ТЕХНОСФЕРА

Лекция №2

Часть 1

Введение в машинное обучение

Спасёнов Алексей

Введение в машинное обучение. Лекция 2

Отмечаемся и оставляем отзывы.

Спасибо!

Введение в машинное обучение. Лекция 2

Содержание лекции

1. Задача регрессии
    1. Линейная регрессия
    2. Регуляризация
2. Метрики
3. Пример

Обучение с учителем

Задачи классификации (classification)

- $F_j = \{true, false\}$ – классификация на 2 класса
- $F_j = \{1, \dots, M\}$ – классификация на M непересекающихся классов
- $F_j = \{0,1\}^M$ - классификация на M классов, которые могут пересекаться

Задача восстановления регрессии (regression)

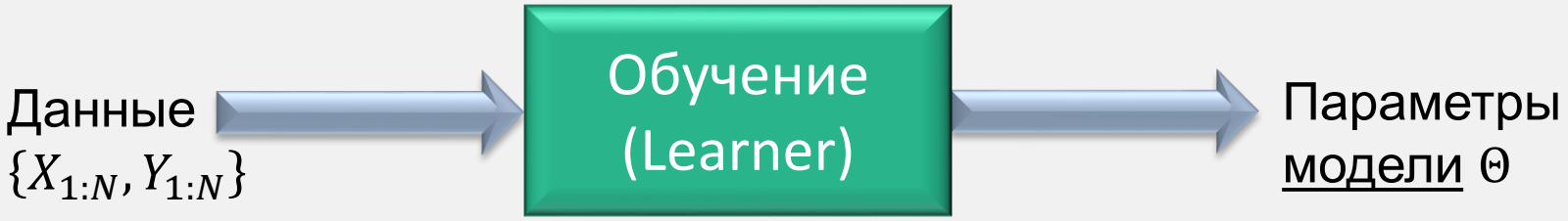
- $F_j = \mathbb{R}$ или $F_j = \mathbb{R}^M$ (ответом является действительное число или числовой вектор)

Задача ранжирования (learning to rank)

- F_j - конечно упорядочено (ответы надо получить сразу на множестве объектов, после чего отсортировать их по значениям ответов)

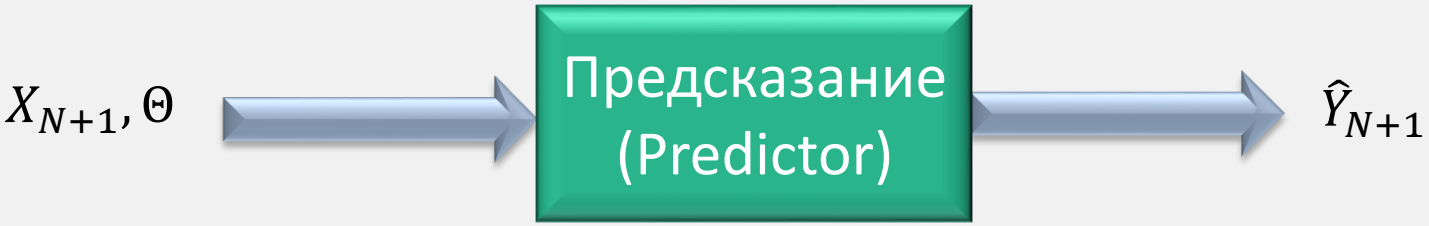
Обучение с учителем

Этап обучения (train)



Необходимо учитывать представительность выборки

Этап применения (test)

