



Методы машинного обучения

Экзамен

Классификация:

1. supervised
2. unsupervised
3. reinforced
4. ансамбльов

32.09

Линейная регрессия

$$\begin{matrix} X \rightarrow Y \\ \text{вход} \quad \uparrow \quad \text{выход} \end{matrix}$$

n - кол-во примеров

d - размерность входа

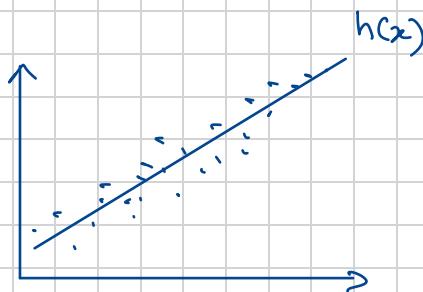
$(x^{(i)}, y^{(i)})$ - i -ий пример из выборки

$$x^{(:)} = (x_1^{(:)}, x_2^{(:)}, \dots, x_d^{(:)})^T$$

$y = \begin{cases} \text{категориальное значение (классификация)} \\ \text{непрерывное число (регрессия)} \end{cases}$

Обучение с учителем

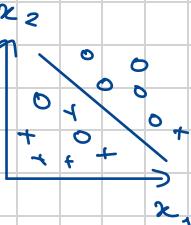
Пример регр	
параметр	f
искусственная выборка dataset	2304
	1600
	2400
	...
	400
	330
	369
	...



Задание: найти "линейку" $h(x)$ на y

Пример бинарной классификации

x_1	x_2	y
-	-	0
-	-	0
-	1	1
-	0	0
1	-	1
⋮	⋮	⋮



Задача: найти решающую границу, такую, что на одн. стороны находятся +, а на другой - 0

Учебная выборка

$$\{x^{(i)}, y^{(i)}\}, i=1$$

аналогично М.О.



$$\bar{x} \in \mathbb{R}^d, y \in \mathbb{R}$$

$$h_{\theta}(\bar{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_d x_d$$

$$h_{\theta}(\bar{x}) = \theta_0 + \sum_{i=1}^d \theta_i x_i = \sum_{i=0}^d \theta_i x_i = \overline{\theta}^T \bar{x}$$

$$\overline{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}$$

Добавим прямойной признак $x_0=1$

стандартное уменьшение векторов

intercept term

$$J(\overline{\theta}) = \frac{1}{2} \sum_{i=1}^n (h_{\theta}(\bar{x}^{(i)}) - y^{(i)})^2 - SE \text{ (Squared error)}$$

функция потерь (loss)
стоимость (cost)

$$y_2(\bar{\theta}) = \frac{1}{n} y(\bar{\theta}) - \text{MSE}_{\text{(mean SE)}}$$

Средний квадратический остаток

$$y_2(\bar{\theta}) = \sqrt{y_2(\bar{\theta})} - \text{RMSE}$$

Среднеиздатческая ошибка

$$\hat{\theta} \in \arg \min_{\theta} \bar{y}(\theta)$$

Справки
Сテンография

$$A_k(x) = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^k}$$

$$\begin{aligned} A_0(x) &= \text{среднее} \\ &\quad \text{геометрическое} \\ A_1(x) &= \text{среднее} \\ &\quad \text{апериодическое} \\ A_2(x) &= \text{MSE} \end{aligned}$$

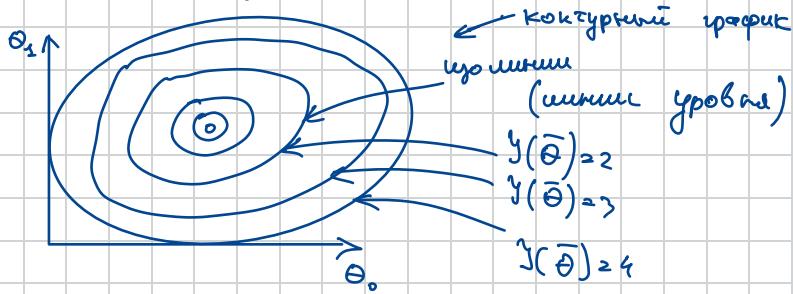
$A_3(x) \geq$ ~~когда~~ кубическое

$A_{-s}(z)$ - среднее
горизонтальное

$$A_{+\infty}(x) = \max\{x_s\}$$

$$A_{-\infty}(x) = \min \{x_1\}$$

Гражданский спуск



Amoratum

2. not. go through:
2. get bcse $j \in \{0, 1, \dots, d\}$ homework now

$$\theta_j^{(i+1)} := \theta_j^{(i)} - \alpha \frac{\partial}{\partial \theta_i} y(\theta^{(i)})$$

Ири съвбасие

d-зинер параметр иоден

Чи є векторної форми

1) //

$$2) \quad \bar{\Theta}^{(i+1)} := \bar{\Theta}^{(i)} - \alpha \nabla_{\bar{\Theta}} y(\bar{\Theta}^{(i)})$$

скорость обучения
(запоминания слова)

Характеристика:

$$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots, \theta^{(t)}, \theta^{(t+1)}, \dots$$

$$\text{нока } \|\theta^{(t+1)} - \theta^{(t)}\| > \varepsilon$$

$$\text{нока } |y(\theta^{(t+1)}) - y(\theta^{(t)})| > \varepsilon$$

$$\text{нока } \|\nabla_{\theta} y(\theta^{(t)})\| >$$

$$\begin{aligned}\bar{\theta}^{(t+1)} &:= \bar{\theta}^{(t)} - \alpha \nabla_{\theta} y(\bar{\theta}^{(t)}) = \bar{\theta}^{(t)} - \alpha \nabla_{\theta} \left[\frac{1}{2} \sum_{i=1}^n (h_{\theta}(\bar{x}^{(i)}) - y^{(i)})^2 \right] = \\ &= \bar{\theta}^{(t)} - \alpha \nabla_{\theta} \left[\frac{1}{2} \sum_{i=1}^n ((\bar{\theta}^{(t)})^T \cdot \bar{x}^{(i)} - y^{(i)})^2 \right] = \bar{\theta}^{(t)} - \alpha \sum_{i=1}^n ((\bar{\theta}^{(t)})^T \cdot \bar{x}^{(i)} - \bar{y}^{(i)}) \cdot \bar{x}^{(i)}\end{aligned}$$

$$\frac{\partial (\bar{\theta}^T \bar{x})}{\partial \theta_i} = \frac{\partial (\theta_0 x_0 + \dots + \theta_d x_d)}{\partial \theta_i} = x_i$$

$$\nabla_{\theta} (\bar{\theta}^T \bar{x}) = \left(\frac{\partial \bar{\theta}^T \bar{x}}{\partial \theta_0}, \dots, \frac{\partial \bar{\theta}^T \bar{x}}{\partial \theta_d} \right) = (x_0, \dots, x_d) = \bar{x}$$

Stochastic GD (с ранд. выбор. счётах)

некоторые градиентные $i \in \{1, 2, \dots, n\}$ генер.

$$\bar{\theta}^{(t+1)} = \bar{\theta}^{(t)} - \alpha ((\bar{\theta}^{(t)})^T \cdot \bar{x}^{(i)} - \bar{y}^{(i)}) \bar{x}^{(i)}$$

$y(\bar{\theta}) = \frac{1}{2} \sum_{i=1}^n (h_{\theta}(\bar{x}^{(i)}) - \bar{y}^{(i)})^2$, где i -мн. выбор. равномерно одно из знач. $\{1, 2, \dots, n\}$

математическая модель

+ матрицы

Mflik

Нормальные уравнения

$$X = \begin{bmatrix} -x_0^{(1)} & \dots & -x_0^{(n)} \\ -x_1^{(1)} & \dots & -x_1^{(n)} \\ \vdots & \ddots & \vdots \\ -x_d^{(1)} & \dots & -x_d^{(n)} \end{bmatrix}, \bar{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \theta = \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_d \end{bmatrix} \Rightarrow X \cdot \bar{\theta} - \bar{y} = \begin{bmatrix} \bar{\theta}^T \cdot x_0^{(1)} - y^{(1)} \\ \vdots \\ \bar{\theta}^T \cdot x_0^{(n)} - y^{(n)} \end{bmatrix}$$

Design matrix

$$a^T b = b^T a$$

$$\mathbb{E}(\bar{\theta}) = \frac{1}{2} (x \cdot \bar{\theta} - \bar{y})^T \cdot (x \cdot \bar{\theta} - \bar{y})$$

$$\boxed{\nabla_{\theta} \mathbb{E}(\bar{\theta}) = \bar{\theta}} \quad \nabla_{\theta} \frac{1}{2} (x \cdot \bar{\theta} - \bar{y})^T (x \cdot \bar{\theta} - \bar{y}) = \nabla_{\theta} \frac{1}{2} ((x \cdot \bar{\theta})^T \cdot (x \cdot \bar{\theta})) - 2 (x \cdot \bar{\theta})^T \bar{y} +$$

$$+ \bar{y}^T \cdot \bar{y}) = \nabla_{\theta} \frac{1}{2} [\bar{\theta}^T x^T \cdot x \cdot \bar{\theta} - 2 \bar{\theta}^T x^T \bar{y} + \bar{y}^T \bar{y}] = \frac{1}{2} [2(x^T \cdot x) \cdot \bar{\theta} - 2x^T \bar{y}] = 0 \\ \Rightarrow (x^T \cdot x) \cdot \bar{\theta} = x^T \bar{y}$$

$$\bar{\theta} = (x^T x)^{-1} \cdot x^T \bar{y} \text{ - normal equation}$$

При предположении, что $y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}(\omega)$, где $\varepsilon^{(i)}(\omega) \sim N(0, \sigma^2)$

$$\text{Тогда } p(y^{(i)} | x^{(i)}, \theta, \sigma^2) = N(\theta^T x^{(i)}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2} \frac{(y^{(i)} - \theta^T x^{(i)})^2}{\sigma^2}\right]$$

Воспользуемся MLE: $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^n$; $\boxed{\theta? \sigma^2?}$

$$p(\mathcal{D} | \theta, \sigma^2) = p(y^{(1)}, \dots, y^{(n)} | x^{(1)}, \dots, x^{(n)}; \theta, \sigma^2) \underset{\theta, \sigma^2}{\rightarrow} \max$$

$$L(\theta, \sigma^2) = p(\bar{y} | \bar{x}; \theta, \sigma^2) = \prod_{i=1}^n p(y^{(i)} | x^{(i)}; \theta, \sigma^2) \text{ - функция правдоподобия}$$

$$l(\theta, \sigma^2) = \ln L(\theta, \sigma^2) = \ln \prod_{i=1}^n p(y^{(i)} | x^{(i)}; \theta, \sigma^2) = \text{ - логарифм вероятности по } y^{(i)}$$

$$\approx \sum_{i=1}^n \ln \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2} \frac{(y^{(i)} - \theta^T x^{(i)})^2}{\sigma^2}\right] \right\} = -\frac{1}{2\sigma^2} \sum_{i=1}^n (y^{(i)} - \theta^T x^{(i)})^2 - \frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2$$

$$\text{MLE}(\theta) = LS(\theta)$$

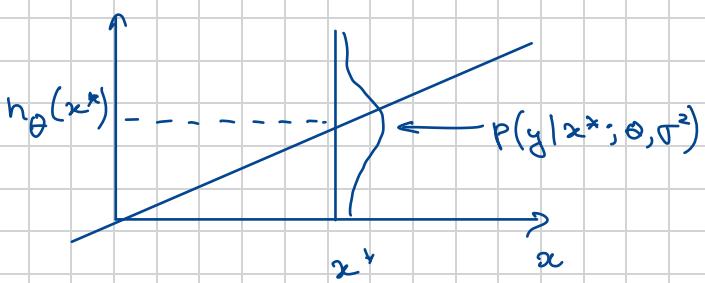
$$\boxed{\max_{\theta} L(\theta, \sigma^2) = \min_{\theta} \frac{1}{2} \sum_{i=1}^n (y^{(i)} - \theta^T x^{(i)})^2}$$

$$\sigma^2 = ?$$

$$\frac{\partial L(\theta, \sigma^2)}{\partial \sigma^2} = \frac{1}{2(\sigma^2)^2} \cdot \sum_{i=1}^n (y^{(i)} - \theta^T x^{(i)})^2 - \frac{n}{2\sigma^2} = 0$$

$$\Rightarrow \boxed{\sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^n (y^{(i)} - \theta^T x^{(i)})^2}$$

$$h_{\theta}(x) = E[y | x; \theta, \sigma^2]$$



Интерпретация через линейную модель

$$J(\theta) = \frac{1}{2} (X \cdot \bar{\theta} - \bar{y})^T (X \cdot \bar{\theta} - \bar{y})$$

$$\min_{\theta} J(\theta)$$

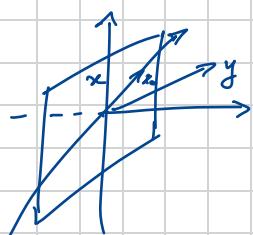
$$J(\theta) \leq 0 \Rightarrow \underset{6 \text{ строк}}{X \cdot \bar{\theta} - \bar{y} = 0} \Rightarrow \boxed{X \cdot \bar{\theta} = \bar{y}} \quad \text{ЧУДО!!!}$$

одно единственное решение

$$X \cdot \hat{\theta} = \hat{y} \Rightarrow y - \hat{y} \text{ будет минимальна!}$$

$$\min J(\theta)$$

$$\text{Пусть } X = \begin{pmatrix} x_{10} & x_{11} \\ x_{20} & x_{21} \\ x_{30} & x_{31} \\ x_0 & x_1 \end{pmatrix} \quad \bar{\theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} \quad \bar{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$



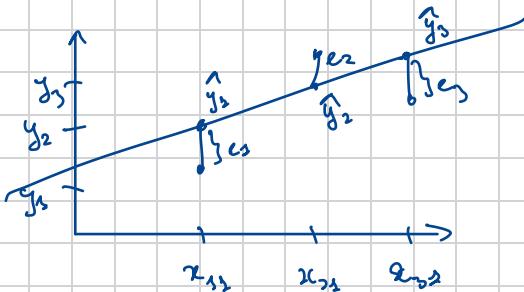
$$X \cdot \bar{\theta} = \bar{x}_0 \cdot \theta_0 + \bar{x}_1 \cdot \theta_1 = \bar{y}$$

$$\hat{y} = P_y$$

$$X \cdot \bar{\theta} = \hat{y} \Rightarrow P_y = \underbrace{X(X^T X)^{-1} X^T}_{P} y$$

$$X \cdot \bar{\theta} = X \cdot (X^T X)^{-1} X^T y \quad P$$

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



$$\|\bar{e}\|^2 = e_1^2 + e_2^2 + e_3^2$$

$$h_{\theta}(\bar{x}) = \bar{\theta}^T \bar{x}$$

$$\bar{X} = \begin{bmatrix} -x^{(1)T} \\ -x^{(2)T} \\ \vdots \\ -x^{(n)T} \end{bmatrix}$$

$$\bar{X}\bar{\theta} \approx \bar{y} \Rightarrow \hat{\theta} = \underset{\theta}{\operatorname{argmin}} J(\theta) = \underset{\theta}{\operatorname{argmin}} \frac{1}{2} \|\bar{X}\bar{\theta} - \bar{y}\|^2 =$$

$$= \frac{1}{2} \sum_{i=1}^n (\bar{\theta}^T \bar{x}^{(i)} - \bar{y}^{(i)})^2$$

множ. прием:

- геом. смысл
- стох. геом. смысл (\rightarrow GP)
- минимизация

коэф. скорости
обучения
оценка
ошибки

$$\bar{\theta}^{(t+1)} := \bar{\theta}^{(t)} - \alpha \nabla_{\theta} J(\bar{\theta}) = \bar{\theta}^{(t)} - \alpha (\bar{\theta}^T \bar{x}^{(i)} - \bar{y}^{(i)}) \cdot \bar{x}^{(i)}$$

$$\theta^{(t+1)} = \theta^{(t)} + \alpha (y^{(i)} - \theta^T x^{(i)}) \cdot x^{(i)}$$

$$\frac{1}{2} \sum_{i=1}^n (\bar{\theta}^T \bar{x}^{(i)} - \bar{y}^{(i)})^2$$

Аналитическое решение

$$(X \cdot X^T) \cdot \bar{\theta} = X^T \bar{y} \Rightarrow \boxed{\bar{\theta} = (X \cdot X^T)^{-1} \cdot X^T \cdot \bar{y}}$$

координатные векторы
нормализ.

$$ax + by + c = 0$$

$$x \cos \alpha + y \sin \alpha + c = 0$$

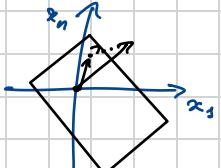
$$X^T \cdot X \cdot \bar{\theta} - X^T \bar{y} = 0 \Rightarrow X^T (X \cdot \bar{\theta} - \bar{y}) = 0$$

$$\bar{y} = \underbrace{\bar{\theta}^T \bar{x}}_{\text{сигнал}} + \underbrace{\varepsilon(\omega)}_{\text{шум}}, \quad \varepsilon(\omega) \sim N(0, \sigma^2)$$

$$\text{MLE} \stackrel{\text{аналог}}{\Leftrightarrow} LS$$

$$NLL \stackrel{\text{аппроксимация}}{\Leftrightarrow} SE$$

$$X\bar{\theta} \approx \bar{y}$$



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



$$y = \theta_0 + \theta_1 x + \theta_2 x^2$$

\downarrow
 x_1

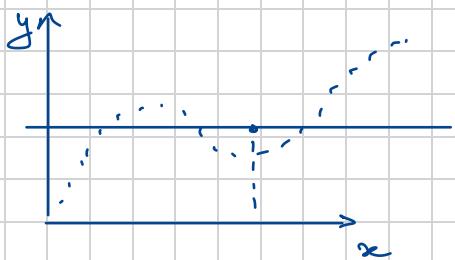
\downarrow
 x_2

$$y = \bar{\Theta}^T \bar{x}'$$

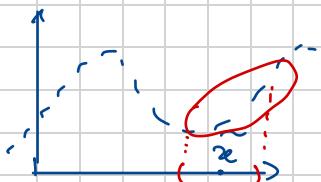
$$\begin{aligned} y_0 &= 1 & x_1' &= x & x_2' &= x^2 \\ x_3' &= \cos(x) & x_4' &= \frac{\sqrt{x}}{\log x+2} \end{aligned}$$

Локально бутылочная пересеч

1.P



$$\underset{\Theta}{\operatorname{argmin}} \quad \frac{1}{2} \sum_{i=1}^n (\bar{\Theta}^T \bar{x}^{(i)} - \bar{y}^{(i)})^2 \quad h_{\bar{\Theta}}(\bar{x})$$



$$\underset{\Theta}{\operatorname{argmin}} \quad \frac{1}{2} \sum_{i=1}^n W^{(i)} (\bar{\Theta}^T \bar{x}^{(i)} - \bar{y}^{(i)})^2, \text{ где } W^{(i)} - \text{весовая оценка, например}$$

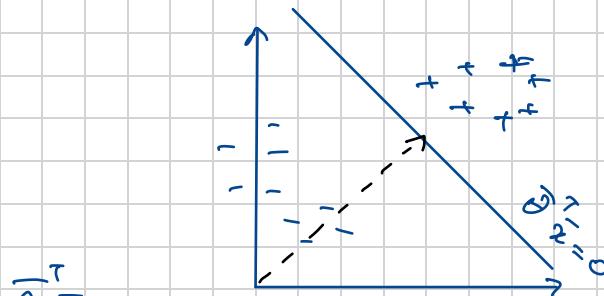
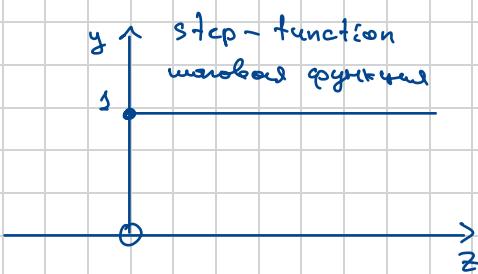
$$W_x^{(i)} = \exp \left\{ - \frac{(x^{(i)} - \tilde{x})^2}{2 \tilde{\sigma}^2} \right\}$$

Персепtron

$x^{(i)} \in \mathbb{R}^{(k+1)}$, $y^{(i)} \in \{0, 1\}$

$$h_\theta(\tilde{x}) = g(\tilde{\theta}^T \tilde{x}), \text{ где}$$

$$g(z) = \begin{cases} 1, & \text{если } z \geq 0 \\ 0, & \text{если } z < 0 \end{cases}$$



$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$$

Алгоритм обучения

$$\theta := \text{init}(\bar{\theta})$$

for i=1,2,3,...

$$\theta := \theta + \alpha (y^{(i)} - h_\theta(x^{(i)})) x^{(i)}$$

размер шага

$\tilde{\theta}$ - вектор нормали к границе классов (перпендикульар)

Нед перспек. более ярок $\tilde{\theta}x$, т.е. с какой стороны от p.2. лежит x.

