МГТУ
им. Н. Э. Баумана
ТЕХНОПАРК
Mail.Ru Group

ПРОГРАММИРОВАНИЕ CUDA C/C++, АНАЛИЗ ИЗОБРАЖЕНИЙ И DEEP LEARNING

Лекция №7

Спасёнов Алексей

Введение в Машинное обучение

Часть вторая

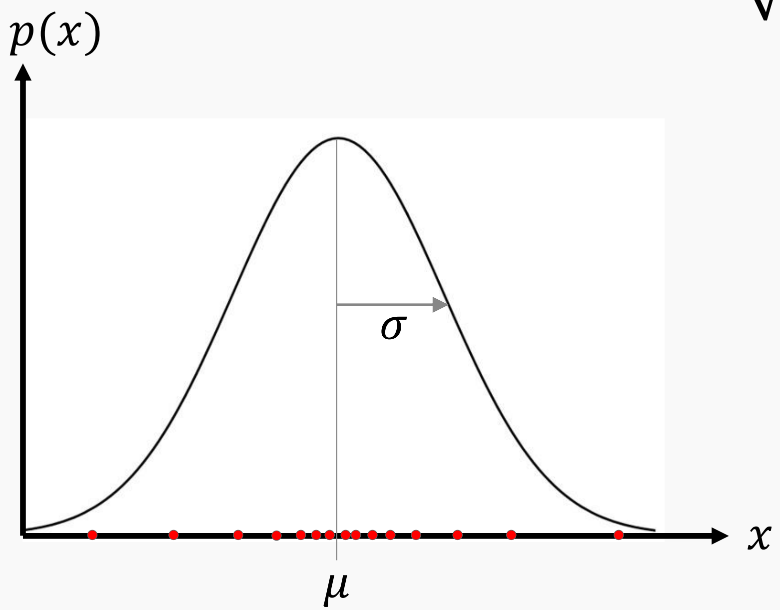
1. Распределение Гаусса
2. Метод максимального правдоподобия
3. Метод максимального правдоподобия для линейной регрессии
4. Распределение Бернулли
5. Энтропия
6. Регуляризация, регуляризации Тихонова, гребневая регрессия
7. Оптимизация, метод градиентного спуска, метод Ньютона
8. Логистическая регрессия

Введение в Машинное обучение

Распределение Гаусса (Нормальное распределение)

Плотность вероятности:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)}$$



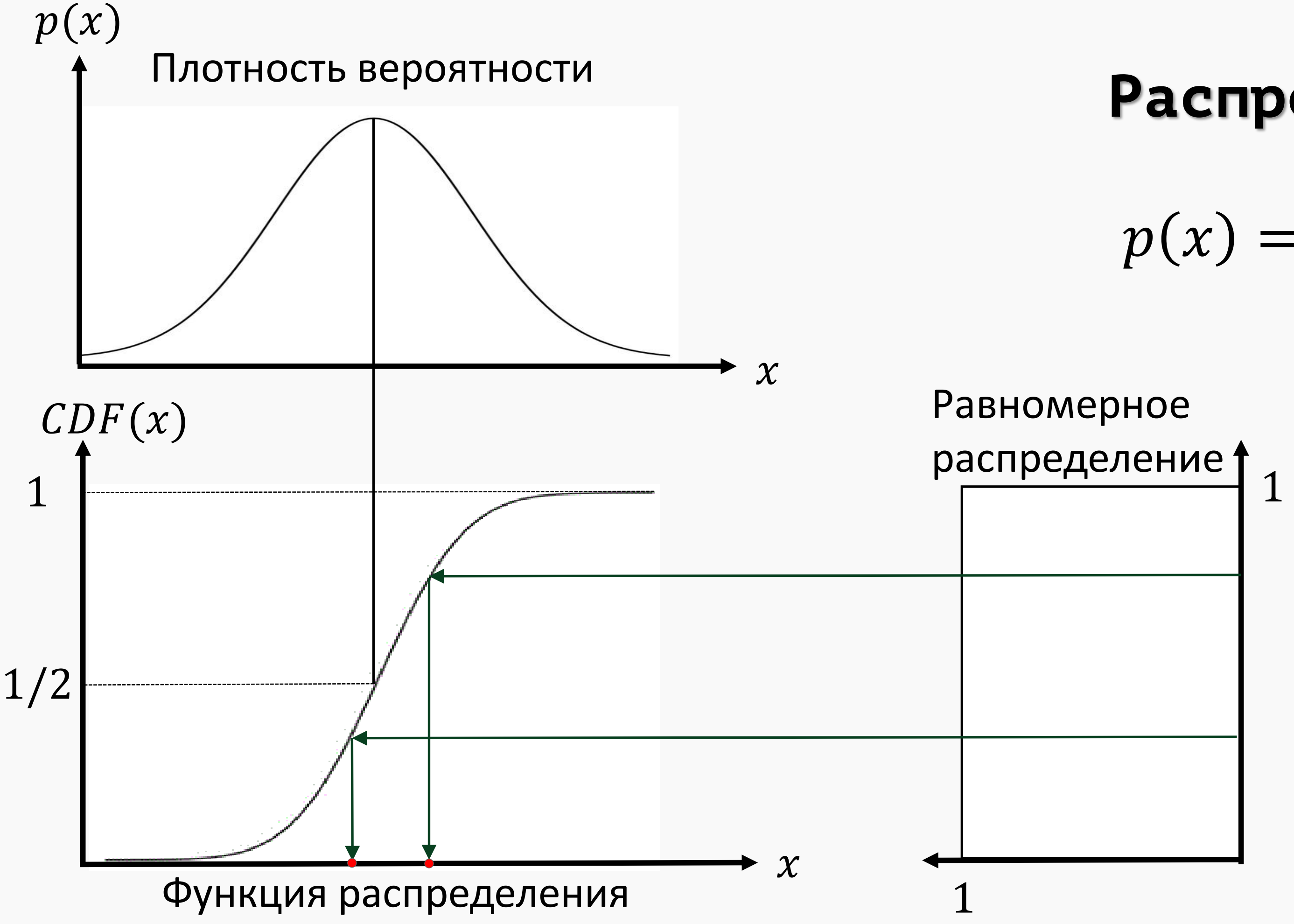
μ - математическое ожидание
(mean or expectation)