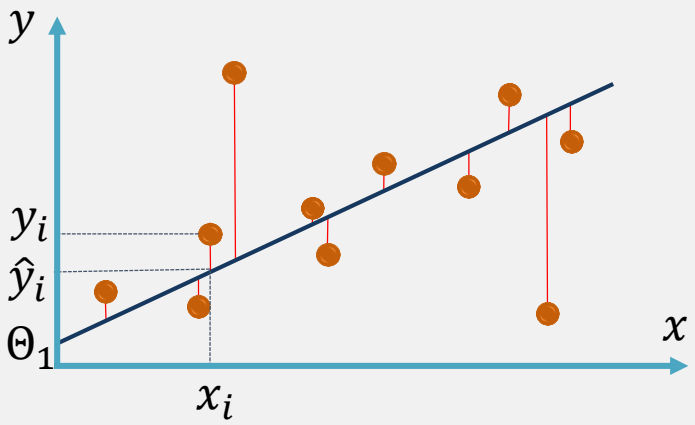


Задача восстановления регрессии

Линейная регрессия

Модель:

$$\hat{y}(X_i) = \Theta_1 + x_i\Theta_2$$



Целевая функция (Objective function, Energy, Loss)

Величина ошибки алгоритма на обучающей выборке

Пример для задачи регрессии:

$$J(\Theta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \Theta_1 - x_i\Theta_2)^2$$

Метод наименьших квадратов (Ordinary Least Squares)

Задача восстановления регрессии

Линейная регрессия

Пример задачи регрессии:

Предсказание стоимости жилья

Целевая переменная

Area	Bedroom	Kitchen	HouseStyle	Neighborhood	YearBuilt	Alley	SalePrice
8450	3	1	2Story	CollgCr	2003	NA	208500
9600	3	1	1Story	Veenker	1976	NA	181500
11250	3	1	2Story	CollgCr	2001	NA	223500
9550	3	1	2Story	Crawfor	1915	NA	140000
14260	4	1	2Story	NoRidge	2000	NA	250000
14115	1	1	1.5Fin	Mitchel	1993	Grvl	143000
10084	3	1	1Story	Somerst	2004	NA	307000
10382	3	1	2Story	NWAmes	1973	NA	200000
6120	2	2	1.5Fin	OldTown	1931	NA	129900
7420	2	2	1.5Unf	BrkSide	1939	Grvl	118000

Признаки

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Задача восстановления регрессии

Линейная регрессия

Модель:

$$\hat{y}_i = \sum_{j=1}^{d} x_{ij}\Theta_j = 1 * \Theta_1 + x_{i2}\Theta_2 + x_{i3}\Theta_3 + \ \dots + x_{id}\Theta_d$$

В матричной форме:

$$\hat{y} = X\Theta$$

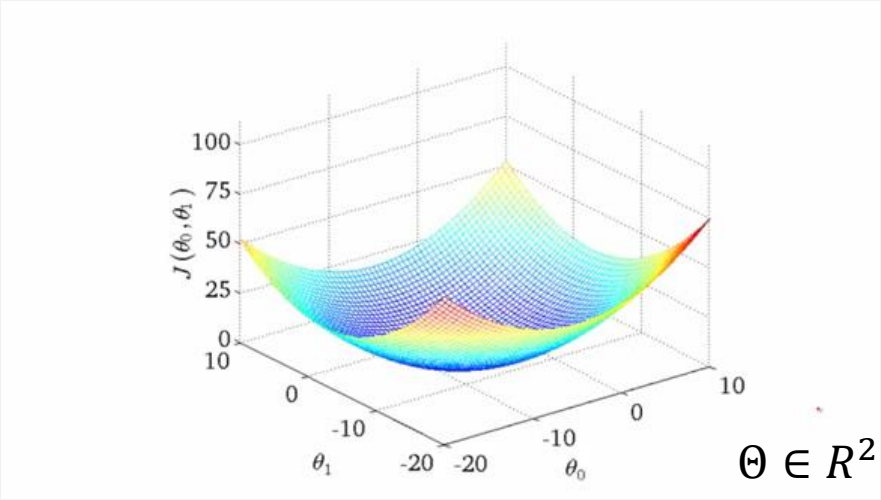
$$\begin{bmatrix} \hat{y}_1 \\ \dots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1d} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nd} \end{bmatrix} \begin{bmatrix} \Theta_1 \\ \dots \\ \Theta_d \end{bmatrix}$$