Если $y^{(i)} = 0 \wedge h_\theta(x^{(i)}) = 0$, то можно не обучать

$y^{(i)} = 1 \wedge h_\theta(x^{(i)}) = 1$, то можно не обучать

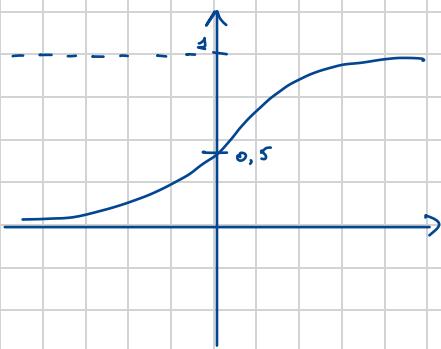
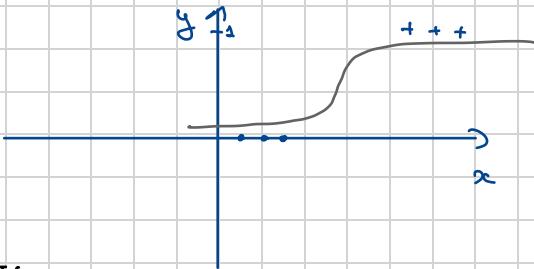
B прут. с.у. либо θ либо пред. либо вин. максимиз θ

Логистическая "рекессия"
(класификация)
(малый персентиль)

$$x^{(i)} \in \mathbb{R}^{d+1}, g^{(i)} \in \{0, 1\}$$

$$h_{\theta}(x) = g(\theta^T x)$$

↑
Интерпретация
как вероятность
логистическая
функция



$$g(0) = 0.5$$

$$\lim_{z \rightarrow \infty} g(z) = 1$$

$$\lim_{z \rightarrow -\infty} g(z) = 0$$

$$P(y^{(i)} = 1 | x^{(i)}; \Theta) = h_{\theta}(x^{(i)})$$

$$P(y^{(i)} = 0 | x^{(i)}; \Theta) = 1 - h_{\theta}(x^{(i)})$$

$$P(y^{(i)} | x^{(i)}; \Theta) = h_{\theta}(x^{(i)})^{y^{(i)}} \cdot (1 - h_{\theta}(x^{(i)}))^{1-y^{(i)}}$$

Как обустроить модель?

Методом макс. правдоподобия (MLE)

$$L(\bar{\Theta}) = P(\emptyset | \Theta) = P(\bar{y} | X; \Theta) = P(y_1, y_2, \dots, y_n | \bar{x}_1, \bar{x}_2, \dots, \bar{x}_n; \Theta) =$$

= Т.к. генерал iid

independent identically distributed

(н.о.п.с.б.)

$$= \prod_{i=1}^n P(y^{(i)} | x^{(i)}; \Theta) = \prod_{i=1}^n \left[h_{\theta}(x^{(i)})^{y^{(i)}} \cdot (1 - h_{\theta}(x^{(i)}))^{1-y^{(i)}} \right]$$

$$L(\theta) = \log L(\bar{\theta}) = \sum_{i=1}^n \left[y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \cdot \log (1-h_\theta(x^{(i)})) \right]$$

$$\begin{aligned} \nabla_\theta g(z) &= \left(\frac{1}{1+e^{-z}} \right)' = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1-1+e^{-z}}{(1+e^{-z})^2} = \frac{1-e^{-z}}{1+e^{-z}} \frac{1}{(1+e^{-z})} = \\ &= \underbrace{\frac{1}{(1+e^{-z})}}_{g(z)} \left(\underbrace{\frac{1+e^{-z}}{1+e^{-z}} - \underbrace{\frac{1}{1+e^{-z}}}_{g(z)} \right) \end{aligned}$$

$$g(z)' = g(z)(1-g(z))$$

$$L(\theta) = y \cdot \log g(\theta^T z) + (1-y) \cdot \log (1-g(\theta^T z))$$

$$\begin{aligned} \nabla_\theta L(\theta) &= \frac{y \cdot g(\theta^T z) \cdot (1-g(\theta^T z)) \cdot z}{g(\theta^T z)} + \frac{(1-y) \cdot g(\theta^T z) \cdot (1-g(\theta^T z)) \cdot z}{1-g(\theta^T z)} = \\ &= y \cdot (1-g(\theta^T z)) \cdot z - (1-y) \cdot g(\theta^T z) \cdot z = [y - y \cdot g(\theta^T z) - g(\theta^T z) + y \cdot g(\theta^T z)] = \\ &= [(y - g(\theta^T z)) \cdot z] \end{aligned}$$

↑ пакетно обновление где np. номб ённа:

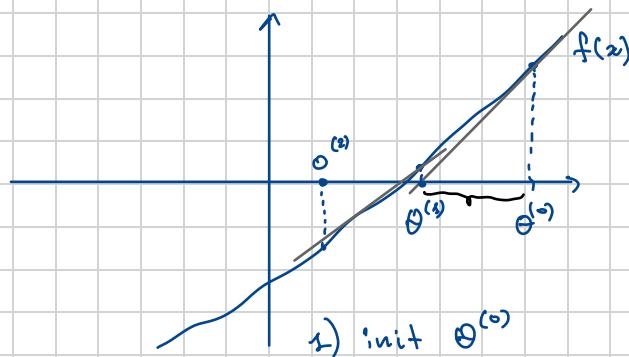
$$\theta := \theta + \alpha \sum_{i=1}^n (y^{(i)} - g(\theta^T z^{(i)})) \cdot z^{(i)}$$

Метод гисторика

$$f(x) = 0$$

↑
корен
уравнения

$$\nabla_\theta L(\theta)$$



MLE:

$$1. \text{ одномерный} \\ \theta = \theta - \frac{L'(\theta)}{f'(\theta)}$$

2. многомерный случай

$$\theta := \theta - H^{-1} \cdot \nabla_\theta L(\theta)$$

$$H_{ij} = \frac{\partial e}{\partial \theta_i \partial \theta_j}$$

- 1) init $\theta^{(0)}$
- 2) move to point go to сходимости
- 3) $\theta^{(t+1)} := \theta^{(t)} - \frac{f(\theta^{(t)})}{f'(\theta^{(t)})}$

Exponential family distribution

Up.

$$N(\mu, \sigma^2) \in \text{Exp Fam}$$

$$p(y; \eta) = b(y) \cdot \exp\{\eta^\top T(y) - a(\eta)\}$$

y - случайная величина

η - естественный параметр модели
(natural parameter)

$T(y)$ - достаточная статистика

$$T(y) = y \quad (\text{б. близкость к нулю примеров})$$

$b(y)$ - ф-я норм-разделения

(log-partition Func)

- нормализующий константа распределения

$$p(y; \eta) = b(y) \cdot e^{\eta y} \quad \text{-exponential tilting}$$

(экспоненц. трансф.)

$$A(\eta) = \int b(y) e^{\eta y} dy = E[e^{\eta y}] \quad \text{Если } b(y) \text{ - плотность распределения, то}$$

$$\int \frac{b(y) \cdot e^{\eta y}}{A(\eta)} = s$$



генерирующая ф-я моментов

$$\frac{d}{d\eta} A(\eta) = E[y]$$

$$\frac{d^2}{d\eta^2} A(\eta) = E[y^2]$$

Например

$$p(y; \phi) = \phi^y \cdot (1-\phi)^{1-y} = e^{\log(\phi^y - (1-\phi)^{1-y})} = e^{y \log \phi + (1-y) \log(1-\phi)} = \exp\left\{y \log \frac{\phi}{1-\phi} +$$

$$+ \log(1-\phi)\right\}$$



$$\eta = \log \frac{\phi}{1-\phi} \Leftrightarrow \phi = \frac{1}{1+e^{-\eta}}$$

(математическая функция)

$$b(y) = 1 \quad T(y) = y$$

$$a(\eta) = -\log(1-\phi) = -\log\left(\frac{e^{-\eta}}{1+e^{-\eta}}\right) = \log\left(\frac{1}{1+e^{\eta}}\right)$$

$$p(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{1}{2} \frac{(y-\mu)^2}{\sigma^2}\right\} = \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{(y-\mu)^2}{2}\right\}$$

$$p(y; \eta) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{y^2 - 2y\mu + \mu^2}{2}\right\} = \underbrace{\frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{y^2}{2}\right\}}_{b(y)} \cdot \exp\left\{y\mu - \frac{\mu^2}{2}\right\} \sim a(\eta)$$

$$b(y) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{y^2}{2}\right\}$$

$$\boxed{\eta = \mu \Leftrightarrow \mu = \eta}$$

$$a(\eta) = \frac{\mu^2}{2} = \frac{\eta^2}{2}$$

$$p(y; \eta) = N(0, 1) \cdot \exp\left\{y\eta - \frac{\eta^2}{2}\right\}$$

Свойства

1) $\log p(y; \eta)$ — борнхорд ф-я

L — борнхорд по η

NLL — борнхорд по y

$$2) E[y; \eta] = \frac{d}{d\eta} a(\eta)$$

$$E[y^2; \eta] = \frac{d^2}{d\eta^2} a(\eta)$$

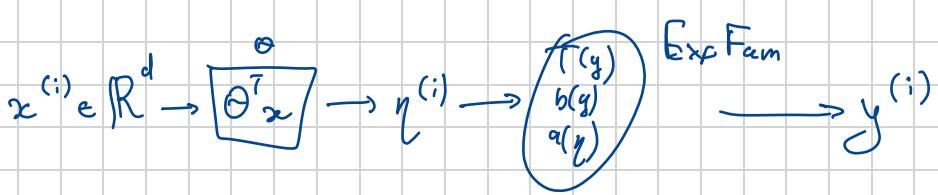
Обобщенное линейное моделирование (GLM)

1. Параметрическое

a) $y | \bar{x}; \Theta \sim \text{ExpFam}$ (где Θ — параметры базиса)

$$\delta) h_\Theta(\bar{x}) = E[y | \bar{x}; \Theta]$$

$$\theta) \eta = \Theta^T \bar{x}^- \quad (\text{линейное моделирование})$$



Линейная регрессия

$\checkmark \in \text{Exp Family}$

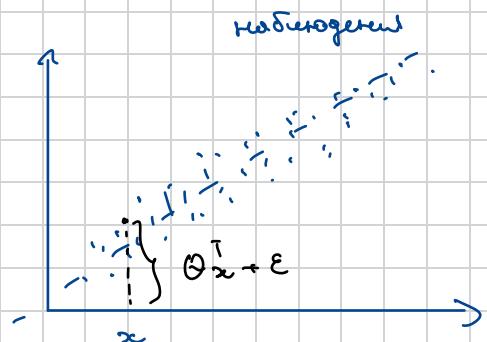
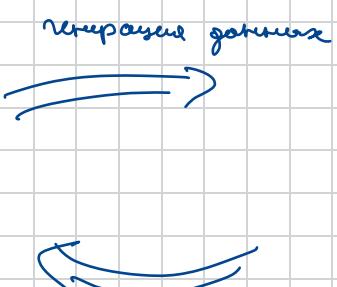
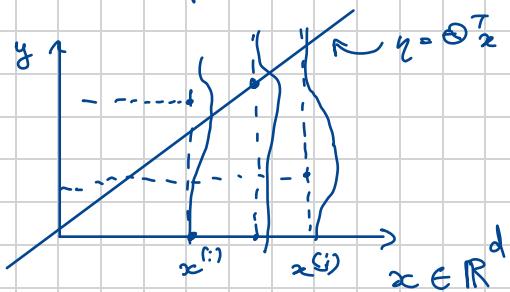
$$h_{\Theta}(x) = E[y|x; \Theta] = \mu = \eta = \Theta^T x$$

Логистическая регрессия

Bernoulli Exp Fam: y

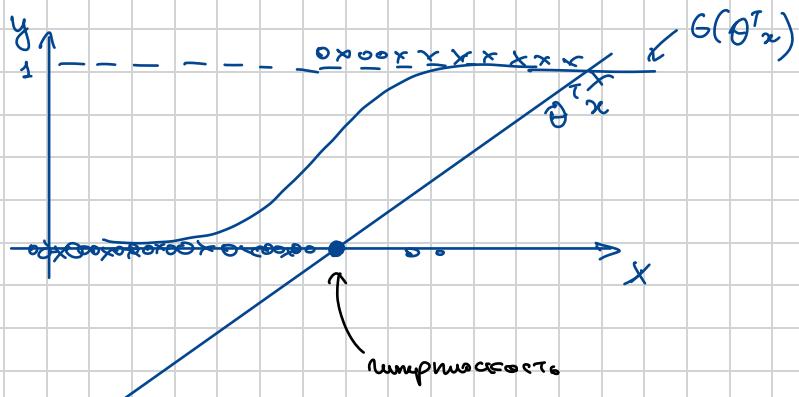
$$h_{\Theta}(x) = E[y|x; \Theta] = p = \frac{1}{1 + e^{-\eta}} = \frac{1}{1 + e^{-\Theta^T x}}$$

Линейная регрессия



подгонка модели (Θ)

- ML = {
1. Подбор модели \leftarrow делает человек
 2. Подгонка модели под данные (обучение) \leftarrow делает компьютер



Параметризация

Параметры модели:

Естественные

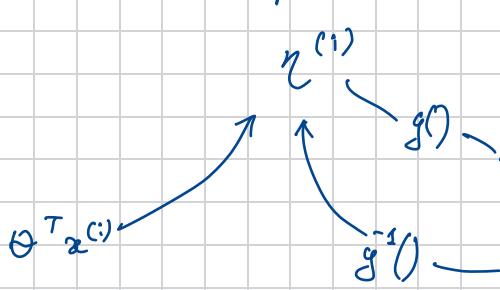
параметры:

Канонические параметры

(для среднего)

$$\theta \in \mathbb{R}^d$$

"обуслови"



- μ (мод)
- ϕ (Берншт)
- λ (Рыбакова)
- a, b (Бета)

$g(\cdot)$ — каноническая φ -л откликса

$g^{-1}(\cdot)$ — каноническая φ -л сбояи

$$h_\theta(x) = E[y | x; \theta] \boxed{= g(\theta^T x)} \Rightarrow y$$

5) Модели

Тип генерации
ген. y

Распр. & Exp Fam

Название мо

$$\mathbb{R}$$

нормальное расп-е
Лапласа

линейная
репрессия
репрессия Лапласа

$$\{0, 1\}$$

Берншт

(бинарная классификация)
модель логистической
репрессии

$$\{1, 2, \dots, k\}$$

Категориальное

(многоклассовая
классификация)
сортимент-репрессия

$$\mathbb{N}_+$$

целое nat. число

Рыбакова

Репрессия Рыбакова

$$\mathbb{R}_+$$

Экспоненциальное
расп.
гамильт

модели спада
важимости

$$[0, 1]$$

вероятность

Бета

$$[0, 1]^k$$

Дерихов

Лекция GLM

1. Автоматическое построение моделей

- баз. расп. (на осн. y)
- баз. через Exp Fam \Rightarrow нал. $g(\eta)$
- ищем линию $h_{\theta}(x) = g(\theta^T x)$

2. Однократное правило обновления весов при SGD:

$$\bar{\theta} := \bar{\theta} + d(y^{(i)} - h_{\theta}(x^{(i)})) \cdot x^{(i)}$$

3. Прогноз

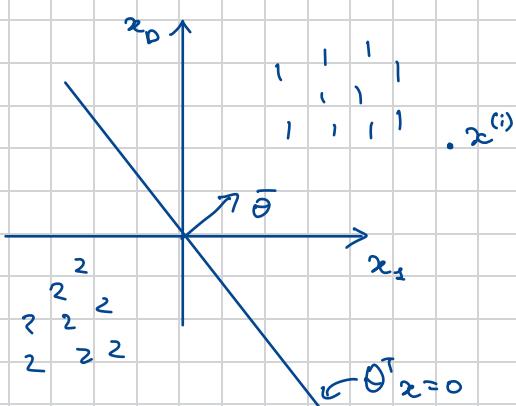
$$h_{\bar{\theta}}(x^*) = g(\bar{\theta}^T x^*)$$