σ - среднеквадратическое отклонение
(standard deviation)

σ^2 - дисперсия
(variance)

$$\int p(x)xd = \int_{-\infty}^{+\infty} p(x)xd = 1$$

Введение в Машинное обучение



Распределение Гаусса

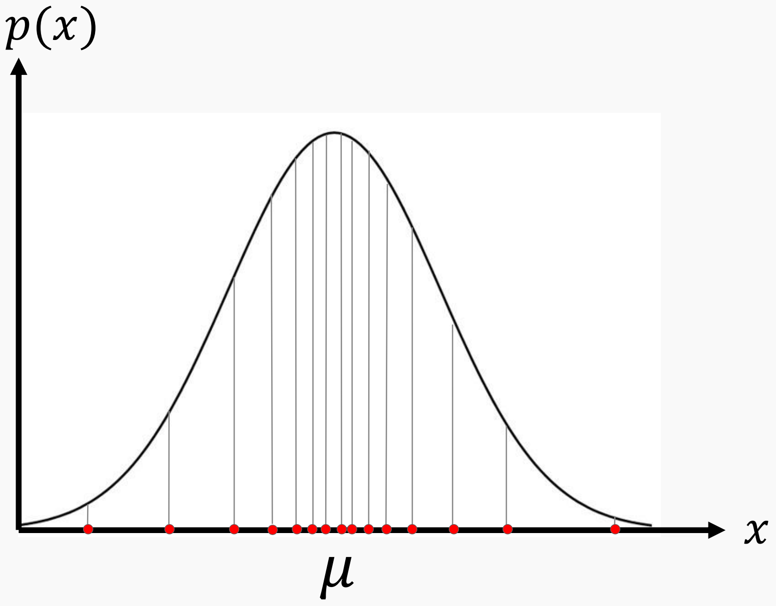
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)}$$

$$N(\mu, \sigma^2) \equiv p(x)$$

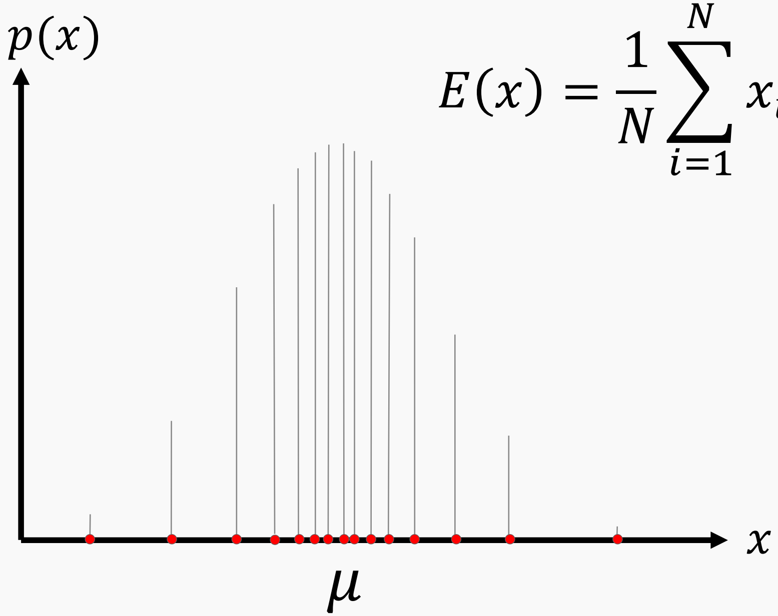
Введение в Машинное обучение

Математическое ожидание

$$E(x) = \int xp(x)dx \ = \mu(\text{для распределения Гаусса})$$



$$E(x) = \frac{1}{N}\sum_{i=1}^{N} x_i$$



Ковариация

Мера линейной зависимости двух случайных величин

$$cov(X,Y) = E[(X - E(X)(Y - E(Y)] = E[X,Y] - E(X)E(Y)$$

$$cov(X,Y) = \frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})(y_t - \bar{y})$$

$$\bar{x} = \frac{1}{N}\sum_{t=1}^{N} x_t \qquad \bar{y} = \frac{1}{N}\sum_{t=1}^{N} y_t$$

$$E[X,Y] = \frac{1}{N}\sum_{t=1}^{N} x_t y_t$$

Если размерность пространства равна d:

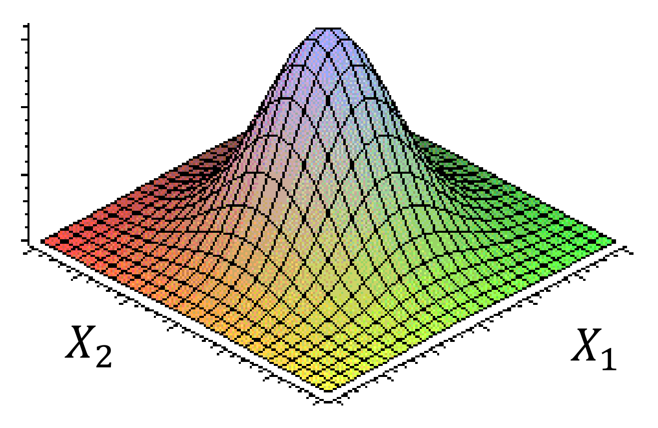
$$cov[X] = E\left[\left(X - E(X)\left(X - E(X)\right)^T\right] =$$

$$\begin{pmatrix} var[X_1] & cov(X_1,X_2) & \dots & cov(X_1,X_d) \\ cov(X_2,X_1) & var(X_2) & \dots & cov(X_2,X_d) \\ \dots & \dots & \ddots & \dots \\ cov(X_d,X_1) & cov(X_d,X_w) & \dots & var(X_d) \end{pmatrix}$$