Задача восстановления регрессии

Линейная регрессия

Целевая функция:

$$J(\Theta) = (y - X\Theta)^T(y - X\Theta) = \sum_{i=1}^{n}(y_i - X_i^T\Theta)^2$$



Задача восстановления регрессии

Линейная регрессия

Целевая функция:

$$J(\Theta) = (y - X\Theta)^T(y - X\Theta) = \sum_{i=1}^{n}(y_i - X_i^T\Theta)^2$$

Поиск решения:

$$\frac{\partial J(\Theta)}{\partial \Theta} = \frac{\partial}{\partial \Theta}\left(y^T y - 2y^T x\Theta + \Theta^T x^T x\Theta\right) = 0$$

$$\Theta = (x^T x)^{-1}x^T y$$

Задача восстановления регрессии

Оптимизация

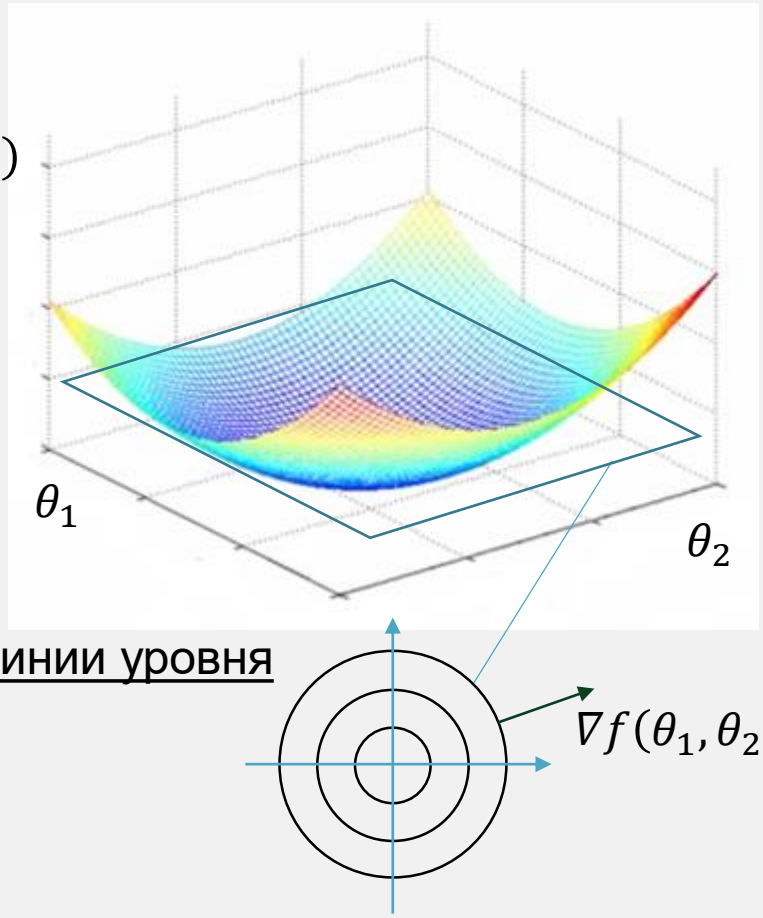
Функция:

$$f(\theta_1, \theta_2) = {\theta_1}^2 + {\theta_2}^2$$

Частные производные:

$$\frac{\delta f(\theta_1,\theta_2)}{\delta \theta_1} = 2\theta_1, \frac{\delta f(\theta_1,\theta_2)}{\delta \theta_2} = 2\theta_2$$

$$\nabla f(\theta_1, \theta_2) = \begin{bmatrix} 2\theta_1 \\ 2\theta_2 \end{bmatrix}$$