4. Быстрее модельный оптимум

5. NLL=Loss Function

Мноклассовая классификация

Бинарный случай



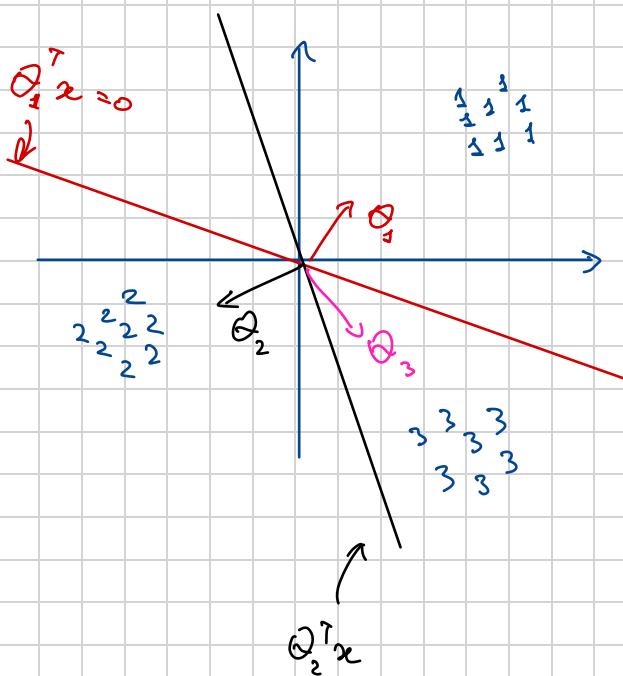
$\theta^T x^{(i)}$

$= 0$, т.е. $x^{(i)}$ на линии	$\theta^T x^{(i)}$ > 0 , т.е. $x^{(i)}$ в области, куда "смотрит" нормаль θ
< 0 , т.е. $x^{(i)}$ в противоположной области	

$$p(y=1; \theta) = g(\theta^T x), \text{ где } g-\text{ист. ф-я}$$

$$p(y=2; \theta) = 1 - g(\theta^T x)$$

Многоклассовое распознавание



$$\text{One: } x: \begin{bmatrix} \theta_1^T x \\ \theta_2^T x \\ \theta_3^T x \end{bmatrix}$$

Other:
 $\arg\max_i \theta_i^T x$

Например:

one-hot $\rightarrow \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, если $\theta_1^T x -$
 - макс.



$$\begin{bmatrix} \theta_1^T x \\ \theta_2^T x \\ \theta_3^T x \end{bmatrix} \xrightarrow{\text{экспоненцируем}} \begin{bmatrix} e^{\theta_1^T x} \\ e^{\theta_2^T x} \\ e^{\theta_3^T x} \end{bmatrix} \xrightarrow{\text{нормируем}} \begin{bmatrix} e^{\theta_1^T x} / \sum \\ e^{\theta_2^T x} / \sum \\ e^{\theta_3^T x} / \sum \end{bmatrix},$$

Мы хотим получить единицу:

$$1) a_i > 0, \forall i=1, \dots, k$$

$$2) \sum_{i=1}^k a_i = 1$$

$$\text{так } \sum = \sum_{i=1}^k e^{\theta_i^T x}$$

$$P(y=i | x; \theta) = \frac{e^{\theta_i^T x}}{\sum_{i=1}^k e^{\theta_i^T x}} - \text{softmax}$$

Пример

$$\begin{bmatrix} 1.2 \\ -0.8 \\ 0.5 \end{bmatrix} \Rightarrow \begin{bmatrix} 3.32 \\ 0.449 \\ 1.649 \end{bmatrix} \Rightarrow \begin{bmatrix} 0.613 \\ 0.083 \\ 0.304 \end{bmatrix}$$

$$\begin{bmatrix} q \\ 0.613 \\ 0.083 \\ 0.304 \end{bmatrix}$$

$$\begin{bmatrix} p \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$E[p] = \sum_i p_i \cdot \log \frac{1}{p_i}$$

$$KL(p||q) = CE(p, q) - E[p]$$

Расхождение

$$CE(p, q) = \sum_i p_i \cdot \log \frac{1}{q_i}$$

Коэффициент - мера близости \Rightarrow ошибка перекрестной энтропии

$$CE(q)_i^{(i)} - \log q_i^{(i)} = \log \frac{1}{p_i^{(i)}} \text{ где } i - \text{категория класса}$$

$$P(y; \phi) = \phi^y \cdot (1-\phi)^{1-y}$$

$$L(\theta) = \prod_{i=1}^n p_i^{y^{(i)}} \cdot (1-p_i)^{1-y^{(i)}}$$

$$L(\theta) = \log(L(\theta)) = \sum_{i=1}^n [y^{(i)} \cdot \log \phi_i + (1-y^{(i)}) \cdot \log(1-\phi_i)]$$

$$\left. \begin{array}{l} \text{если } y^{(i)} = 1 : -\log \phi_i \\ \text{если } y^{(i)} = 0 : -\log(1-\phi_i) \end{array} \right\} \begin{array}{l} \text{CE-терминал} \\ \text{на } i\text{-м } \text{шаге} \end{array}$$

loop-суммы "кенн" классов

$y \sim N \Rightarrow$ лин. пред.
с квадратич. оши. SE

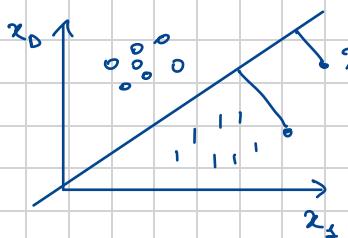
$y \sim \text{Bern} \Rightarrow$ лин. пред. с ошибкой CE

$y \sim \text{Cat} \Rightarrow$ softmax пред.

с CE-loss

Генеративные модели

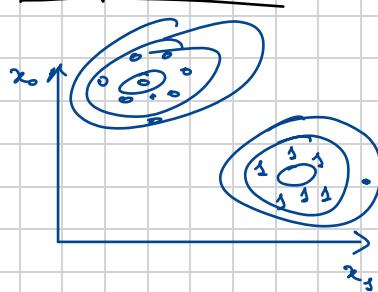
3) Генеративные модели

дискретно-непрерывные

Справка

$$p(y|x) - \text{(нор. расп.)}$$

или

 $h_\theta(x)$ непрерывн. (персептрон)генеративные

Справка

$$[p(x,y)] = p(x|y) \cdot p(y)$$

обум. расп.

$$p(y|x) \cdot p(x)$$

условн. расп.

априорное расп.

Примечание

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)}$$

$$p(x) = p(x|y=0) \cdot p(y=0) + p(x|y=1) \cdot p(y=1)$$

Процесс генерации

1. Опред. с помощью $p(y)$ класс K2. Генерируючи с помощью $p(x|y=k)$ новый образец

Обе модели

1. GDA $\rightarrow x \in \mathbb{R}^n$ 2. NB $\rightarrow x$ -дискретное

2) Гауссовский дискриминантный анализ

Предположение

$$y \sim \text{Bern}(\phi) \sim N(\bar{\mu}_1, \Sigma)$$

$$x|y=0 \sim N(\bar{\mu}_2, \Sigma)$$

$$\text{Bern}(\phi) = \phi^y \cdot (1-\phi)^{1-y}$$

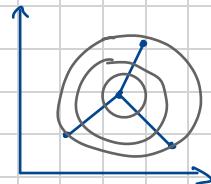
$$N(\bar{\mu}_1, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \cdot \exp \left\{ -\frac{1}{2} (\bar{x} - \bar{\mu}_1)^T \Sigma^{-1} (\bar{x} - \bar{\mu}_1) \right\}$$

$$N(\bar{\mu}_2, \Sigma) = \dots$$

Δ^2 — расстояние Manhattan

Параметры модели

$$\phi, \bar{\mu}_1, \bar{\mu}_2, \Sigma$$



Определение

$$\text{Будем} \quad X = \{(\bar{x}^{(1)}, y^{(1)}), \dots, (\bar{x}^{(n)}, y^{(n)})\}$$

$$l(\phi) = \log p(X; \phi) = \log \prod_{i=1}^n p(\bar{x}^{(i)}, y^{(i)}; \phi) =$$

$$= \sum_{i=1}^n \log(p(\bar{x}^{(i)} | y^{(i)}; \phi) \cdot p(y^{(i)}; \phi)) \Rightarrow \boxed{\nabla_{\phi} l(\phi) = 0}$$

Определение:

$$\phi^{ML} = \frac{\sum_{i=1}^n \prod_{j \neq i} \{y_j^{(j)} = 1\}}{N}$$

$$\sum_{i=1}^n \bar{\mu}_1 = \frac{\sum_{i=1}^n (\bar{x}^{(i)} - \bar{\mu}_1) (\bar{x}^{(i)} - \bar{\mu}_1)^T}{N}$$

$$\bar{\mu}_2^{ML} = \frac{\sum_{i=1}^n \prod_{j \neq i} \{y_j^{(j)} = 1\} \cdot \bar{x}^{(i)}}{\sum_{i=1}^n \prod_{j \neq i} \{y_j^{(j)} = 1\}}$$

уравнение

$$\bar{\mu}_1^{ML} = \dots$$

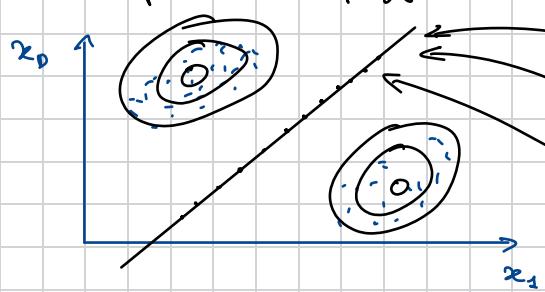
3) Промоделирование

$$P(y=1|x) = \frac{P(x|y=1) \cdot P(y=1)}{P(x|y=1) \cdot P(y=1) + P(x|y=0) \cdot P(y=0)}$$

$$P(y=0|x) = \dots$$

Объект: $\arg \max_y P(y|x)$

Интерпретация результатов



$$\hat{\theta}^T x$$

$$P(x|y=0) = P(x|y=1)$$

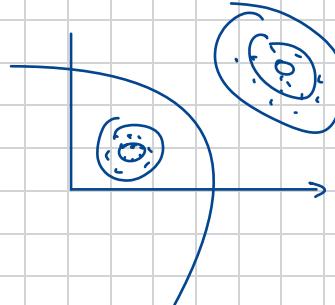
$$P(y=1|x) = 0.5$$

$$\Rightarrow$$

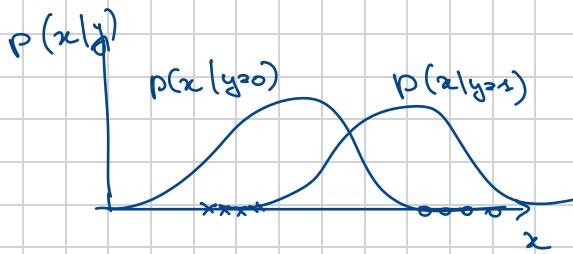
$$\frac{1}{1 + e^{-\hat{\theta}^T x}},$$

$$\text{так } \hat{\theta} = f(\phi)$$

если $\Sigma_1 \neq \Sigma_2$



Выражение $P(y=1|x)$ как функция от x



4) Сравнение GDA и лог. регрессии

Генеративный (GDA) номогр

$$\begin{aligned} \textcircled{1} \quad & z|y=0 \sim N(\mu_1, \Sigma) \\ & z|y=1 \sim N(\mu_2, \Sigma) \\ & y \sim \text{Bern}(\phi) \end{aligned}$$

самостоят номогр.

Дискриминат. номогр (LR)

$$p(y=1|x) = \frac{\phi}{1 + e^{-\theta^T x}}$$

никаких
номогр. о генерике
самостоят
номогр

HO
~~z~~

$$\begin{aligned} \textcircled{2} \quad & z|y=0 \sim \text{Poisson}(\lambda_0) \\ & z|y=1 \sim \text{Poisson}(\lambda) \\ & y \sim \text{Bern}(\phi) \end{aligned}$$

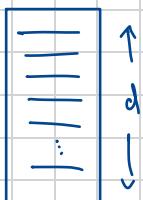
\Rightarrow —//—

NBC

наивные Байесовский классификатор

1. Модели (Бернoulli)

Строки символь



\Rightarrow номогр. текст т.к.:

$$\bar{x} = \{0, 1\}^d \mid x_j = \begin{cases} 1, \text{ если } \text{символ } j \text{ бх. в } t \\ 0, \text{ если нет} \end{cases}$$

Тройкеты модели:

$$p(y) = \text{Bern}(\phi_y) - 1 \text{ нт}$$

$$p(x_j | y=0) = \text{Bern}(\phi_{j|0}) - d \text{ нт.}$$

$$p(x_j | y=1) = \text{Bern}(\phi_{j|1}) - d \text{ нт.}$$

2d+1 параметров

$$\begin{aligned} p(\bar{x}) &= p(x_1, x_2, \dots, x_d) = p(x_1|y) \cdot p(x_2|x_1, y) \cdot \\ &\dots \cdot p(x_d|x_1, x_2, y) = \prod_{i=1}^d p(x_i|y) \end{aligned}$$

Сумм. фнк. номогр. (3D в кубе
"наивнейшее" NBC)

$$p(a, b | y) = p(a | y) \cdot p(b | a, y)$$

установлено номогр.
 $p(b | a, y) = p(b | y)$

$$\downarrow$$

$$p(a, b | y) = p(a | y) \cdot p(b | y)$$

$$\phi = (\phi_y, \bar{\phi}_{j/0}, \bar{\phi}_{j/1})$$

2. Обучение

$$l(\phi) = \log p(X; \phi) \stackrel{i.i.d.}{=} \log \prod_{i=1}^n p(x^{(i)}, y^{(i)}; \phi) = \log \prod_{i=1}^n p(y^{(i)}; \phi_y) \cdot p(x^{(i)} | y^{(i)}; \phi) =$$

независимое
предпол.

$$= \log \prod_{i=1}^n p(y^{(i)}; \phi_y) \cdot \left[\prod_{j=1}^d p(x_j^{(i)} | y^{(i)}; \phi) \right] \Rightarrow \nabla_\phi l(\phi) = 0 \Rightarrow \bar{\phi}_{j/0} = \frac{\sum_{i=1}^n \mathbb{1}\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^n \mathbb{1}\{y^{(i)} = 0\}}$$

$\phi_y = \frac{\sum_{i=1}^n \mathbb{1}\{y^{(i)} = 1\}}{n}$

$\sum_{i=1}^n \mathbb{1}\{y^{(i)} = 1\}$

Храним в виде частных

Вместо ϕ храним

$$b = (b_1, b_2, \dots, b_d) - частные \phi_{j/1}$$

$$g = (g_1, g_2, \dots, g_d) - частные \phi_{j/0}$$

s - # слова, h - # не слова

\uparrow
как-то

$$\phi_{j/0} = \frac{g_j}{h}, \quad \phi_{j/1} = \frac{b_j}{s}$$

$$\phi_y = \frac{s}{h+s}$$

3. Редукция модели

\bar{x}

$$\arg \max_y p(y | \bar{x}; \phi) = \arg \max_y$$

$$\frac{p(\bar{x} | y; \phi) \cdot p(y; \phi_y)}{p(\bar{x}; \phi)} =$$

использует $\arg \max$
не считает

$$= \dots \Rightarrow \begin{cases} (1 - \phi_y) \cdot \prod_{j \in S} \phi_{j/0}^{x_j} \cdot (1 - \phi_{j/0})^{1-x_j}, & \text{если } y = 0 \\ \phi_y \cdot \prod_{j \in S} \phi_{j/1}^{x_j} \cdot (1 - \phi_{j/1})^{1-x_j}, & \text{если } y = 1 \end{cases}$$

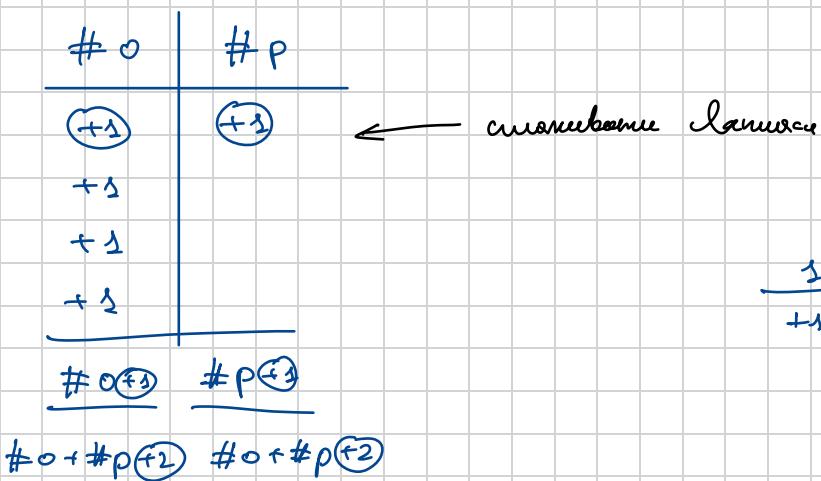
Вер-тб, что j -ое слово
пол. в снах, т.е. кон-бо
снаш - сообн., в кот. выражении
слово j , значение не кон.
снаш - сообн.

