|}{\sum_{j=1}^{m} |y^{(j)} - \bar{y}|}$$

Регрессия: оценка качества

- Среднеквадратичная ошибка (Root mean squared error)

$$RMSE = \sqrt{\frac{\sum_{j=1}^{m}(y^{(j)} - \mathrm{h}^{(j)})^2}{m}} = \sqrt{\frac{SSE}{m}}$$

- Относительная среднеквадратичная ошибка (Relative squared error)

$$RSE = \frac{\sum_{j=1}^{m}(y^{(j)} - \mathrm{h}^{(j)})^2}{\sum_{j=1}^{m}(y^{(j)} - \bar{y})^2} = \frac{SSE}{SST}$$

Регрессия: оценка качества

- Коэффициент детерминации (R^2)

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$, где:

$SST = \sum_{j=1}^{m}(y^{(j)} - \bar{y})^2$ - общая дисперсия

$SSR = \sum_{j=1}^{m}(\mathrm{h}^{(j)} - \bar{y})^2$ - объясненная дисперсия

$SSE = \sum_{j=1}^{m}(y^{(j)} - \mathrm{h}^{(j)})^2$ - остатки

R^2 принимает значения от 0 до 1,

чем ближе к 1, тем лучше