Задача восстановления регрессии

Проблема переобучения

1)Обучающая выборка

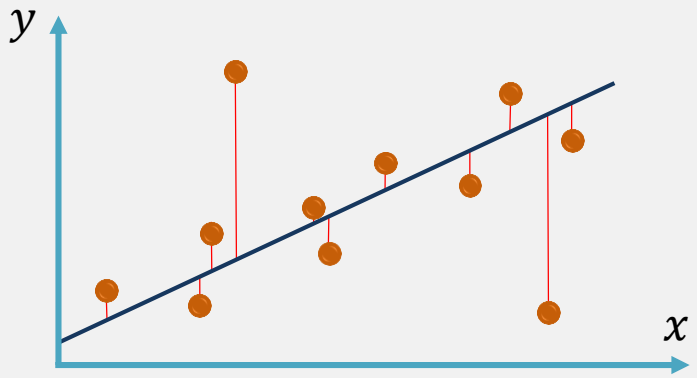
2)Контрольная выборка

Пример

Модель: $h(X,\theta) = \theta_0 + \theta_1 \cdot x + \cdots + \theta_n x^n$

Целевая функция: $J(X,\Theta) = \sum_{i=0}^{n}(\theta_0 + \theta_1 \cdot x_i + \cdots + \theta_n x_i^n - y_i)^2$

Что будет, если увеличить n?



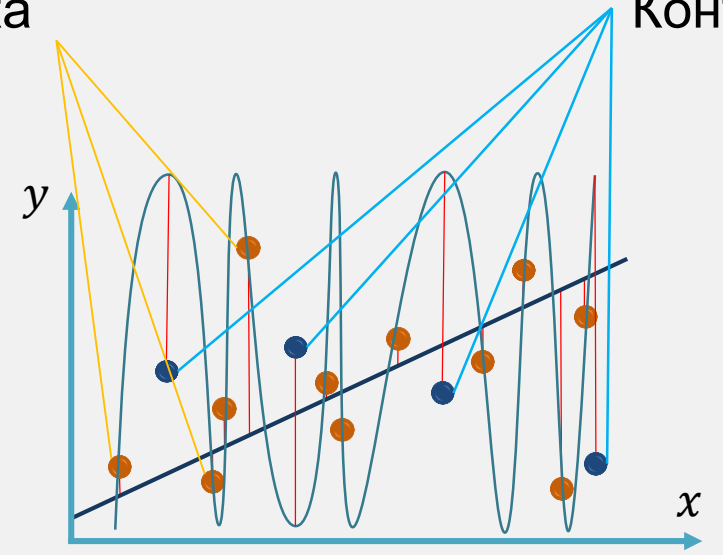
Задача восстановления регрессии

Проблема переобучения

1)Обучающая выборка

2)Контрольная выборка

Обучающая выборка
(Ошибка = 0)