Проблема 3: $d=10000$, $\phi \approx 10^{-3} \Rightarrow$ анс. $b_{sp} \approx 10^{30000}$

решение: $\log_{10}(10^{30000}) = 30000$

Проблема 2:

Сокращение Ланчара



$$\begin{array}{c|c|c|c|c|c} 1 & 2 & 3 & 4 & 5 & 6 \\ \hline +1 & +3 & +8 & +8 & +8 & +1 \\ \end{array}$$

$$\phi_{j/s} = \frac{b_j + s}{s + 2}$$

$$\phi_{j/o} = \frac{g_j + s}{h + 2}$$

$$\phi_j = \phi_y$$

NB с мультиплексированной моделью

Пусть t_i — путь текста t

x_j — id j-го символа в тексте

$$x^{(i)} \in \{1, 2, \dots, d\}^{t_i}$$

$$\begin{matrix} 218 & : & \text{мама} \\ 520 & : & \text{мама} \\ 2054 & : & \text{речь} \end{matrix}$$

мама мама речь

$$x = (218, 520, 2054)$$

модели:

$$y \sim \text{Bern}(\phi) \rightarrow \text{ур.}$$

$$\bar{x}|y=0 \sim \text{Cat}(\bar{\phi}_{k/o}) \rightarrow \text{ур.}$$

$$\bar{x}|y=s \sim \text{Cat}(\bar{\phi}_{k/t}) \rightarrow \text{ур.}$$

$$\bar{\phi}_{k/s} = \frac{\sum_{i=1}^n \sum_{j=1}^{t_i} \mathbb{1}\{x_j^{(i)} = k \wedge y^{(i)} = s\}}{\sum_{i=1}^n \mathbb{1}\{y^{(i)} = s\} \cdot t_i + d}$$

сумм. баланс

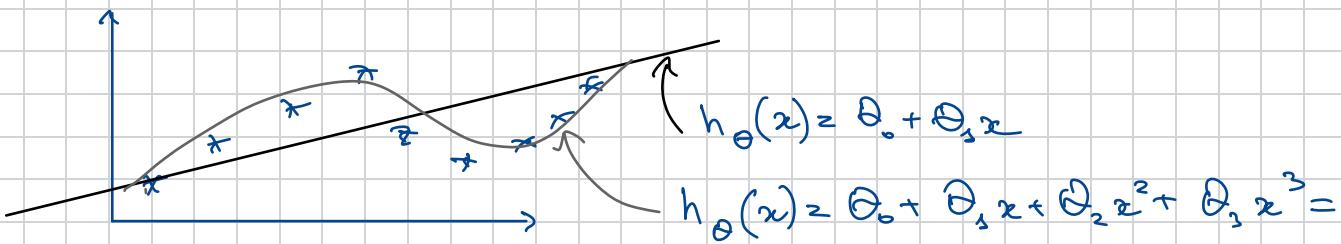
$$\Phi_{k10} = \dots$$

$$\Phi_y = \dots$$

10.10

Лекция SVM

1. Трансформация признаков в ядро



$$\phi(x) = x$$

$$\phi: \mathbb{R}^d \rightarrow \mathbb{R}^p$$

$$= \theta_0 + \theta_1 \phi(x)_1 + \theta_2 \cdot \phi(x)_2 + \theta_3 \cdot \phi(x)_3 =$$

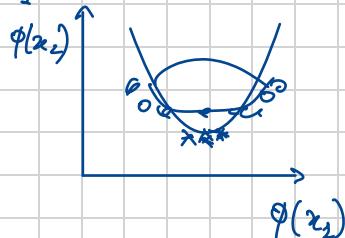
$$= \bar{\theta}^\top \phi(x), \text{ где } \phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix}$$

$$\{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\} \xrightarrow{\phi} \{(\phi(x^{(1)}), y^{(1)}), \dots, (\phi(x^{(n)}), y^{(n)})\}$$

$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$x_2$$

$$\phi(\bar{x}) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_1 x_2 \end{bmatrix}$$



Ядро - функция от данных:

$$K(\bar{x}, \bar{z}) \triangleq \phi(\bar{x})^\top \phi(\bar{z})$$

Пример 1:

$$x, z \in \mathbb{R}^d; \underbrace{\phi(x), \phi(z) \in \mathbb{R}^{d^3}}_{O(d^3)} \Rightarrow k(x, z) \in \mathbb{R} \underbrace{O(d)}$$

Пример 2

$$x, z \in \mathbb{R}^d \quad \underbrace{\phi(x), \phi(z) \in \mathbb{R}^p}_{O(p)} \Rightarrow k(x, z) \in \mathbb{R} \underbrace{O(d)}$$

2. Градиентный метод

$$\text{init} \rightarrow \theta := \overline{0} \quad (\text{на примере линейной регрессии})$$

$$\theta := \theta + \lambda \sum_{i=1}^n (y^{(i)} - \theta^T \phi(x^{(i)})) \cdot \phi(x^{(i)})$$

$O(p)$

$$\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \\ x_1 x_2 \\ x_1 x_3 \\ \vdots \\ x_d x_1 \\ \vdots \\ x_1 x_2 x_3 \\ \vdots \end{bmatrix} \begin{array}{c} d \\ \\ d^2 \\ \\ d^2 \\ \\ d^3 \end{array}$$

$$x \in \mathbb{R}^d$$

$$\phi(x) \in \mathbb{R}^p$$

$$\theta \in \mathbb{R}^p$$

$$\text{Если } d = 10^3$$

$$p \approx 10^9$$

$$p = 1 + d + d^2 + d^3$$

Теорема на основе метода GD θ может быть представлена как в.к. бз.

где:

$$\theta = \sum_{i=1}^n \beta_i \cdot \phi(x^{(i)})$$

Док-бо:

но что же это?

$$\text{Соум: } t \geq 0 : \theta = \sum_{i=1}^n \alpha_i \cdot \phi(x^{(i)})$$

$$t = 1 : \theta = \sum \alpha_i y^{(i)} \cdot \phi(x^{(i)}) \underbrace{\beta_i}_{\beta_i}$$

Числ. пригн.

$$\Theta = \sum_{i=1}^n \beta_i \phi(x^{(i)})$$

$$\left| \begin{aligned} \Theta &= \sum_{i=1}^n \beta_i \phi(x^{(i)}) + \lambda \cdot \sum_{i=1}^n (y^{(i)} - \Theta^T \phi(x^{(i)})) \\ &= \sum_{i=1}^n \underbrace{\left(\beta_i + \lambda (y^{(i)} - \Theta^T \phi(x^{(i)})) \right)}_{\beta_i} \cdot \phi(x^{(i)}) \end{aligned} \right.$$

$$\Theta = \sum_{i=1}^n \beta_i \phi(x^{(i)})$$

$$\beta_i := \frac{\Theta - \sum_{j \neq i} \beta_j \phi(x^{(j)})^T \phi(x^{(i)})}{\sum_{j \neq i} \phi(x^{(j)})^T \phi(x^{(i)})} = \beta_i + \lambda (y^{(i)} - \sum_{j \neq i} \beta_j \cdot k(x^{(i)}, x^{(j)}))$$

Держимо для GD:

0) Бон. K на охоле X , ван. $K(\cdot, \cdot)$

1) Числ., $\bar{\beta} = \overline{\Theta}$

2) Підсв. до складення $O(n^2)$
на штрафах

$$\bar{\beta} := \overline{\beta} + \lambda (\bar{y} - K \cdot \bar{\beta})$$

бекторизаційна
форма

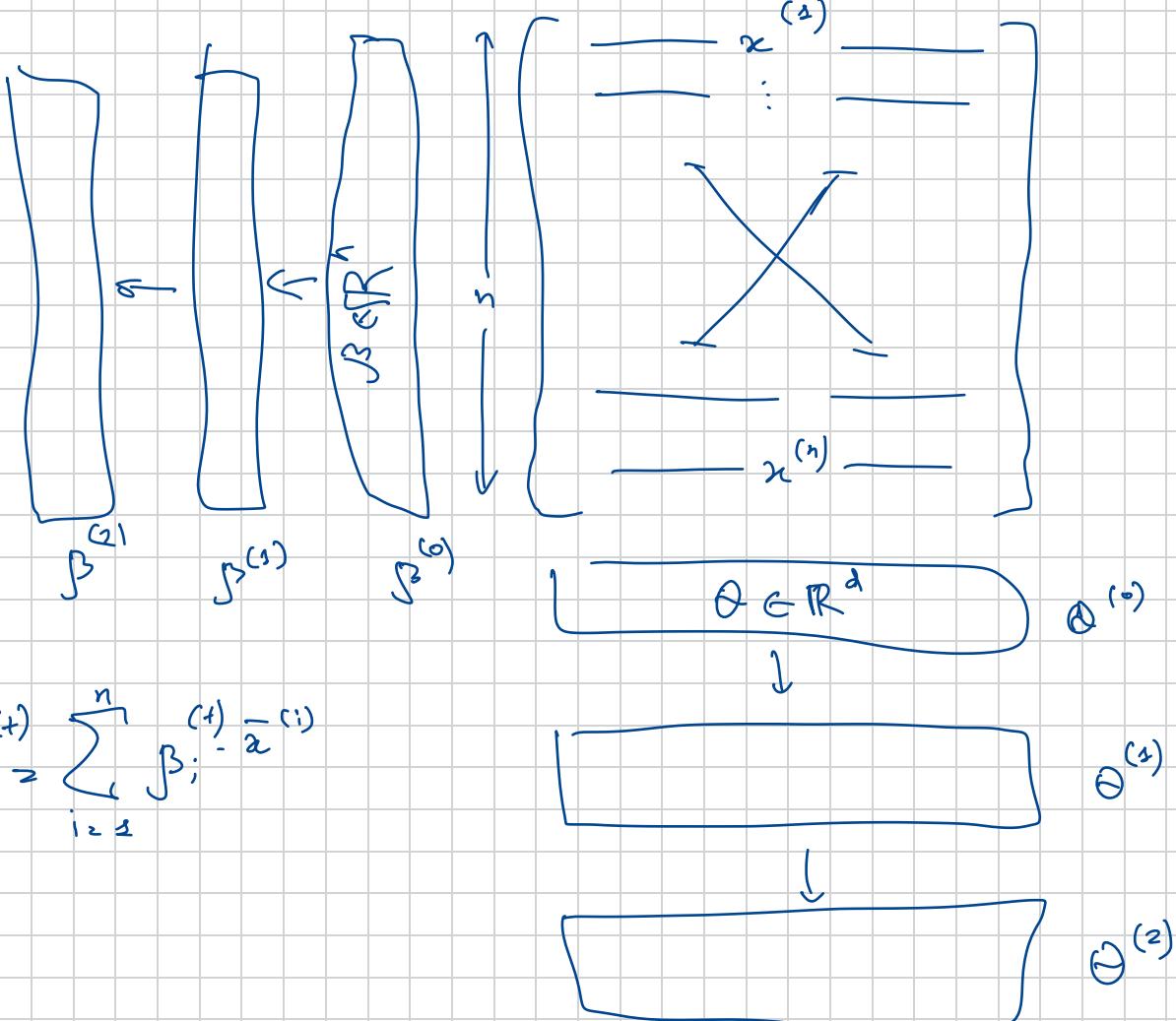
Во броях будо же

$$h_\Theta(\bar{x}) = \bar{\Theta}^T \phi(\bar{x}) = \sum_{i=1}^n \beta_i \cdot \phi(x^{(i)})^T \cdot \phi(\bar{x}) = \sum_{i=1}^n \beta_i \cdot K(\bar{x}^{(i)}, \bar{x})$$

	однакове щукомануєше пр. чи др.	будо же	нашого
числ. чесн.	$O(np) + O(T_{np})$	$O(p)$	$O(p)$ згд Q
дедукції чесн.	$O(n^2 d) + O(T_n^2)$	$O(nd)$	$O(nd) + O(n)$ згд X згд B

$$\Phi(\bar{x})^T \Phi(z) = \begin{bmatrix} 1 \\ \vdots \\ z_1 \\ \vdots \\ x_i x_j \\ \vdots \\ x_i x_j z_k \end{bmatrix}^T \begin{bmatrix} 1 \\ \vdots \\ \bar{z}_1 \\ \vdots \\ \bar{z}_j \\ \vdots \\ \bar{z}_i \bar{z}_j \\ \vdots \\ \bar{z}_i \bar{z}_j z_k \end{bmatrix} = 1 + \sum_{i=1}^n z_i \bar{z}_i + \sum_{i=1}^n \sum_{j=1}^n x_i x_j z_i \bar{z}_j + \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n x_i x_j z_k \bar{z}_i \bar{z}_j = 1 + \bar{x}^T z + (\bar{x}^T z)^2 + (\bar{x}^T z)^3$$

Выделение



3. Примеры ядер в задачах по реконструкции

$$a) \phi(\bar{z}) = \begin{bmatrix} \vdots \\ z_i z_j \\ \vdots \end{bmatrix} \Rightarrow k(\bar{x}, \bar{z}) = (\bar{x}^\top \bar{z})^2$$

$$b) \phi(z) = \begin{bmatrix} \vdots \\ z_i z_j \\ \vdots \\ \bar{x}^\top z_i \\ \vdots \\ c \end{bmatrix} \Rightarrow k(\bar{x}, \bar{z}) = (\bar{x}^\top \bar{z} + c)^2$$

б) полиномиальное ядро d

$$k(\bar{x}, \bar{z}) = (\bar{x}^\top \bar{z} + c)^d - бе однозначно до параметра d$$

в) линейное ядро

$$k(x, z) = \bar{x}^\top \bar{z}$$

$$g) Гауссово ядро: k(\bar{x}, \bar{z}) = \exp \left\{ \frac{-\|\bar{x} - \bar{z}\|^2}{2\sigma^2} \right\}$$

$$\begin{aligned} & \rightarrow \text{если } d \rightarrow \infty \\ & \text{то } k(\bar{x}, \bar{z}) \rightarrow 0 \end{aligned}$$

- $\phi(\bar{z})$ - бесконечно мерно

- это ядро есть мерно и мерой "склоняется" к векторам

$$k(\bar{x}, \bar{z}) \approx 1$$

$$k(\bar{x}, \bar{z}) \rightarrow 0$$

$$\|\bar{x} - \bar{z}\| \rightarrow \infty$$

Теорема Шварца:

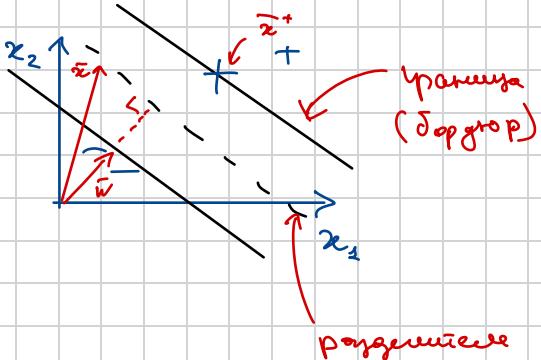
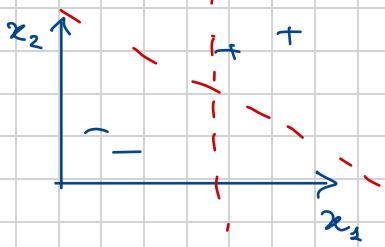
Рядущее $k(\cdot, \cdot)$ эл. ядро $\Leftrightarrow \forall X = \{\bar{x}^{(1)}, \dots, \bar{x}^{(n)}\}$

матрица $K_{ij} = k(\bar{x}^{(i)}, \bar{x}^{(j)})$ обладает положительной полуопределенностью

Матрица A эл. PSD $\Leftrightarrow \forall \bar{z}: \bar{z}^\top A \bar{z} \geq 0$

4. Машинна опорних векторів (Support Vector Machine)

Приклад: Дискретні класи суперпозицій (примір лінійної розмежуваності даних)



$$\bar{x}^T \bar{w} = c \Rightarrow \bar{x}^T \bar{w} + b = 0$$

Чис. 6 обчислених

$$\bar{\Theta} = \begin{bmatrix} \Theta_0 \\ \Theta_1 \\ \vdots \\ \Theta_d \end{bmatrix} \quad \bar{z} = \begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_d \end{bmatrix}$$

$$\bar{\Theta}^T \bar{z} = \bar{w}^T \bar{z} + b$$

$$y \in \{-1, 1\}$$

$$\underbrace{\bar{x}_+^T \bar{w} + b}_k \geq 1, \quad \forall x_+ \in \text{найменшій бірдгор} \\ \text{заряд}$$

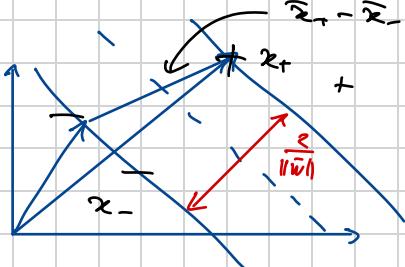
Максимізуватись ф.з. no \bar{w} та b необхідно!!!

$$\bar{w}^T \bar{x} + b = 0 \Rightarrow d\bar{w}^T \bar{x} + db = 0$$

$\underbrace{d\bar{w}^T \bar{x}}_{w'}, \underbrace{db}_{b'}$

Ограниченні: $\bar{x}^T \bar{w} + b \geq 1, \quad y^{(i)} = +1$ $\bar{x}^T \bar{w} + b \leq -1, \quad y^{(i)} = -1$ $\Rightarrow \begin{cases} \bar{x}_+^T \bar{w} + b \geq 1, & y^{(i)} = +1 \\ \bar{x}_-^T \bar{w} + b \leq -1, & y^{(i)} = -1 \end{cases} \Rightarrow$
заряд

$$\Rightarrow y^{(i)} (\bar{x}_+^T \bar{w} + b) - 1 \geq 0$$



$$\text{ширина} = (\bar{x}_+ - \bar{x}_-) \cdot \frac{\bar{w}}{\|\bar{w}\|}$$

$$= \frac{(\bar{x}_+^T + \bar{w})}{\|\bar{w}\|} + \frac{(-\bar{x}_-^T - \bar{w})}{\|\bar{w}\|} = 1 + b = \frac{2}{\|\bar{w}\|}$$

$$\max_{\bar{w}, b} \frac{\frac{1}{2} \|w\|^2}{\|w\|} \Rightarrow \min_{\bar{w}, b} \|w\| \Rightarrow \min_{\bar{w}, b} \frac{\frac{1}{2}}{2} \|w\|^2$$

$\min_{\bar{w}, b} \frac{1}{2} \|w\|^2$, upu ychubem

$$y^{(i)}(w^T x^{(i)} + b) - 1 \geq 0, \forall i = 1, n$$

5. Метод квадратов

$$L(\bar{w}, b, \bar{\alpha}) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \lambda_i [y^{(i)}(w^T x^{(i)} + b) - 1]$$

Примечание:

$$\min_{\bar{w}, b} \max_{\bar{\alpha}} L(\bar{w}, b, \bar{\alpha})$$

Карун-Кун-Такея

Обобщенное ядро:

$$\max_{\bar{\alpha}} \min_{\bar{w}, b} L(\bar{w}, b, \bar{\alpha})$$

эквивалентно тому, что имеем
primal = dual

Решение

$$\nabla_{\bar{w}} L = \bar{w} - \sum_{i=1}^n \lambda_i y^{(i)} x^{(i)} = 0 \Rightarrow \bar{w} = \sum_{i=1}^n \lambda_i y^{(i)} x^{(i)}$$

$$\boxed{\nabla_b L = \sum_{i=1}^n \lambda_i y^{(i)} = 0}$$

$$L(\bar{w}, b, \bar{\alpha}) = \frac{1}{2} \left(\sum_i \lambda_i y^{(i)} x^{(i)} \right) \left(\sum_j \lambda_j y^{(j)} x^{(j)} \right) - \sum_{i=1}^n \lambda_i y^{(i)} \left(\sum_{j=1}^n \lambda_j y^{(j)} x^{(j)} \right)^T x^{(i)}$$

$$\underbrace{-b \cdot \sum_{i=1}^n \lambda_i y^{(i)}}_{=0} + \sum_{i=1}^n \lambda_i = \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y^{(i)} y^{(j)} x^{(i)} x^{(j)}$$

Дл. экл. ядра:

$$\max_{\bar{\alpha}} \sum_i \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y^{(i)} y^{(j)} x^{(i)} x^{(j)}$$