Многомерное распределение Гаусса

$$\Sigma = cov(X)$$

$$\mathrm{N}(X|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu))$$



Введение в Машинное обучение

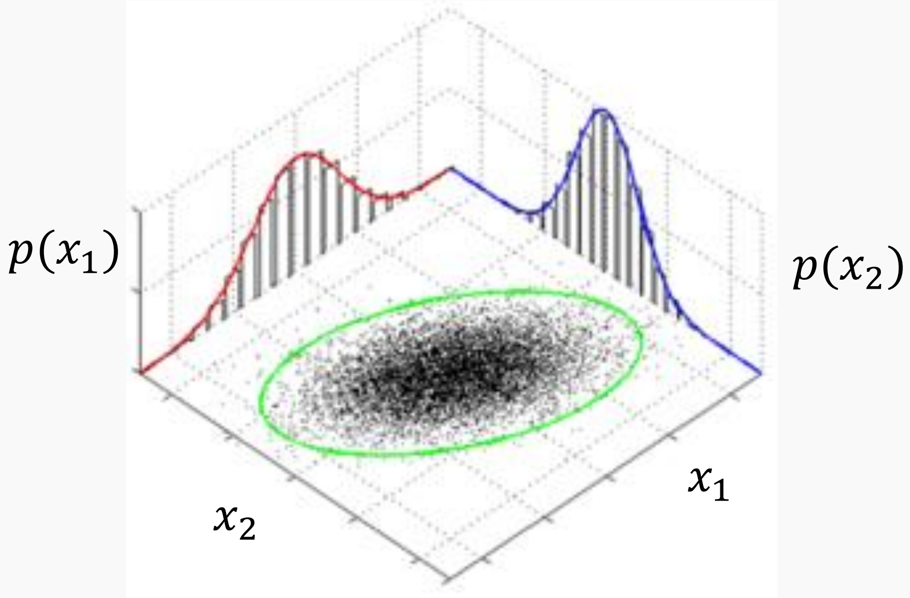
Если имеется 2 независимые переменные

$$x_1 = N(\mu_1, \sigma^2), \quad x_2 = N(\mu_2, \sigma^2)$$

$$p(x_1, x_2) = p(x_1|x_2)p(x_2\ ) = p(x_1\ )p(x_2) =$$

$$= |2\pi\Sigma|^{-1/2} e^{-1/2[X-\mu]^T \Sigma^{-1} [X-\mu]}$$

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$

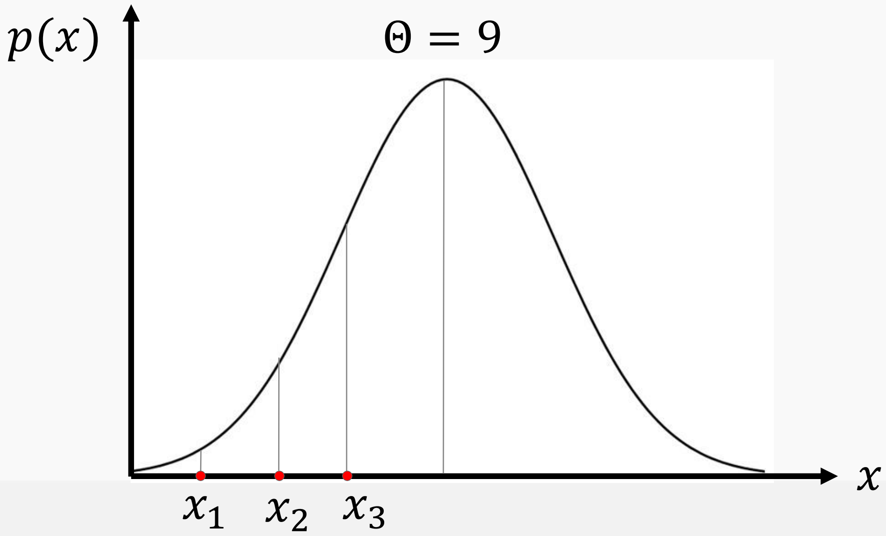


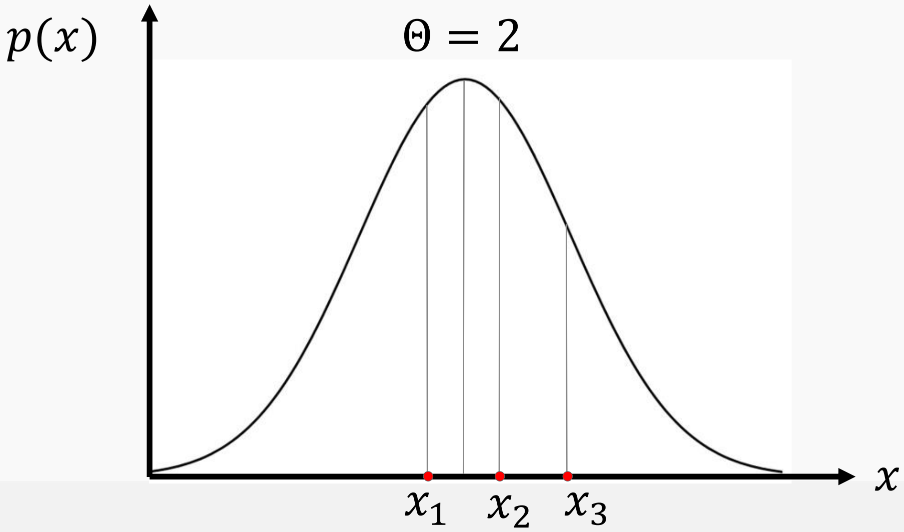
Правдоподобие (Likelihood)

Имеется 3 независимые величины:

$x_1 = 1, \; x_2 = 3, \; x_3 = 5$

Необходимо найти такое Θ для $N(\Theta, \sigma^2)$, $\sigma^2 = 1$, чтобы $p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3)$ было максимальным.