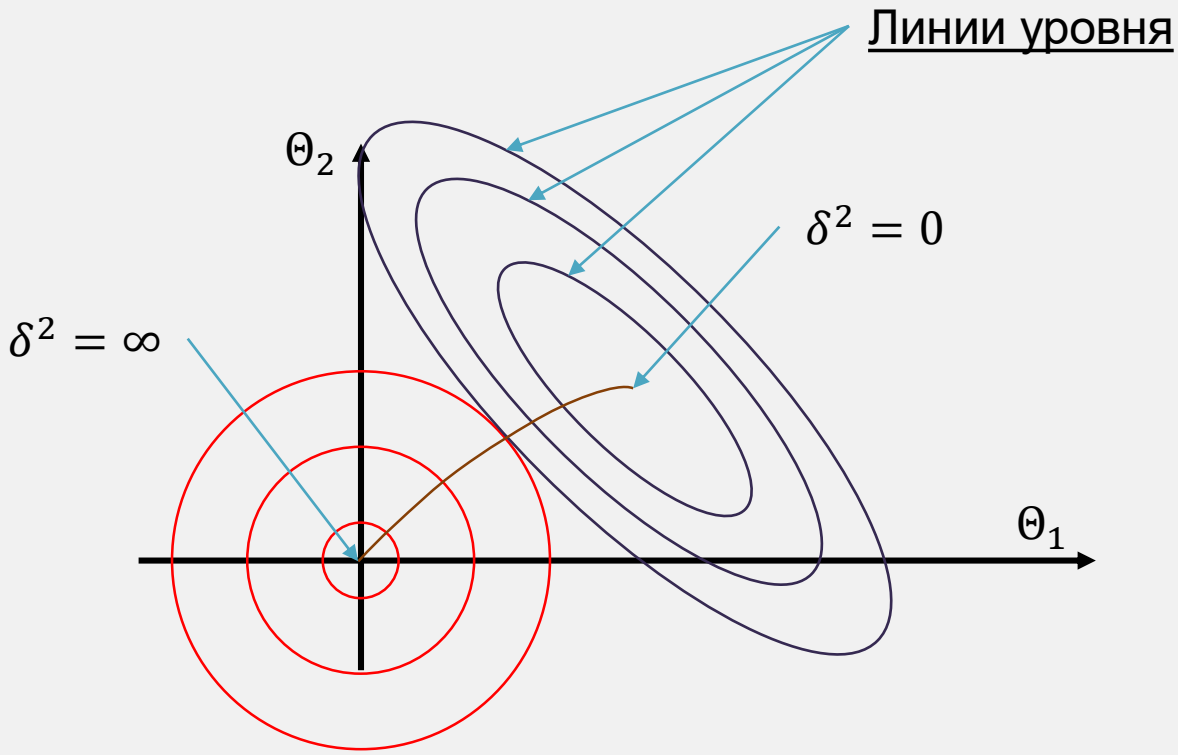
Контрольная выборка

Задача восстановления регрессии

Регуляризация

Целевая функция:

$$J(\Theta) = (y - X\Theta)^T(y - X\Theta) + \delta^2\Theta^T\Theta$$



Задача восстановления регрессии

Регуляризация

L1 регуляризация:

$$J(\Theta) = \sum_j (y_j - \sum_i (x_i^j * \Theta_i))^2 + \delta^2 \sum_i |\Theta_i|$$

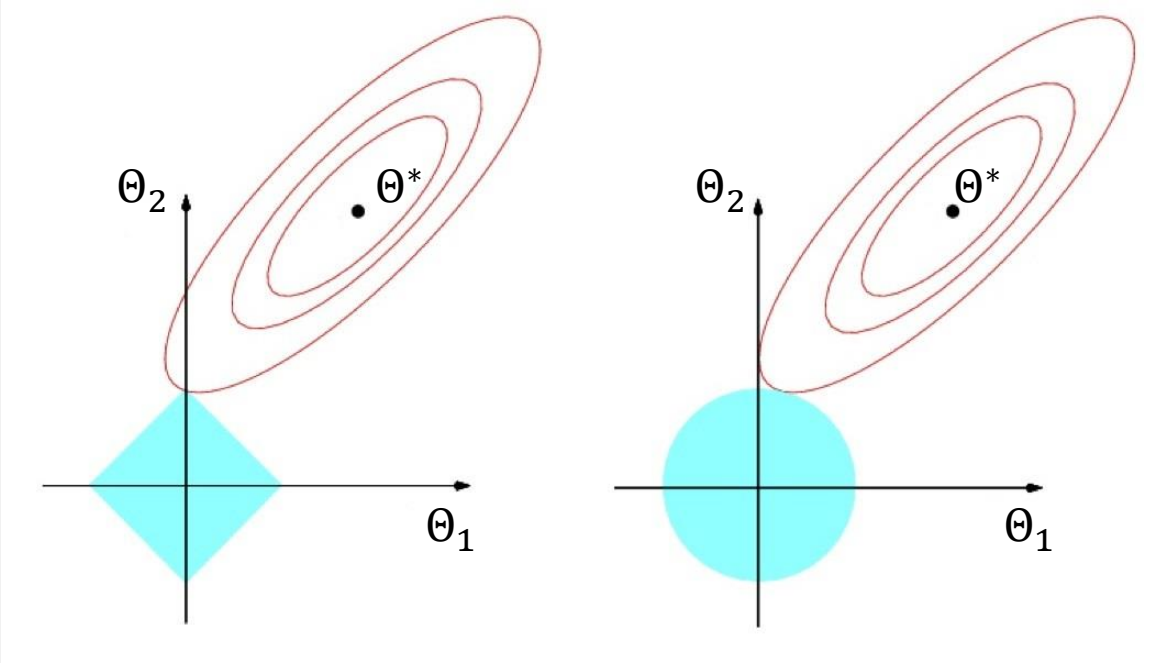
L2 регуляризация или регуляризация Тихонова:

$$J(\Theta) = \sum_j (y_j - \sum_i (x_i^j * \Theta_i))^2 + \delta^2 \sum_i (\Theta_i)^2$$

Elastic Net

$$J(\Theta) = \sum_j (y_j - \sum_i (x_i^j * \Theta_i))^2 + \alpha\rho \sum_i |\Theta_i| + \alpha \frac{(1-\rho)}{2} \sum_i (\Theta_i)^2$$

Задача восстановления регрессии



Регуляризация

L1 регуляризация

L2 регуляризация

Метрики

Регрессия

1. MAE
2. MSE
3. RMSE
4. RMSLE
5. R2

Метрики

Регрессия

MAE

Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{y_i} - y_i|$$

Лучшее константное предсказание - медиана

$\hat{Y}$	Y
0.1	0
0.5	1
0.6	1
0.5	1
0.3	0

$$MAE = \frac{1}{5}(|0.1 - 0| + |0.5 - 1| + |0.6 - 1| + |0.5 - 1| + |0.3 - 0|)$$

Метрики

Регрессия

MSE (Mean Squared Error)

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

$RMSE$ (Root Mean Squared Error)

$$RMSE = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Лучшее константное предсказание - среднее

Метрики

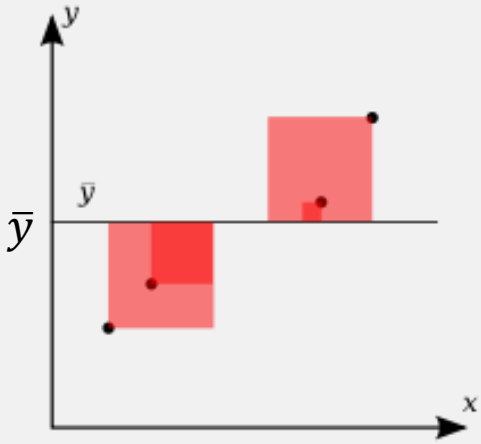
Регрессия

$$R2 - score$$

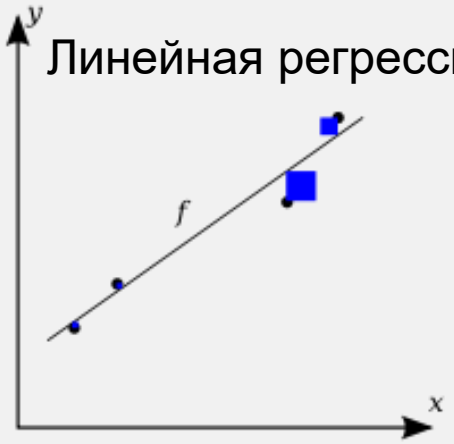
$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$SS_{tot} = \sum_{i=1}^{N} (y_i - \bar{y})^2 \ - variance$$

$$SS_{res} = \sum_{i=1}^{N} (y_i \ - f_i)^2$$



SS_{tot}



SS_{res}

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

ТЕХНОСФЕРА

Спасибо за внимание!

Спасёнов Алексей

a.spasenov@corp.mail.ru