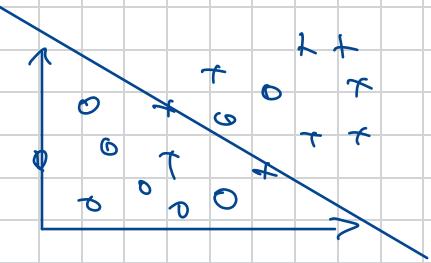
$$\lambda_i \geq 0$$

$$\sum_{i=1}^n \lambda_i y^{(i)} = 0, \forall i = 1, n$$

Тонкое градиентное бокопад $\alpha_i \neq 0$!

$$\text{Быстро: } w^T x + b = \left(\sum_{i=1}^n d_i y^{(i)} x^{(i)} \right)^T x + b = \sum_{i=1}^n d_i y^{(i)} x^{(i)T} x + b$$

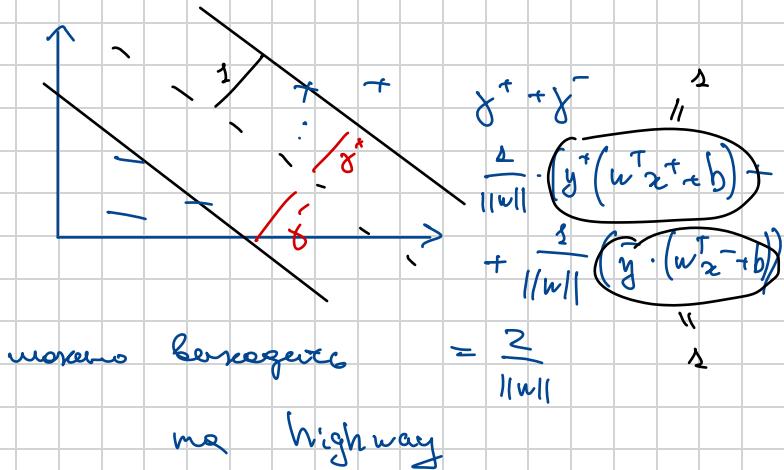
6. Равнодулярный и неравнодулярный бокорпаса



$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \zeta_i;$$

$$y^{(i)} (w^T x^{(i)} + b) \geq 1 - \zeta_i;$$

$$\zeta_i \geq 0, i = 1, n$$



Soft Margin SVM

$$\left\{ \begin{array}{l} \text{Об. экв. задача:} \\ \max_{\bar{\lambda}} \sum_i d_i - \frac{1}{2} \sum_i \sum_j d_i d_j y^{(i)} y^{(j)} \bar{\lambda}^{(i)} \bar{\lambda}^{(j)} \\ \text{с.t. } d_i \geq 0 \\ \sum_{i=1}^n \bar{\lambda}_i y^{(i)} = 0, \forall i = 1, n \end{array} \right.$$

Геометрический ядер

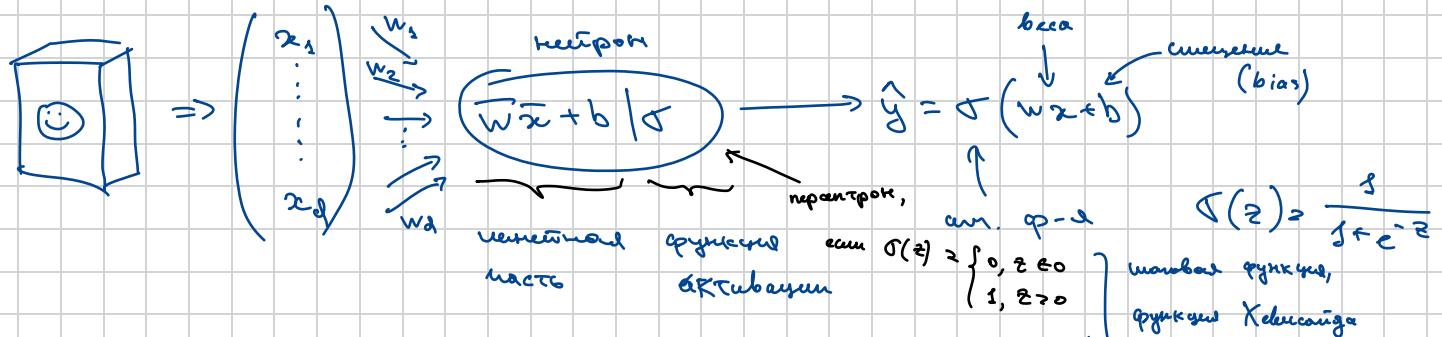
$$f^{(i)} = y^{(i)} \cdot \left(\frac{w^T}{\|w\|} \cdot \bar{x}^{(i)} + \frac{b}{\|w\|} \right)$$

$$\max_{w, b} \min_i f^{(i)}$$

Нейронные сети

1. Однослойные модели

Пример: бинарная классификация



Нейрон = линейная часть + нелинейная активация

Процесс обучения:

- 1) нач. параметры \tilde{w} и b
 - 2) макс. оптим. значение $\tilde{w}^* + b^*$ $\xrightarrow{\text{GD}}$
 - 3) учн. $\hat{y} = \sigma(x \cdot w^* + b^*)$ в кон. приближение
- $$\alpha(\tilde{w}, b) = -[y \cdot \log \hat{y} + (1-y) \cdot \log(1-\hat{y})]$$
- $$\left\{ \begin{array}{l} \tilde{w} := \tilde{w} - \alpha \frac{\partial \ell}{\partial \tilde{w}} \\ b := b - \alpha \frac{\partial \ell}{\partial b} \end{array} \right.$$

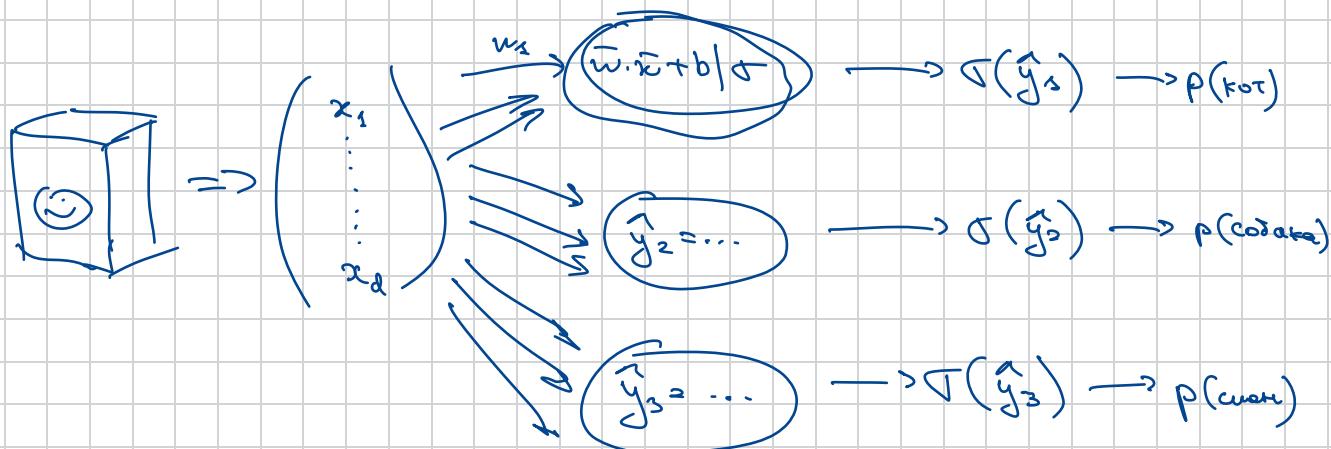
\hat{y} - прогноз модели
 y - исходная метка

История +

- 1943г. Мак-Каллоу, Питтс
- 1960г. Розенблатт (Макл 1, перцептрон)
- 1969г. Рашет, Медисон

{ AI Winter }

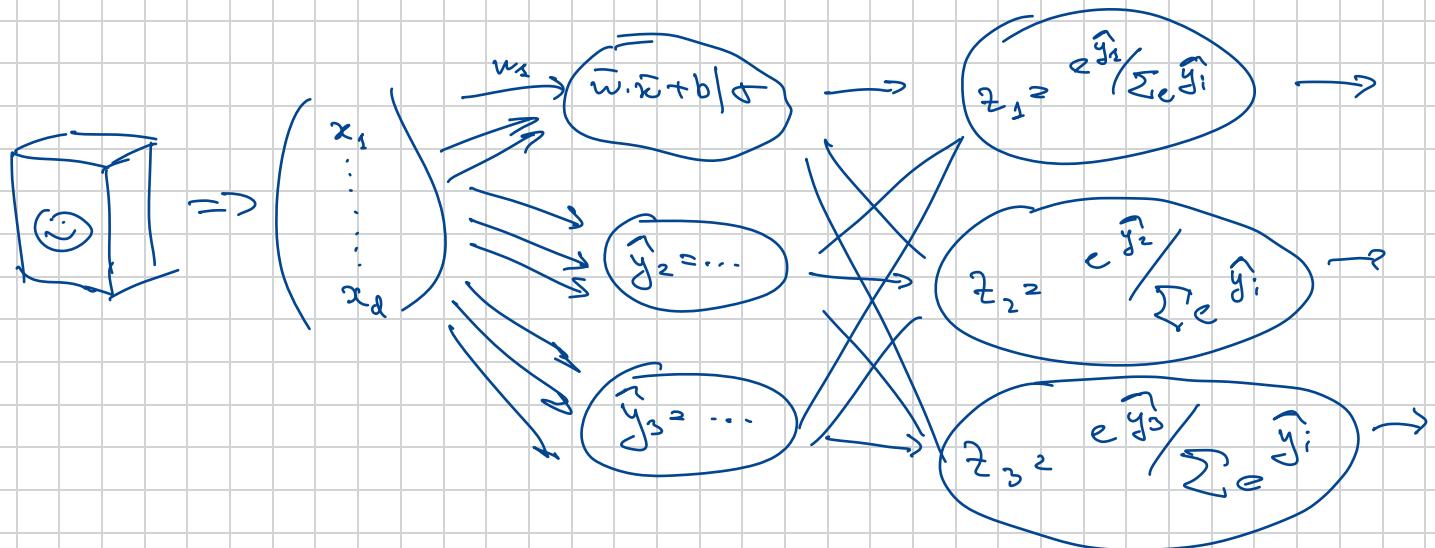
Пример: найти кошку, собаку, смока



$$L_3 = (w, b) = - \sum_{i=1}^n \sum_{j=1}^3 \left\{ y_j^{(i)} \cdot \log \hat{y}_j^{(i)} + (1 - y_j^{(i)}) \log (1 - \hat{y}_j^{(i)}) \right\}$$

нечетверичной сетью предела распространения

Fully Connected Feed Forward Neural Network

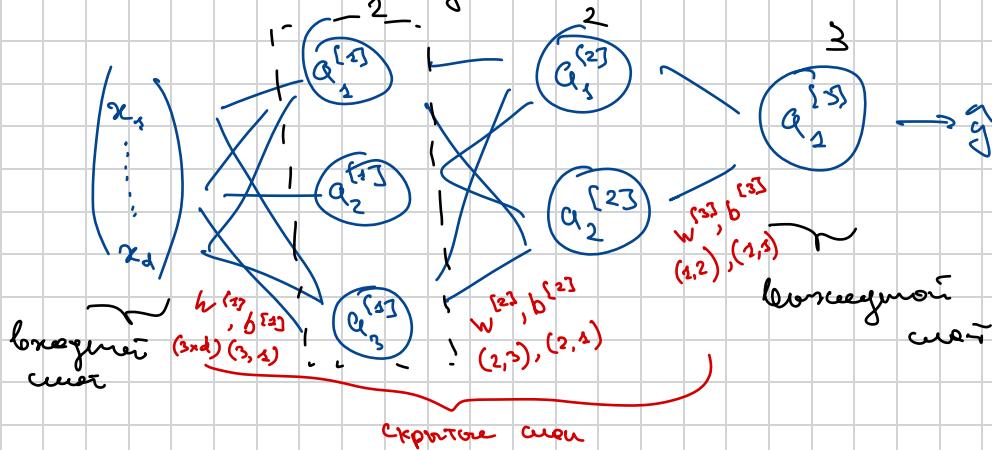


softmax

Например:

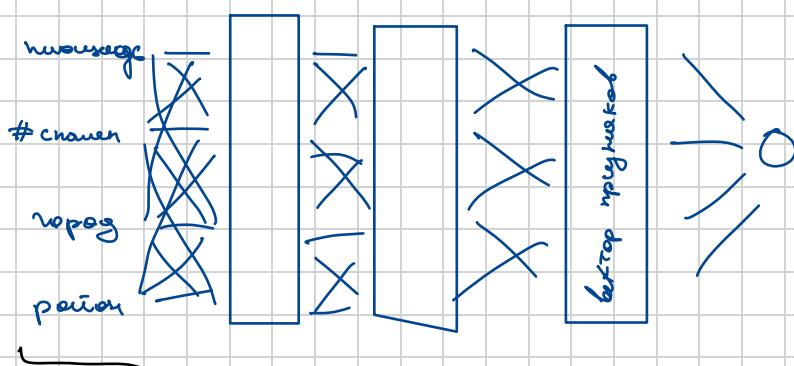
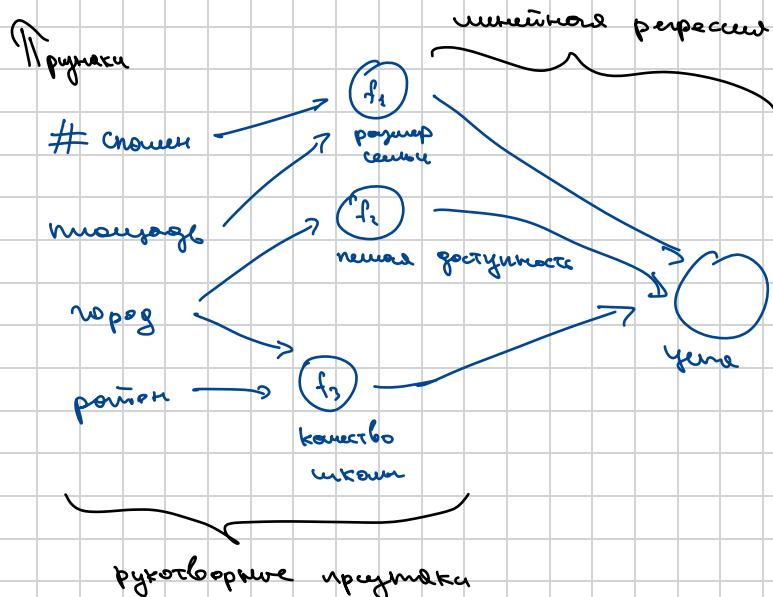
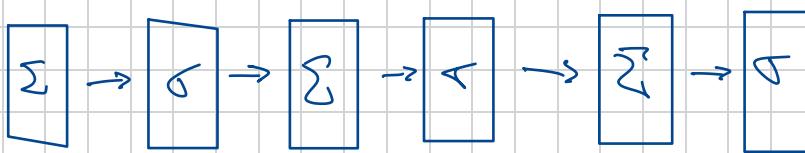
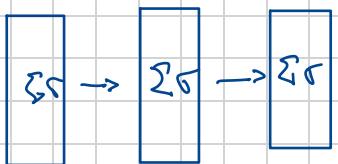
$$\begin{pmatrix} 0.7 \\ 0.2 \\ 0.1 \end{pmatrix} \text{ кот} \\ \text{собака} \\ \text{слон}$$

2. Многослойные модели



шар-шар-шар - это однотипные нейроны связанные между собой

модель = архитектура + параметры



representation learning

автоматическое
построение пространства признаков

последовательное
представление

Уравнение прямого распространения

Обозначение

$w_{ij}^{(l)}$ — вес связи j -го нейрона на $(l-1)$ -м слое в i -го на l -м слое

$b_j^{(l)}$ — сдвиг j -го на l -м слое

$z_i^{(l)}$ — линейная часть

$a_i^{(l)}$ — активация

$\hat{y} = b_0 + b_1 x$ сечу

$y = \text{vertical метка}$

$$\bar{x} \equiv a^{[x]} - b_{\text{key}}$$

$g(\cdot)$ - kreisförmig

Пустыня Картинок:

$$X = \begin{pmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(n)} \end{pmatrix}^T \Rightarrow \begin{pmatrix} | & & | \\ x^{(1)} & \dots & x^{(n)} \\ | & & | \end{pmatrix}$$

$$Z^{[z]} = w^{[z]} \sum_{(s,n)} + b^{[z]}$$

$(s,d) (d,n)$

(s,z)

подпись

Uncorrelated Ticks: broadcasting

3. Руководство активации

A hand-drawn diagram on lined paper showing a circle representing a particle. Inside the circle, the letter 'S' is written with an arrow pointing to it from the left, indicating spin. To the right of the circle, the letter 'g' is written, representing the magnetic moment.

а виноват кемеровского операцено маг з

Pregnancy, w/o $g(z) = z$

$$\hat{y} = a^{[3]} = w^{[3]} a^{[2]} + b^{[3]} = w^{[3]} (w^{[2]} a^{[2]} + b^{[2]}) + b^{[3]} = w^{[2]} (w^{[2]} (w^{[1]} x + b^{[1]}) + b^{[2]}) +$$

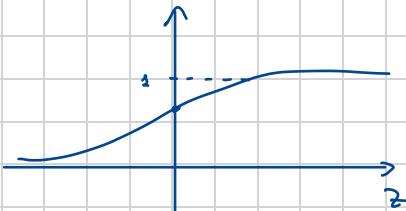
$$+ b^{[3]} = \underbrace{w^{[3]} w^{[2]} w^{[1]} x}_{w} + \underbrace{w^{[3]} w^{[2]} b^{[1]} + w^{[3]} b^{[2]} + b^{[3]}}_{b}$$

Быть нелинейностью буде сюда называемой

Руга:

1. Логистическая ф-я

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



a) $\sigma'(z) = \sigma(z) \cdot (1 - \sigma(z))$

б) симметрична $(-\infty; \infty)$ & $(0, 1)$

в) нелинейна

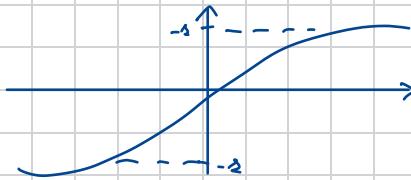
г) насыщается (они же насыщено!!)