Введение в Машинное обучение

Правдоподобие (Likelihood) для линейной регрессии

Предполагаем, что величина y_i принадлежат нормальному распределению с математическим ожиданием ${x_i}^T\Theta$ и дисперсией σ^2, записанным в виде:

$$y_i = N({x_i}^T, \Theta, \sigma^2) = {x_i}^T\Theta + N(0, \sigma^2)$$

Тогда:

$$p(y|{x_i}^T, \Theta, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - {X_i}^T\Theta)^2}$$

Целевая функция

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

Метод максимального правдоподобия (Maximum Likelihood)

Будем максимизировать правдоподобие (Likelihood)

$$p(y|{x_i}^T, \Theta, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - {X_i}^T\Theta)^2}$$

Возьмём логарифм от правдоподобия (likelihood). Положение глобального максимума не изменится, поскольку функция логарифма является монотонно возрастающей.

$$\log(\Theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\Theta)^T(y - X\Theta)$$

Для решения задачи минимизации умножим $\log(\Theta)$ на (-1)

$$\text{neg } \log(\Theta) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(y - X\Theta)^T(y - X\Theta)$$

Распределение Бернулли

$$p(X|\Theta) = \begin{cases} \Theta, & \text{если } x = 1 \\ 1 - \Theta, & \text{если } x = 0 \end{cases}$$

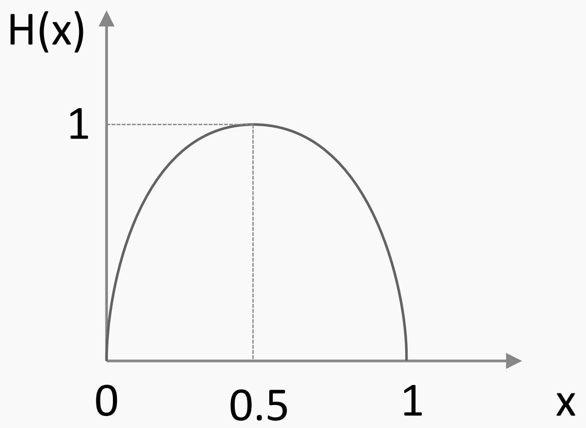
Параметр $\Theta \in (0,1)$

$$p(X|\Theta) = \Theta^X(1-\Theta)^{1-X} = \begin{cases} \Theta, & \text{если } x = 1 \\ 1 - \Theta, & \text{если } x = 0 \end{cases}$$

Энтропия

Мера неопределённости случайной величины

$$\mathrm{H}(X) = -\sum_{x} p(X|\Theta) \log p(X|\Theta)$$



Для распределения Бернулли:

$$\mathrm{H}(X) = -\sum_{x=0}^{1} \Theta^{X}(1-\Theta)^{1-X} \log[\Theta^{X}(1-\Theta)^{1-X}] = -[(1-\Theta)\log(1-\Theta) + (\Theta)\log(\Theta)]$$

Регуляризация

Предположим, что найдены параметры модели:

$$\overline{\Theta} = (X^T X)^{-1} X^T Y$$

Задачу поиска Y можно интерпретировать как поиск решения СЛАУ.

Матрица может быть плохо обусловлена.

Решение: добавим элементы на диагональ:

$$\overline{\Theta} = (X^T X + \delta^2 I_d)^{-1} X^T Y$$

Регуляризация (Регуляризации Тихонова)

Целевая функция

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

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

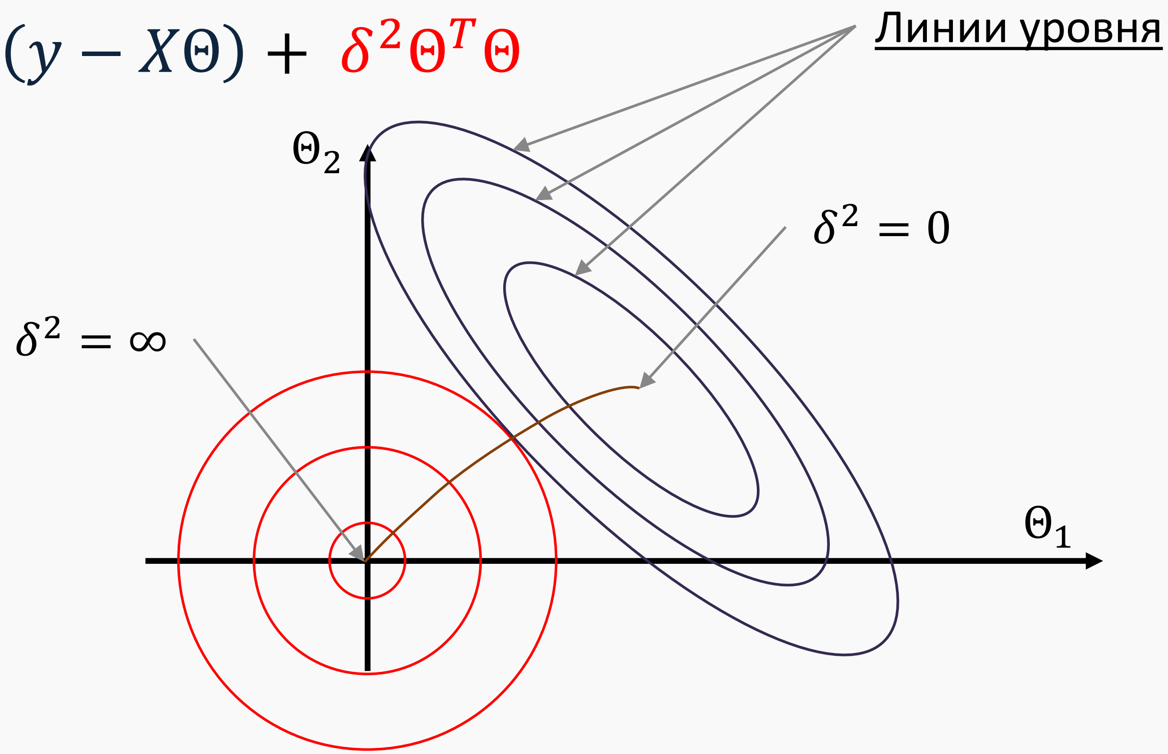
Получаем:

$$\overline{\Theta}_{ridge} = (X^T X + \delta^2 I_d)^{-1} X^T Y$$

Регуляризация

Целевая функция

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



Введение в Машинное обучение

Оптимизация

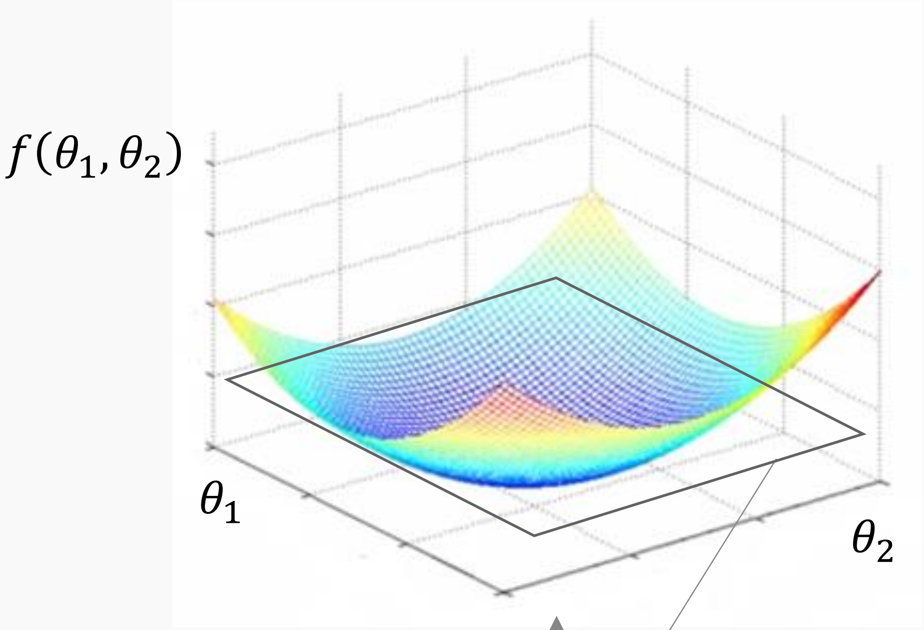
Функция:

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

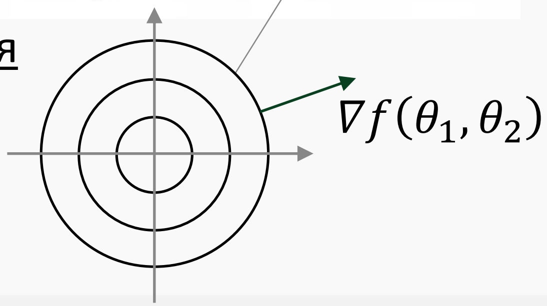
Частные производные:

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

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



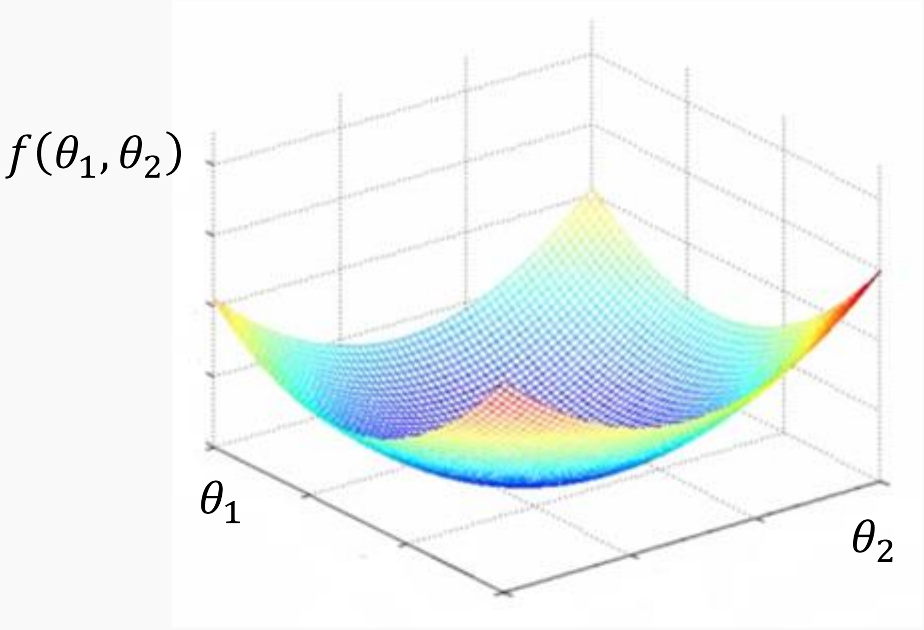
Линии уровня



Оптимизация

Функция:

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



Гессиан функции

$$\frac{\delta f(\theta_1,\theta_2)}{\delta\theta_1\delta\theta_1} = 2, \frac{\delta f(\theta_1,\theta_2)}{\delta\theta_1\delta\theta_2} = 0$$

$$\frac{\delta f(\theta_1,\theta_2)}{\delta\theta_2\delta\theta_1} = 0, \frac{\delta f(\theta_1,\theta_2)}{\delta\theta_2\delta\theta_2} = 2$$

$\mathrm{H} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ - Матрица Гессе

Оптимизация

Линейная модель:

$$y(X_i) = \Theta_1 + x_i\Theta_2$$

Целевая функция:

$$J(\Theta) = \sum_{i=1}^{n}(y_i - \Theta_1 - x_i\Theta_2)^2 + \delta^2\Theta^T\Theta$$

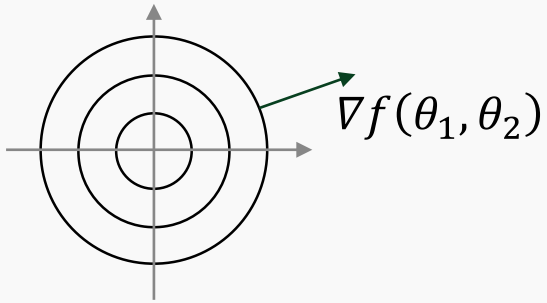
$$\nabla J(\Theta) = \begin{bmatrix} \sum_{i=1}^{n} 2(y_i - \Theta_1 - x_i\Theta_2) * (-1) \\ \sum_{i=1}^{n} 2(y_i - \Theta_1 - x_i\Theta_2) * (-x_i) + 2\delta^2\Theta_1 \end{bmatrix}$$