если $z=10$, то $\sigma(z) \approx 0.99995$ $\sigma'(z) \approx 0.00004$

Общее насыщение (saturation regime)

2) Гиперболический тангенс

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



$\tanh'(z) = 1 - \tanh^2(z)$

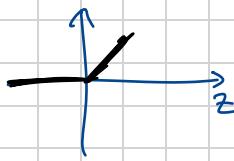
Симметрична $(-\infty, \infty)$ & $(-1, 1)$

— / —

3) ReLU (rectified linear unit)

$$\text{relu}(z) = \max\{0, z\}$$

- не линейн. & одн. +



$$\text{relu}'(z) = \begin{cases} 1, & \text{если } z > 0 \\ 0, & \text{если } z \leq 0 \end{cases}$$

4. Обратное распространение

1) Однослойные

2) нейр. веса $w^{[L]}, b^{[L]}, L=1, L$

2) bias. go изменения (не $\sigma!$, не const!)

$$w^{[L]} := w^{[L]} - \alpha \frac{\partial y}{\partial w^{[L]}}, L=3, L$$

$$b^{[L]} := b^{[L]} - \alpha \frac{\partial y}{\partial b^{[L]}}$$

Учн. членное правило:

$$f(g(h(x)))'_x = f'_g \cdot g'_h \cdot h'_x$$

$f \leftarrow g \leftarrow h$ один
x один на 1, то
настолько уменьшится h

Производное - чувствительность
функции к изм. ее
аргументов

Например $y = y(g_1, \dots, g_k)$

$$g_j = g_j(\theta_1, \dots, \theta_p)$$

$$\boxed{\frac{\partial y}{\partial \theta_i} = \sum_{j=1}^k \frac{\partial y}{\partial g_j} \cdot \frac{\partial g_j}{\partial \theta_i}}$$

Также правило 1

$$1) x \in \mathbb{R}, f \in \mathbb{R}, \frac{\partial F}{\partial x} \in \mathbb{R}$$

Например

$$2) x \in \mathbb{R}^n, f \in \mathbb{R}$$

$$\nabla_x f = \left(\frac{\partial F}{\partial x_1}, \dots, \frac{\partial F}{\partial x_n} \right) - \text{градиент}$$

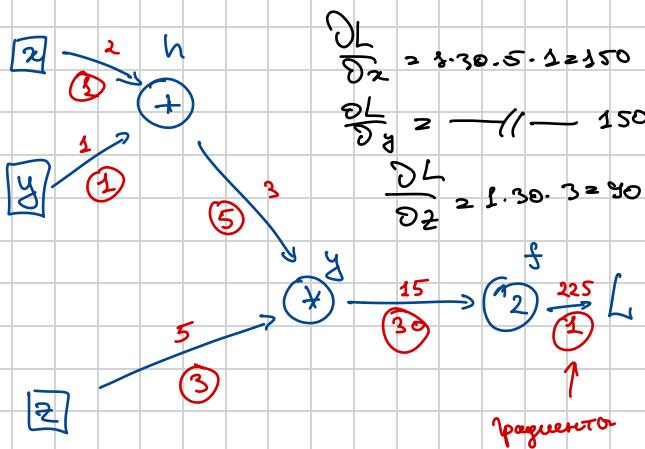
$$3) x \in \mathbb{R}^n, f \in \mathbb{R}^m$$

$y \in \mathbb{R}^{m \times n}$ - матрица логик

$$y = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \dots & \frac{\partial F_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F_m}{\partial x_1} & \dots & \frac{\partial F_m}{\partial x_n} \end{pmatrix}$$

Високоградиентное backprop

$$L(x, y, z) = [(x+y)+z]^2 \text{ при } x=2, y=3, z=5$$



Алгоритм Backprop

1. Прямой ход (forward)

2. Обратный ход

- идёт от L к аргументам
- включает окр. яп. в узлах
- производит окр. яп. по ходу сокращения

$y_{\text{ист}}:$

$$f(g) = g^2$$

$$g(h, z) = h \times z$$

$$h(x, y) = x + y$$

Лок. градиенты

$$\frac{\partial F}{\partial g} = 2g$$

$$\frac{\partial L}{\partial f} = 1$$

$$\frac{\partial g}{\partial h} = z$$

$$\frac{\partial g}{\partial z} = h$$

$$\frac{\partial h}{\partial x} = 1$$

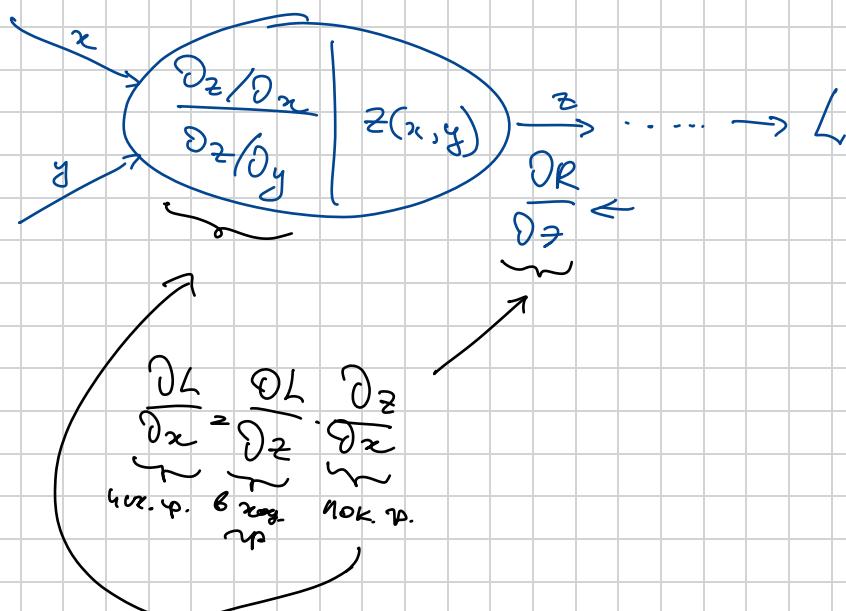
$$\frac{\partial h}{\partial y} = 1$$

Результат

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial h} \cdot \frac{\partial h}{\partial x}$$

$$\frac{\partial L}{\partial y} = \dots \cdot \frac{\partial h}{\partial y}$$

$$\frac{\partial L}{\partial z} = \dots \cdot \frac{\partial g}{\partial z}$$

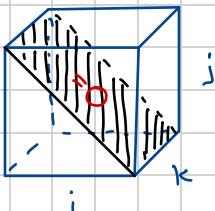


При мер

$$\frac{\partial L}{\partial w^{[2]}} = \underbrace{\frac{\partial L}{\partial z^{[2]}}}_{(2,1)} \cdot \underbrace{\frac{\partial z^{[2]}}{\partial w^{[2]}}}_{(2,2,3)} = \boxed{\frac{\partial L}{\partial z^{[2]}} \cdot a_i^{[2]T}}$$

коэффициент
столбца

$$\frac{\partial z^{[2]}}{\partial w_{j,k}^{[2]}} \neq 0, \text{ только если } i=j$$



$$y(x) = \sum_{i=1}^n L(\hat{y}^{(i)}, y^{(i)}), \text{ где}$$

$$L(\hat{y}^{(i)}, y^{(i)}) = -[y^{(i)} \cdot \log \hat{y}^{(i)} + (1-y^{(i)}) \cdot \log (1-\hat{y}^{(i)})]$$

$$\left\{ \begin{array}{l} \sigma' = \sigma(1-\sigma) \\ (\ln x)' = \frac{1}{x} \end{array} \right.$$

$$\begin{aligned} \frac{\partial L}{\partial w^{[2]}} &= - \left[y^{(i)} \frac{\partial}{\partial w^{[2]}} \underbrace{\log \sigma(w^{[2]} a^{[2]} + b^{[2]}) + (1-y^{(i)}) \frac{\partial}{\partial w^{[2]}} \log (1-\sigma(w^{[2]} a^{[2]} + b^{[2]}))}_{\hat{y}} \right] = \\ &= - \left[y^{(i)} \frac{1}{\hat{y}} \cdot \cancel{\hat{y}} \cdot (1-\hat{y}) a^{[2]T} + (1-y^{(i)}) \cdot \frac{1}{1-\hat{y}} \cdot (-1) \cancel{\hat{y}} \cdot (1-\cancel{\hat{y}}) \cdot a^{[2]T} \right] = \\ &= \dots (\hat{y}^{(i)} - y^{(i)}) a^{[2]T} \end{aligned}$$

$$\frac{\partial y}{\partial w^{[2]}} = \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)}) \cdot a_i^{[2]T}$$

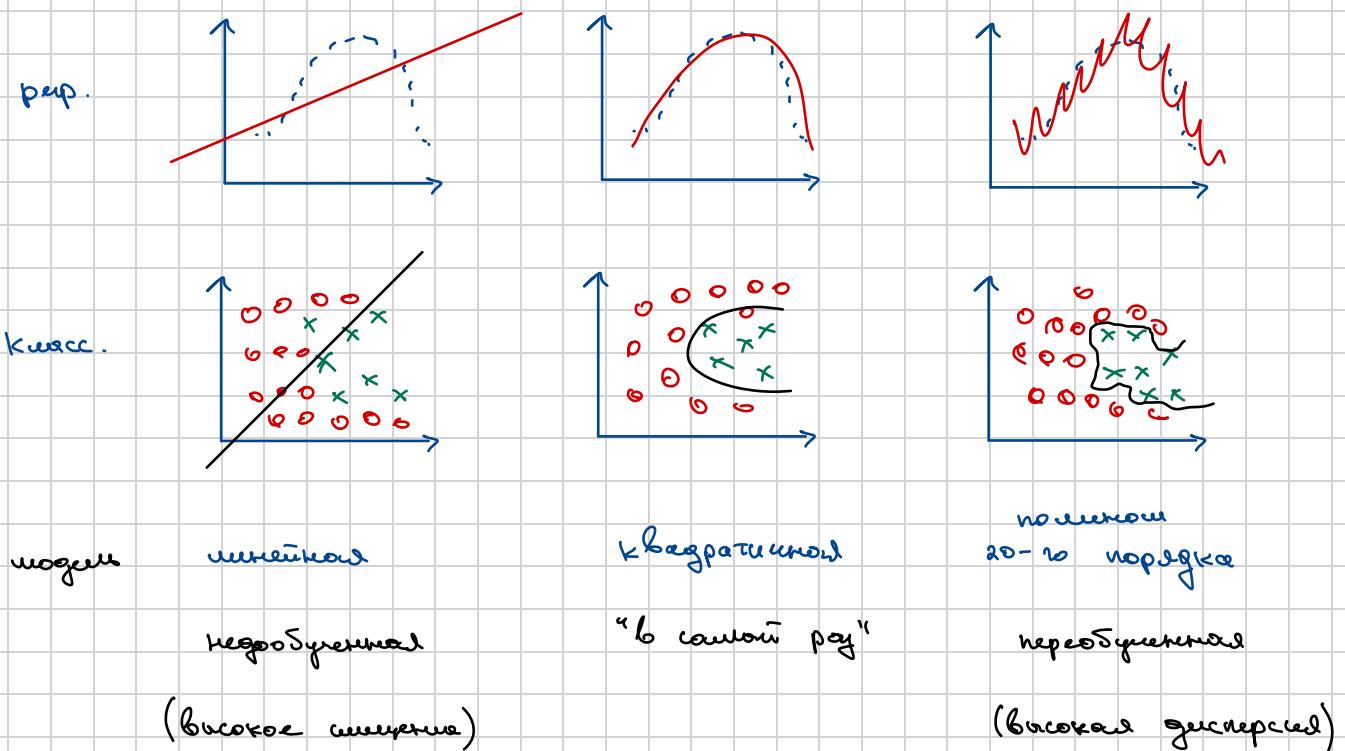
$$\begin{aligned} \frac{\partial L}{\partial w^{[2]}} &= \underbrace{\frac{\partial L}{\partial \hat{y}}} _{(y^{(i)} - \hat{y}^{(i)})} \cdot \underbrace{\frac{\partial \hat{y}}{\partial z^{[2]}}} _{w^{[2]T}} \cdot \underbrace{\frac{\partial z^{[2]}}{\partial a^{[2]}}} _{a^{[2]}} \cdot \underbrace{\frac{\partial a^{[2]}}{\partial w^{[2]}}} _{a^{[2]T}(1-a^{[2]T})} = \\ &\quad \underbrace{(\hat{y}^{(i)} \cdot y^{(i)})}_{(1,1)} \underbrace{w^{[2]T}}_{(2,2)} \underbrace{a^{[2]}}_{(2,1)} \underbrace{(1-a^{[2]T})}_{(1,2)} \cdot a^{[2]T} = \end{aligned}$$

$$\Rightarrow \underbrace{\frac{\partial L}{\partial w^{[2]}}}_{(2,2)} = \sum_{i=1}^n w^{[2]T} \cdot a_i^{[2]} \left(\frac{1}{1-a_i^{[2]}} \right) \cdot a_i^{[2]T}$$

Соотношение между предсказанием и фактическим значением. Регуляризация

① Bias-Variance Trade off

1. "Вдвух чюбах"



2. Ошибка оценки

$$J(\theta) = \frac{1}{n} \sum_{i=1}^n [y^{(i)} - h_\theta(x^{(i)})]^2$$

— ошибка на гр. выборке

Нас интересует:

$$L(\theta) = E[(y - h_\theta(x))^2] \quad \text{при } (x, y) \sim D$$

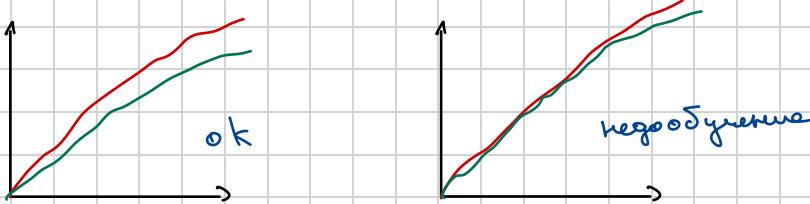
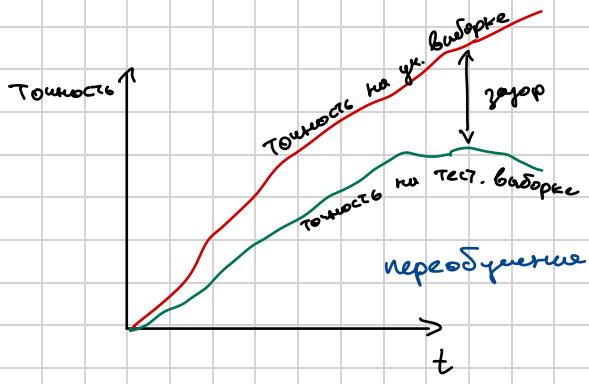
Можно оценить эмп.:

$$\{(x_{\text{test}}^{(1)}, y_{\text{test}}^{(1)}), \dots, (x_{\text{test}}^{(m)}, y_{\text{test}}^{(m)})\} \sim D \Rightarrow L(\theta) \approx \frac{1}{m} \sum_{i=1}^m (y^{(i)} - h_\theta(x^{(i)}))^2$$

приближ. модель в к. мира

Задача оценки: $J(\theta) - L(\theta)$

Несколько оценивается ком. модели не видение



Ошибка Смещение (bias) — ошибка на тестовой выборке, если учебная выборка бесконечного объема

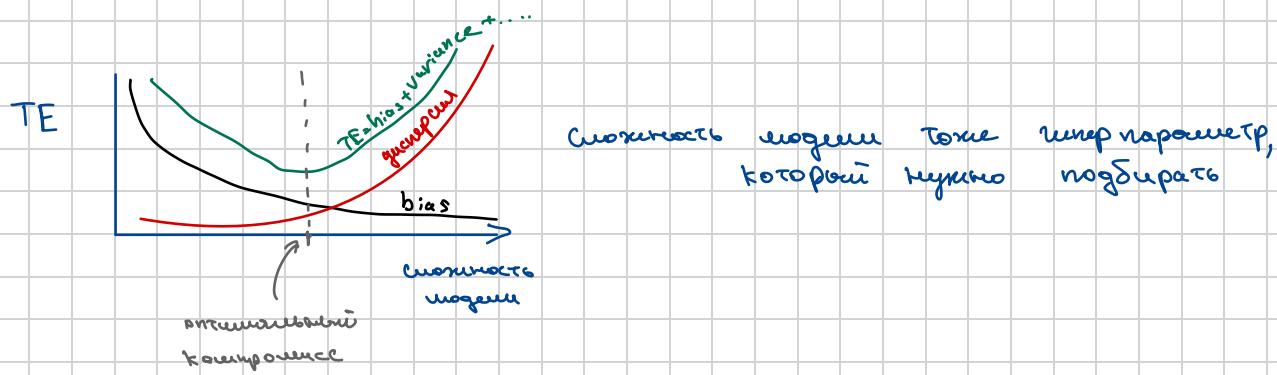
Ошибка Дисперсии — на шансах объема датчиков или уменьшении количества измерений

3. Компромисс смещения и дисперсии

Точность модели может управляться двумя способами:

1. увеличением размера учебной выборки

2. увеличением сложности мод



HO

в трудном обучении наил. фиксирует двойного спуска

в качестве

сложности

будет кол-во параметров



Due next. prep.

Модель

- генерируем $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, где $y^{(i)} = h^*(x^{(i)}) + \zeta^{(i)}$, $\zeta^{(i)} \sim N(0, \sigma^2)$
 - обучаем модель на S и получим h_s истинная модель
 - берём тестовый пример (x, y) , сгенерированный по той же правилам и найдём $E_{\zeta, y} \{ (y - h_s(x))^2 \}$

$$\hat{h}_{\text{ycro}} = h_{\arg(\mathbf{x})} = \mathbb{E}_s [h_s(\mathbf{x})]$$

Лемма: $A \cup B$ — независимое событие. Рассмотрим в $E[A] = 0$, то

$$E[(A+B)^2] = E[A^2] + E[B^2]$$

$$\text{(negativer E)} \quad E[(A+c)^2] = E[A^2] + c^2, \text{ wobei } c - \text{konstante Wk}$$

$$\begin{aligned}
 L(x) &= E_{s,z} \left\{ (y - h_s(x))^2 \right\} = E \left[z + (h^*(x) - h_s(x))^2 \right] = \underbrace{E[z^2]}_{\sigma^2} + E[(h^*(x) - h_s(x))^2] = \\
 &= \sigma^2 + E \left[\underbrace{(h^*(x) - h_{\text{avg}}(x))}_{\text{const}} + \underbrace{h_{\text{avg}}(x) - h_s(x)}_{\text{err.}} \right]^2 = \sigma^2 + \underbrace{(h^*(x) - h_{\text{avg}}(x))^2}_{\text{bias}} + \underbrace{\text{var}[h_s(x)]}_{\text{variance}}
 \end{aligned}$$

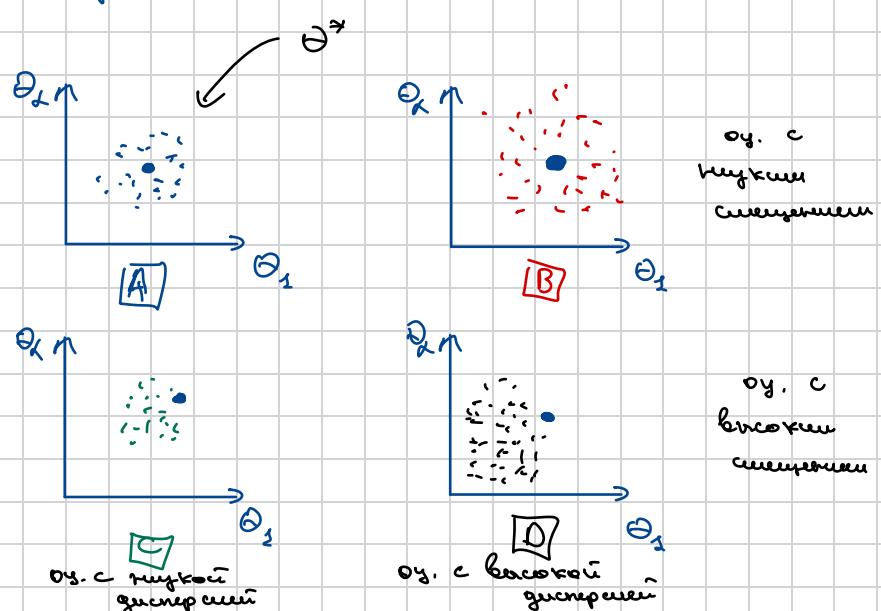
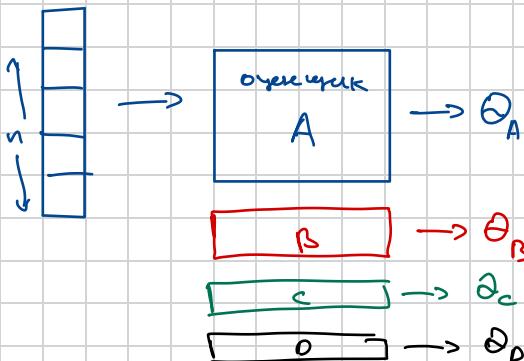
Cursive

Синтез наименуем классы методов относ. системной модели

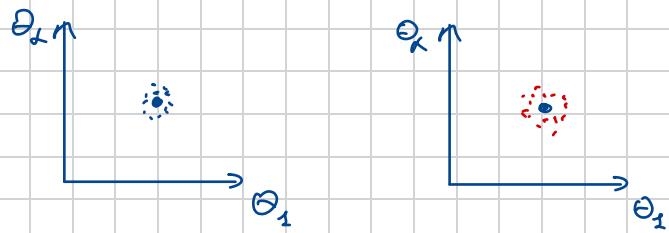
дисперсия — мадели на ин-ве ун. Воробка

5. Высокогорье

$$X \sim D(\theta^*)$$

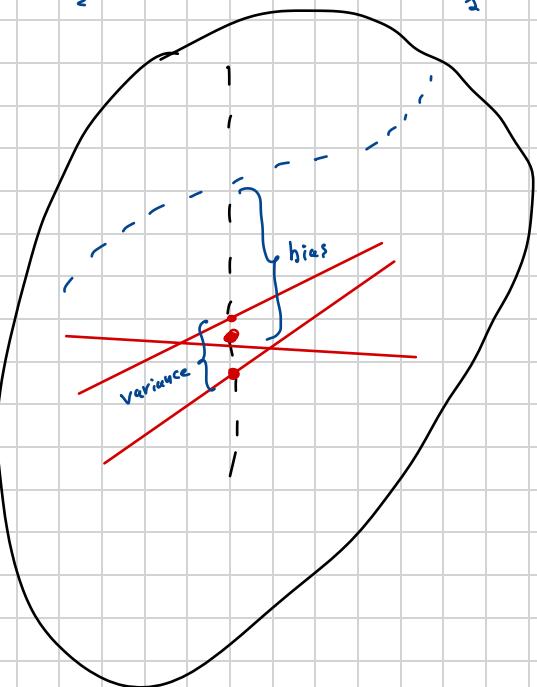
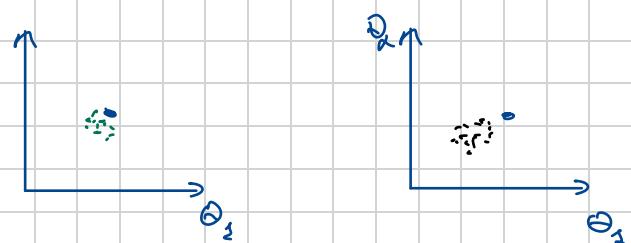
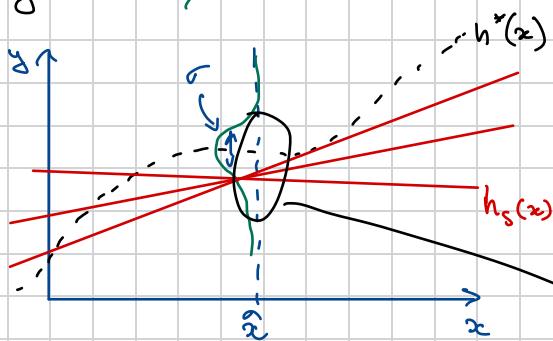


Возможен $N > n$



Для линейного обобщения

$$y = h^*(x) + \epsilon$$



② Валидация моделей

3. Разделение данных на выборки

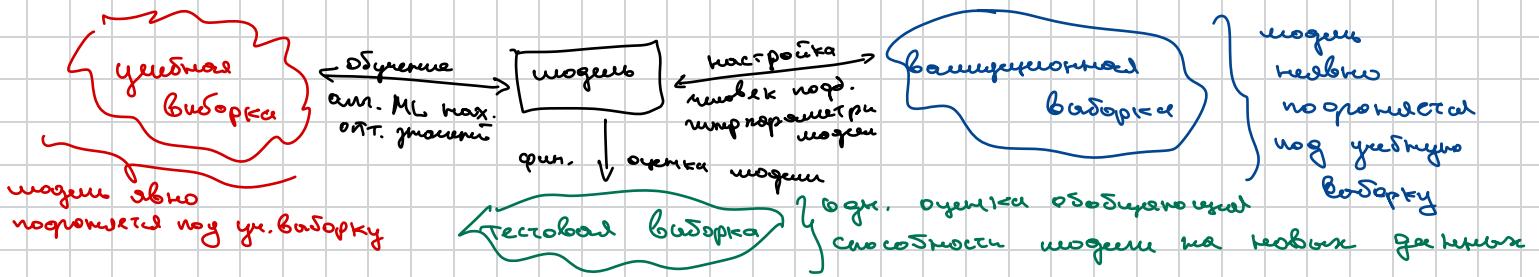
	train	val	test
train	60	70	80
dev	20	20	10
test	20	10	10

учебная выборка прав. выборка подтверждая выборка
прав. выборка Тестовая выборка

Пропорции:

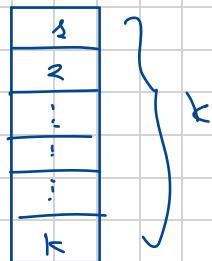
train	60	70	80	- ит. общ. обобщение
dev	20	20	10	- ит. общ. обобщение
test	20	10	10	- ит. общ. обобщение

подтверждение модели



2. Кросс-валидация

- 1) Применяется выборку
- 2) Оставшуюся разбивается на k -частей
- 3) Перебираем i от 1 до k



a) исч. все, кроме i -й части, в训. ун. Выборка
5) исч. i -ю часть для оценки точности

- 4) Изменение среднего у k точностей и вычисление
- матем. ожидания $k=n \Rightarrow$ Leave-one-out

38.50

③ Регуляризация моделей

$$J_\lambda = J(\theta) + \lambda R(\theta)$$

↓ регуляризация
 сума
 регуляризации

если $\lambda = 0$ не регуляризуется

свободные параметры

Будет:

$$3) R(\theta) = \|\theta\|_2^2 - l_2 \text{ норма (евклидово расстояние)}$$

если $\theta = (\theta_1, \dots, \theta_d)$; $\|\theta\|_2^2 = \sum_{i=1}^d \theta_i^2$

Линейная регрессия с L_2 -пер. лог. предикторов
регрессии (Ridge Regression)

Все параметры пер. одинаково. Нет предпочтений.

$$\theta = \begin{vmatrix} 1 & 1 & 1 & 1 \\ \theta_1 & \theta_2 & \dots & \theta_d \end{vmatrix}$$

$$2) R(\theta) = \|\theta\|_1 - l_1 \text{-норма (расстояние Manhattanских квартанов)}$$

$$\|\theta\|_1 = \sum_{i=1}^d |\theta_i|$$

Лин. пер. с L_1 -пер. лог. пер. LASSO (least absolute shrinkage and selection order)
регресс. LASSO, как np., удаляет некоторые веса

$$\begin{vmatrix} \dots & | & | & \dots & | & \dots \end{vmatrix}$$

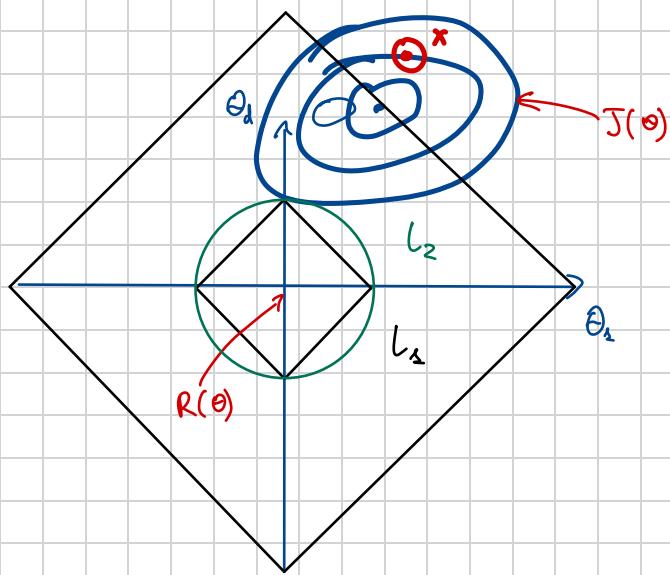
Пример:

$$x = (1, 1, 1, 1)$$

$$w_1 = (1, 0, 0, 0) \text{ - "негн." } L_1$$

$$w_2 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right) \text{ - "негн." } L_2$$

$$w_1^T x = w_2^T x$$



3) Байесовский метод поиска

$$S = \{(x_i, y_i)\}_{i=1}^n$$

$$P(S|Q) = \frac{P(S|\theta) \cdot P(\theta)}{P(S)} \Rightarrow \theta^* = \arg \max_{\theta} P(\theta|S) \approx \arg \max_{\theta} \frac{P(S|\theta) \cdot P(\theta)}{P(S)}$$

$$\propto \arg \max_{\theta} P(S|\theta) \cdot P(\theta) \propto \arg \max_{\theta} [\underbrace{\log P(S|\theta)}_{\text{автоматическоеование}} + \underbrace{\log P(\theta)}_{\text{регуляризация}}]$$

норм-привн. $\propto \lambda \cdot \|\theta\|_2^2$, если $P(\theta) \sim \mathcal{N}(0, \sigma^2)$

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \left\{ -\frac{(\theta)^2}{2\sigma^2} \right\}$$

- оптимизация (методом)

$$\text{MLE} = \arg \max_{\theta} P(S|\theta)$$

- байесовский

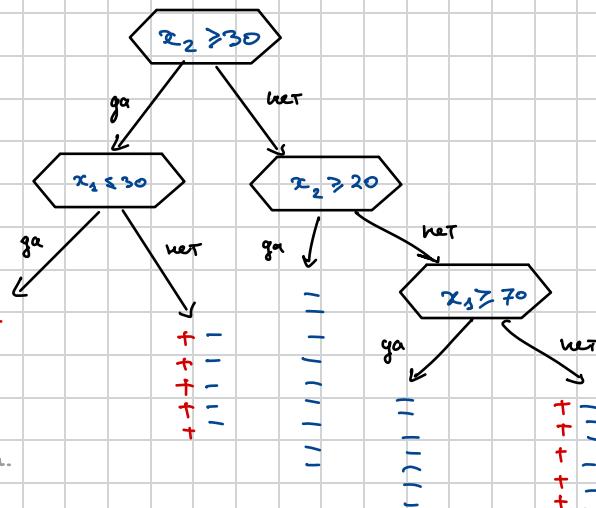
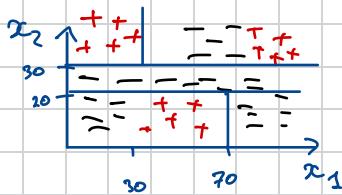
$$\text{MAP} = \arg \max_{\theta} P(\theta|S)$$

maximum a posteriori probability est.

9. Деревья принятия решений

(3) Деревья принятия решений

1) Пример задачи классификации



множества, рекурсивные, исходящие от

Алгоритм построения

1. Выбираем наилучший ТРЕТ

2. Делаем текущее эл-во примером на подмножество с его помощью

3. Рекурсивно повторяем для каждого получившегося эл-ва

Отличие от линейного классификатора!

Мы строим не одну решавшую границу, а целое дерево

У модели очень высокий дисперсия, она часто переобучается

Когда останавливается процесс обучения решавших деревьев

- 1) min размер чиста
- 2) max глубина
- 3) max кол-во узлов
- ⋮

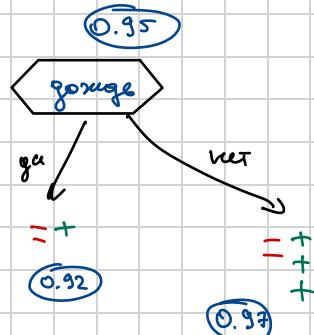
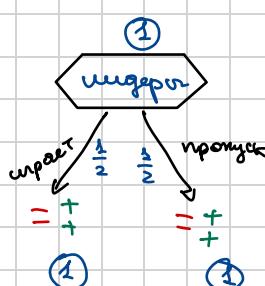
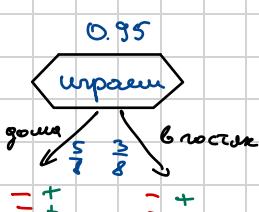
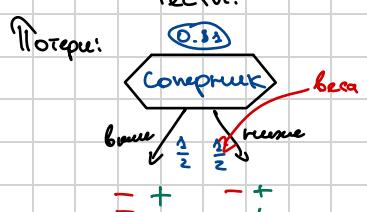
Работа алгоритма:

- 1) спускаемся по дереву, отвечая на вопросы
- 2) в чисте находим класс

2) Оценка текста (Проверка текста)

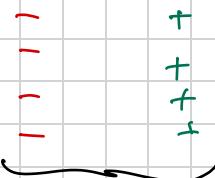
бесе	рэг	недор	зонг	результат
ga	гома	га	га	-
нет	гома	нет	нет	+
нет	гости	го	нет	+
гу	гости	нет	нет	-
га	гома	га	га	-
га	гома	га	нет	+
нет	гома	нет	гу	+
нет	гости	нет	нет	-

Тести:

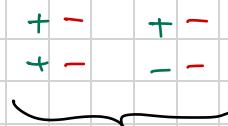


Эксперимент:

оценка качества - однородность ответа



полноценный порядок
наилучший вариант



наиболее худ
наихудший вариант

Уровень 1

Счит. кол-во элементов в однор. группах Плюс

Уровень 2

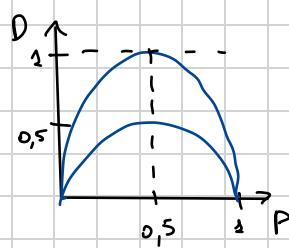
Пусть S -мн-бо, p -голос '+', n -голос '-'

$D(S)$ - мера хаоса мн-бо S

$$1) \text{ Экспрессия: } D(S) = p \cdot \log_2 \frac{1}{p} + n \cdot \log_2 \frac{1}{n}$$

$$2) \text{ Числ. выражение: } D(S) = p(s-p) + n(s-n)$$

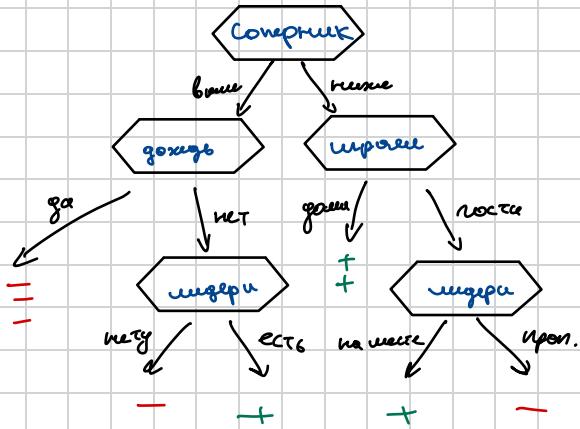
$$\text{Если } k \text{ классов, то } D(S) = \sum_{i=1}^k p_i \log \frac{1}{p_i}; D(S) = \sum_{i=1}^k p_i(1-p_i)$$



Потеря при разде теста

Пусть тест T разб. S на $\{S_1, \dots, S_t\}$

$$L(T) = \sum_{i=1}^t \frac{|S_i|}{|S|} \cdot D(S_i)$$



Минусы:

- Сложное представление
- высокая чувств.

Дерево регрессии

x_d	78	0	0	9 50
g_{10}	0	0	0	15 12
0	0	0	0	0
0	0	9	10	0
0	0	15	5	0

$$\text{Прим.: } \hat{y}_m = \frac{\sum_{i \in R_m} y_i}{|R_m|}$$

Ошибк.:

$$L(m) = \frac{\sum_{i \in R_m} (y_i - \hat{y}_m)}{|R_m|}$$

2. Аддитивн.