Для сложных функций глобальный минимум найти достаточно сложно

Метод градиентного спуска

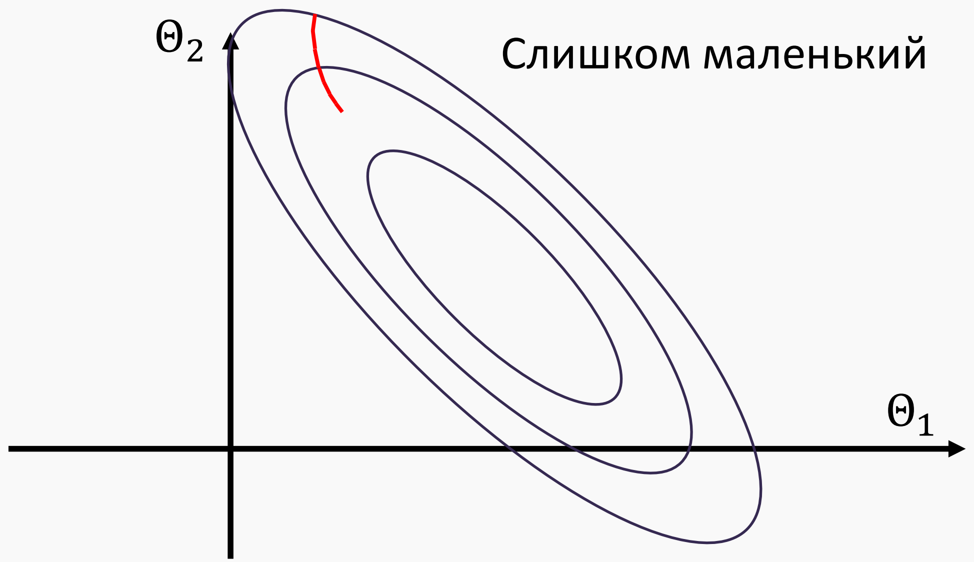
$$\Theta_{k+1} = \Theta_k - \eta_k g_k = \Theta_k - \eta_k \nabla f(\theta_k)$$

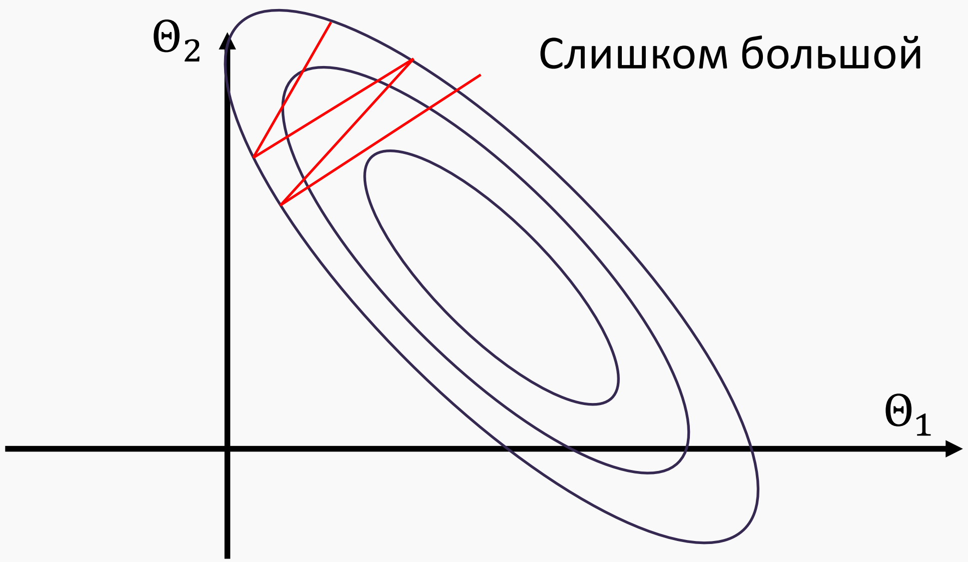
Коэффициент обучения
(learning rate)



Метод градиентного спуска

Как выбрать η_k?





Метод Ньютона

Используем матрицу Гессе

$\Theta_{k+1} = \Theta_k - H_k^{-1} g_k$

$$f_{quad}(\Theta) = f(\Theta_k) + g_k^T(\Theta - \Theta_k) + \frac{1}{2}(\Theta - \Theta_k)^T H_k(\Theta - \Theta_k)$$
$$\frac{\partial}{\partial\Theta} f_{quad}(\Theta) = 0 + g_k^T + H_k(\Theta - \Theta_k) = 0$$

Введение в Машинное обучение

Типы обучения

1) Offline обучения

$\Theta_{k+1} = \Theta_k - \eta_k \sum_{i=1}^{n} X_i^T (y_i - X_i \Theta_k)$, где n - размер выборки

2) Online обучения

$\Theta_{k+1} = \Theta_k - \eta_k X_k^T (y_k - X_k \Theta_k)$

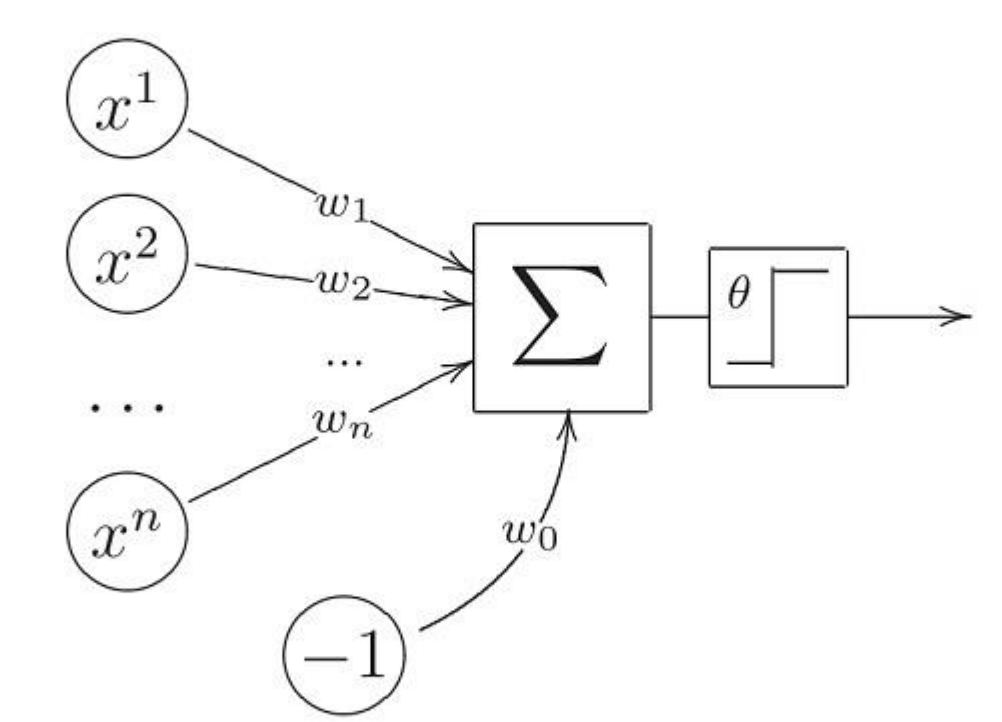
3) Обучение минибатчами

$\Theta_{k+1} = \Theta_k - \eta_k \sum_{j=1}^{20} X_j^T \left(y_j - X_j \Theta_k\right)$, где 20 - размер минибатча

Введение в Машинное обучение

Нейрон МакКаллока-Питтса

Решаем задачу классификации



$$y(X) = f\Big(\sum_{i=1}^{n}(w_i x^i) - w_0\Big)$$

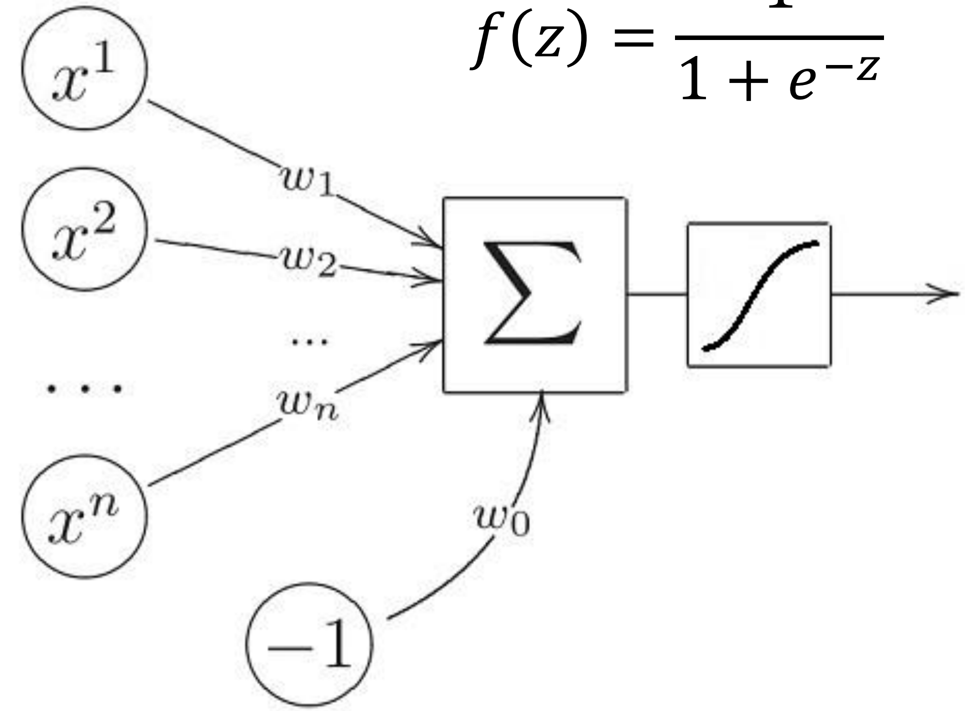
$f(z)$- ступенчатая функция Хевисайда.

Модель МакКаллока-Питтса эквивалентна пороговому линейному классификатору

Логистическая регрессия

В качестве функции $f(z)$ возьмём функцию:

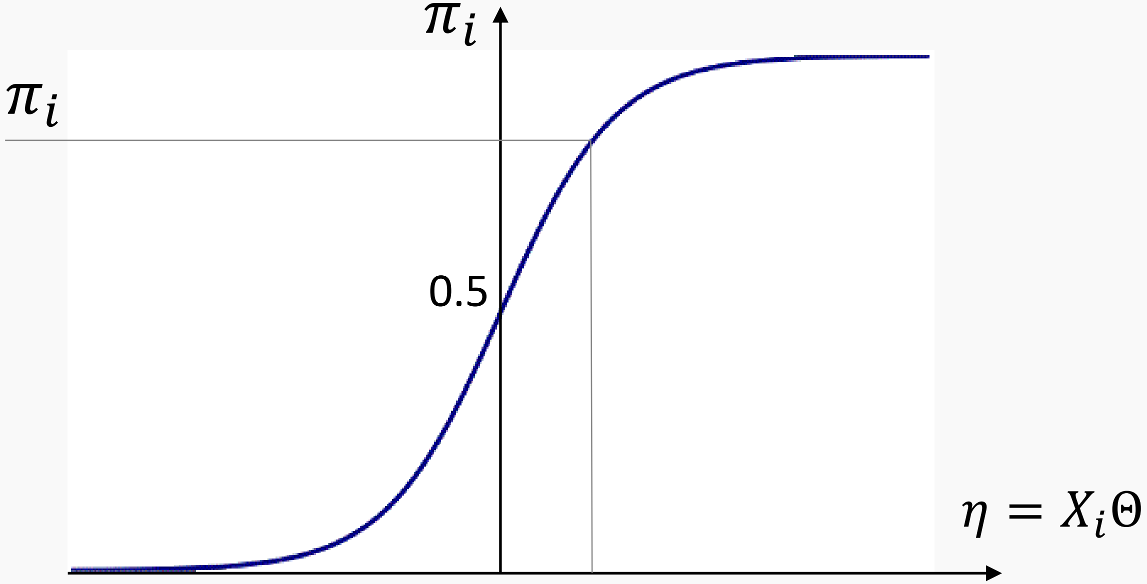
$$f(z) = \frac{1}{1 + e^{-z}}$$



Функция сигмоида

$$sigm(\eta) = \frac{1}{1 + e^{-\eta}} = \frac{e^{\eta}}{e^{\eta} + 1}$$

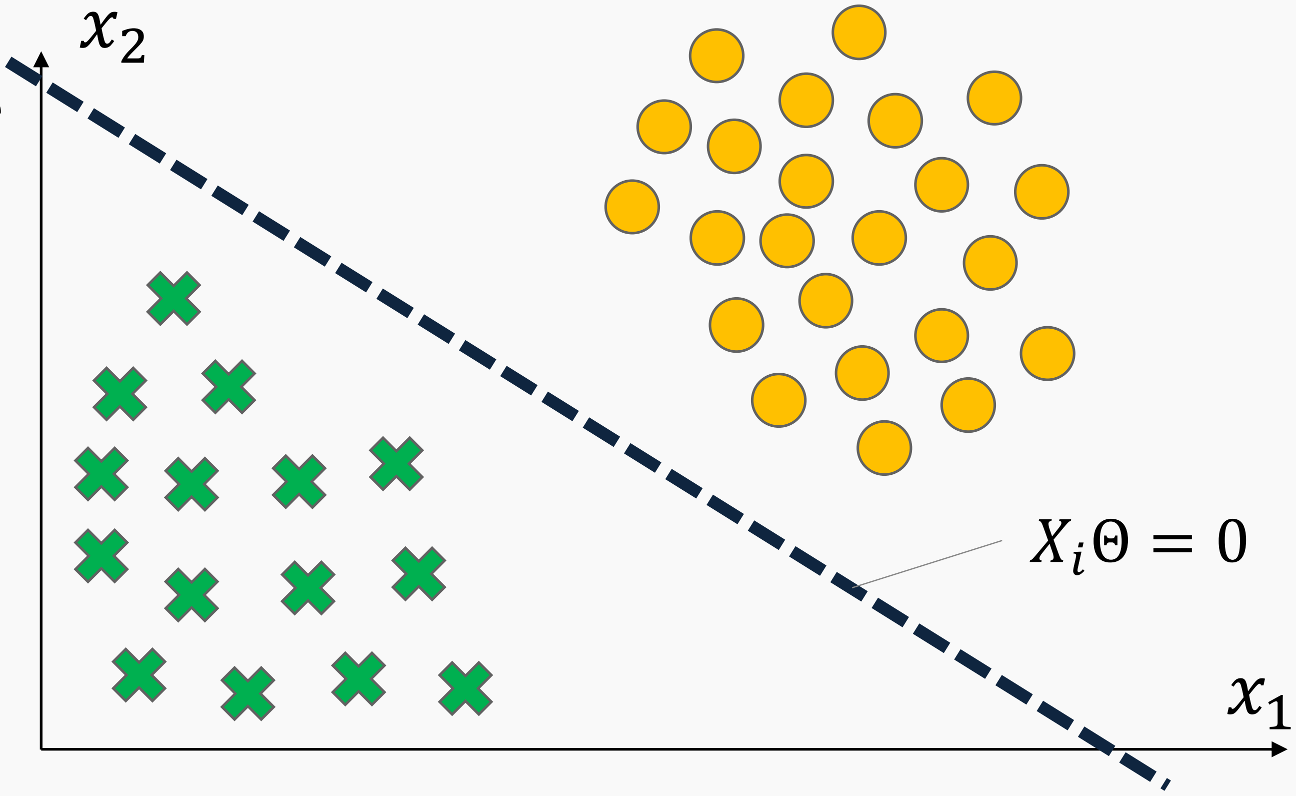
$$p(y_i = 1 | X_i \Theta) = \pi_i$$



Введение в Машинное обучение

Функция сигмоида

Линейное разделение гиперплоскостью



Распределение Бернулли

$$p(X|\Theta) = \begin{cases} \Theta, & \text{если } x = 1 \\ 1 - \Theta, & \text{если } x = 0 \end{cases}$$

Параметр $\Theta \in (0,1)$

$$p(X|\Theta) = \Theta^X(1-\Theta)^{1-X} = \begin{cases} \Theta, & \text{если } x = 1 \\ 1 - \Theta, & \text{если } x = 0 \end{cases}$$

Введение в Машинное обучение

Для логистической регрессии

$$p(y|X\Theta) = \prod_{i=1}^{n} Ber(y_i|sigm(X_i\Theta)) = \prod_{i=1}^{n} \left[\frac{1}{1+e^{-X_i\Theta}}\right]^{y_i} \left[1 - \frac{1}{1+e^{-X_i\Theta}}\right]^{1-y_i},$$

где $X_i\Theta = \Theta_0 + \sum_{j=1}^{d} \Theta_j x_j$.

$$C(\Theta) = -\log\big(p(y|X\Theta)\big) = -\sum_{i=1}^{n} (y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i))$$

Для логистической регрессии

Поиск параметров модели

Целевая функция:

$$J(\Theta) = -\log\big(p(y|X\Theta)\big)$$

Градиент:

$$g(w) = \frac{\partial}{\partial\Theta}J(\Theta) = \sum_{i=1}^{n} X_i^T(\pi_i - y_i) = X^T(\pi - y)$$

Матрица Гессе:

$$H = \frac{\partial}{\partial\Theta}g(w) = \sum_{i=1}^{n} \pi_i(1-\pi_i)X_iX_i^T = X^T diag\big(\pi_i(1-\pi_i)\big)X$$

Ищем решение любым известным методом оптимизации

МГТУ им. Н.Э. Баумана
ТЕХНОПАРК
Mail.Ru Group

Контакты:
a.spasenov@corp.mail.ru
alex_spasenov (Skype)

Спасибо за внимание!