Способы:

3. Бэйтинг (Bootstrap Aggregation)

Пусть есть некотор. выборк. P и случайная выборка $S \sim P$

Предн. что $S = P$, генерируемые неб. ул. выборки: $z_1, z_2, \dots, z_k \sim \mathcal{Z}$

Обучаем k моделей

$$\text{И результатом будет модель: } G(x) = \frac{\sum_{i=1}^k G_i(x)}{k}$$

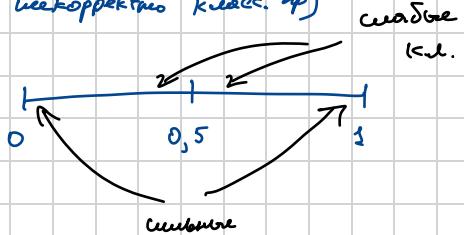
(равновероятно с повторениями)

Random Forest — Бэйтинг для DT, в кот. на каждом шаге выб. случайные независимые тесты

4. Бустинг (AdaBoost)

Пусть $h = \{-1, 1\}$. Козр. ошибка ε :

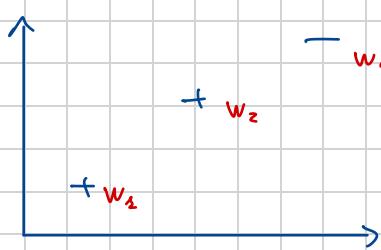
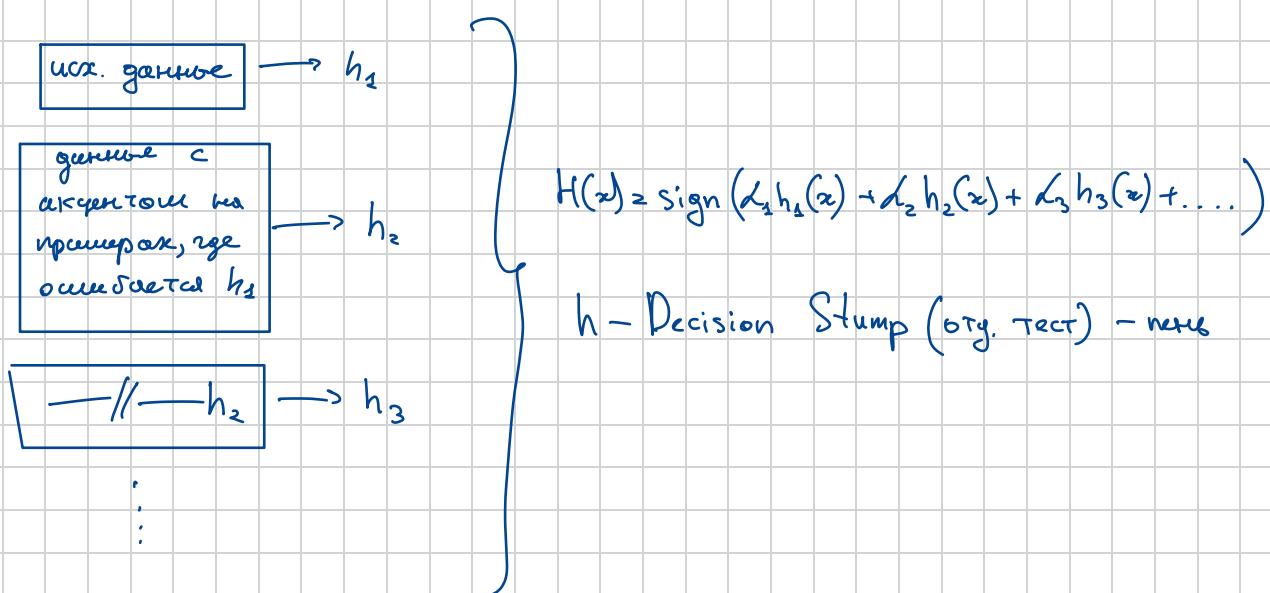
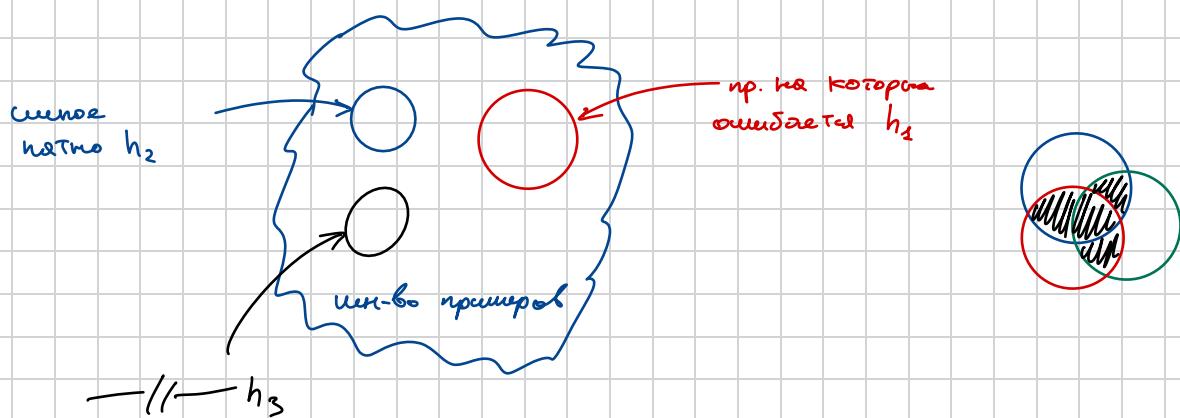
(напр. доп. некорректно класс. np.)



Гипотеза о бустинге

Можно ли соединить несколько классификаторов, обладающих некоторыми свойствами

$$H(x) = \text{sign} \{ h_1(x) + h_2(x) + h_3(x) \}$$



$$\varepsilon(\text{без весов}) = \sum_{\text{ошиб}} \frac{1}{N}, \text{ где } N-\text{к-во просперов бтг. выборке}$$

$$\varepsilon(\text{с весами}) = \sum_{\text{ошиб}} w_i, \text{ при ус.}$$

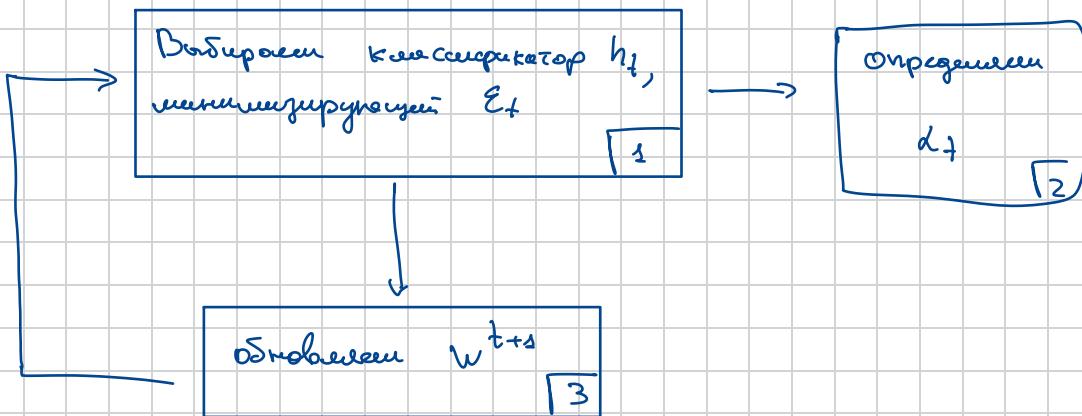
$$\left[\sum_{i=1}^N w_i = 1 \right]$$

w-веса!!!

на первом шаге:

$$w_i = \frac{1}{N}, i = 1, N$$

Общая схема:



Как обновлять веса?

$$w_i^{t+1} = \frac{w_i^t}{z_t} \cdot \exp \left\{ -d + h_t(z) y \right\}$$

↑
истинная метка
вероятн

$$z_t = \sum_i w_i^t \cdot \exp \left\{ -d + h_t(z) y \right\}$$

Если исти. метка $h(z)$, то получим:

$$d_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$$

Анализ

$$① w_i^{t+1} = \frac{w_i^t}{z_t} \cdot \begin{cases} \sqrt{\frac{\epsilon_t}{1-\epsilon_t}}, & \text{если не ошиблись} \\ \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}, & \text{если ошиблись} \end{cases}$$

$$② \text{Найдем } z_t: \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} \cdot \underbrace{\sum_{i \in \text{верно}} w_i^t}_{\text{верно}} + \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \cdot \underbrace{\sum_{i \in \text{ошибка}} w_i^t}_{\text{ошибки}} = 2\sqrt{\epsilon_t(1-\epsilon_t)}$$

Подставим ② в ①

$$w^{t+1} \rightarrow \begin{cases} \frac{w_i^t}{2} \cdot \frac{1}{1-\epsilon_t}, & \text{если верно} \\ \frac{w_i^t}{2} \cdot \frac{1}{\epsilon_t}, & \text{если ошиблись} \end{cases}$$

Если можем w^t подсчитать из первых при меров

$$\frac{1}{2} \cdot \frac{1}{1-\epsilon_t} \cdot \underbrace{\sum_{i \in \text{верно}} w_i^t}_{1-\epsilon_t} = \frac{1}{2} \Rightarrow \boxed{\sum_{i \in \text{верно}} w_i^t = \frac{1}{2}}$$

EM-алгоритм. Пакториеллі алгоритм

$p(x, z; \theta)$, z - скрытая

1. Старт. иниц. θ
2. Повторение до сходимости

"E-шаг"

$$Q_j(z) := p(z^{(i)} | x^{(i)}; \theta)$$

"M-шаг"

$$\theta := \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^n \text{ELBO}(x^{(i)}; \theta, Q)$$

$$\mathbb{E}_{z \sim Q} \left[\log \frac{p(x, z; \theta)}{Q(z)} \right]$$

1) EM для GMM

$$x \in \mathbb{R}^D$$

Модель $z \sim \text{Cat}(\phi) = \begin{bmatrix} \Phi_1 \\ \vdots \\ \Phi_k \end{bmatrix}$

$x | z \sim N(\mu_z, \Sigma_z)$

$p(x, z; \phi, \mu, \Sigma) = p(x|z; \mu, \Sigma) \cdot p(z; \phi)$

E-шаг:

$$2) w_j^{(i)} = Q_j(z^{(i)}=j) = p(z^{(i)}=j | x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)} | z^{(i)}=j; \mu, \Sigma) \cdot p(z^{(i)}=j; \phi)}{\sum_{l=1}^k p(x^{(i)} | z^{(i)}=l; \mu_l, \Sigma_l) \cdot p(z^{(i)}=l; \phi)}$$

M-шаг:

$$\underset{\phi, \mu, \Sigma}{\operatorname{argmax}} \sum_{j=1}^n \sum_{z^{(i)}} w_j^{(i)} \cdot \log \frac{\frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \cdot \exp \left\{ -\frac{1}{2} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right\}}{w_j^{(i)}} \boxed{\phi_j}$$

$$\nabla_{\phi} (\dots) = 0 \Rightarrow \phi = \dots$$

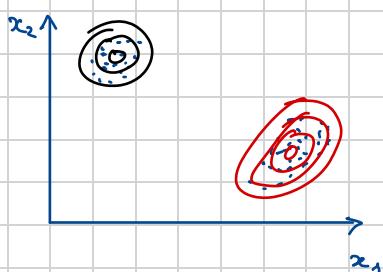
$$\nabla_{\mu_j} (\dots) = 0 \Rightarrow \mu_j = \dots$$

...

Распространение оценки

1) Мотивация

Пусть $d=2$, $n=500$
т.е. $n \gg d$



GMM подходит

Пусть $d \approx n$ или $n < d$

Например, $n=30$, $d=300$

$$\mathbf{x} = \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}_{100}, \rho(\mathbf{x})$$

Если хотим подогнать гауссово

$$\text{MLE: } \mu = \frac{1}{n} \sum_{i=1}^n \mathbf{x}^{(i)}$$

$$\Sigma = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}^{(i)} - \mu)(\mathbf{x}^{(i)} - \mu)^T$$

↔

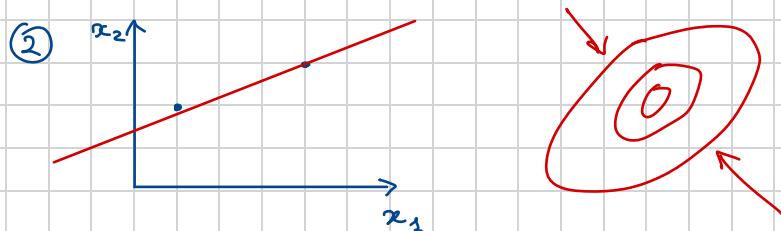
$$N(\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \cdot \exp \left\{ -\frac{1}{2} (\mathbf{x}^{(i)} - \mu)^T \Sigma^{-1} (\mathbf{x}^{(i)} - \mu) \right\}$$

Если $\bar{\alpha} \in \mathbb{R}^d$, то $\bar{\alpha} \bar{\alpha}^T \rightarrow$ то матрица размера $d \times d$ и ранга 1

Пример

① \mathbf{x}_2 ↑
• \mathbf{x}
 \mathbf{x}_1 $\mu = \frac{1}{1} \cdot \mathbf{x} = \mathbf{x}$

$$\Sigma = \frac{1}{3} (\mathbf{x} - \mathbf{x}) \cdot (\mathbf{x} - \mathbf{x})^T$$



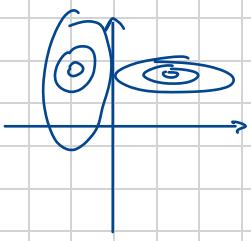
$$\begin{pmatrix} c & 0 \\ 0 & 0 \end{pmatrix}$$

2) Простые решения

Введем опр. на Σ

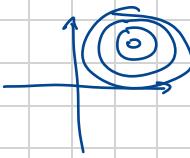
a) Пусть Σ будет диагональной

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & & & & 0 \\ & \sigma_{22}^2 & & \dots & \\ & & \ddots & & \sigma_{dd}^2 \end{pmatrix}$$



$$\text{Тонга: MLE } \hat{\sigma}_{jj}^2 = \frac{s}{n} \cdot \sum_{i=1}^n (x_j^{(i)} - \mu_j)^2$$

5) $\hat{\Sigma}_{\text{гетеро}} \quad \hat{\Sigma} = \sigma^2 I = \begin{pmatrix} \sigma^2 & & \\ & \sigma^2 & \\ & & \ddots & \\ & & & \sigma^2 \end{pmatrix}$



$$\sigma^2 = \frac{s}{n \cdot d} \sum_{i=1}^n \sum_{j=1}^d (x_j^{(i)} - \mu_j)^2$$

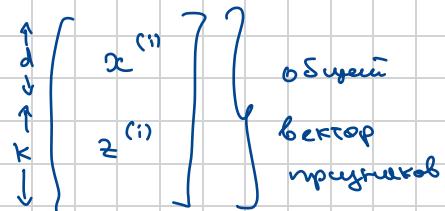
3) Модель с факторным анализом

$$z \in \mathbb{R}^k, k < n$$

$$x^{(i)} = (x_1^{(i)}, \dots, x_d^{(i)}) \quad z^{(i)} = (z_1^{(i)}, \dots, z_k^{(i)})$$

общая часть

специфическая



Модель:

$$p(x, z) = p(x|z) \cdot p(z)$$

Предположения:

$$z \sim N(0, I), z \in \mathbb{R}^k, k < d$$

$$x = \mu + L \cdot z + \epsilon, \text{ где } \epsilon \sim N(0, \Psi), \text{ где } \Psi - \text{ диагональная}$$

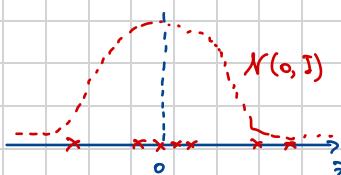
Параметры: $\mu \in \mathbb{R}^d, L \in \mathbb{R}^{d \times k}, \Psi \in \mathbb{R}^{d \times d}$ — диагональные

Модель переформулируем в виде:

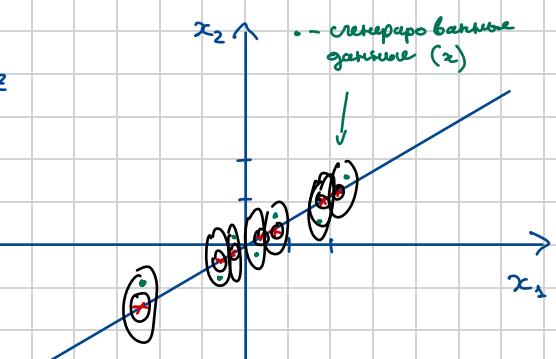
$$x|z \sim N(\mu + Lz, \Psi)$$

Пример 1

$$z \in \mathbb{R}^2, x \in \mathbb{R}^2, d=2 \\ k=1 \\ n=7$$



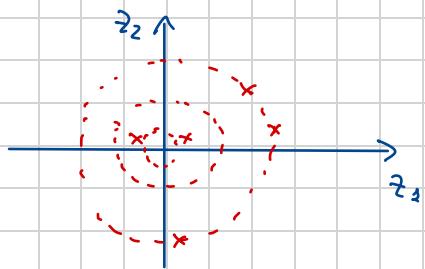
$$\mu + Lz \Rightarrow$$



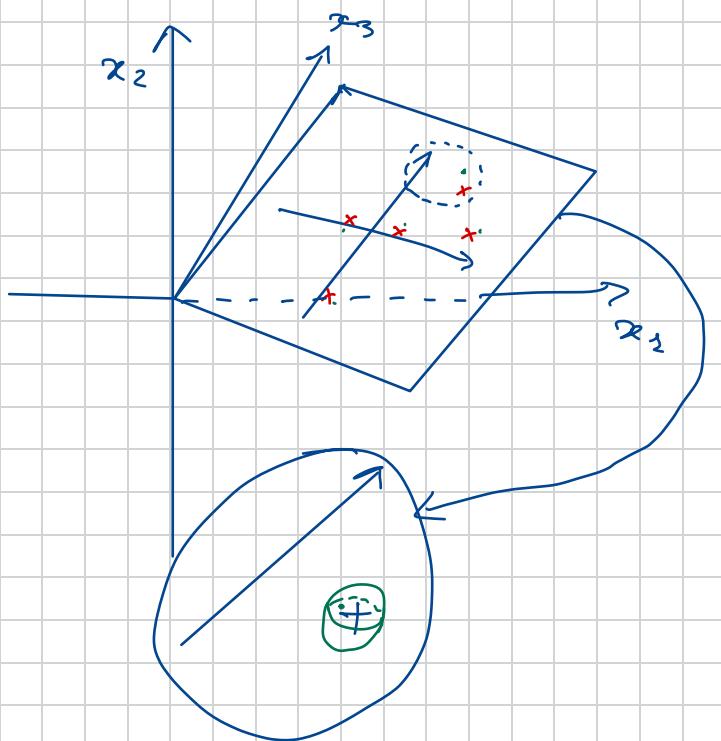
Пусть $L = \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}, \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

Пусть $\Psi = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \Rightarrow$
 \Rightarrow

Пример 2: $z \in \mathbb{R}^2, x \in \mathbb{R}^3$, $d=3$
 $k=2$
 $n=5$



$$\Rightarrow \mu, L_z \Rightarrow$$



4) Нек. об-ва умнож. норм. расп.

Пусть $x = \begin{bmatrix} x_1 \\ \vdots \\ x_r \\ x_{r+1} \\ \vdots \\ x_s \end{bmatrix}, x_1 \in \mathbb{R}^r, x_2 \in \mathbb{R}^s, x \in \mathbb{R}^{r+s}$

$$x \sim N(\mu, \Sigma) \Rightarrow \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_r \\ \mu_{r+1} \\ \vdots \\ \mu_s \end{bmatrix}, \Sigma = \begin{pmatrix} \sum_{11} & \sum_{12} \\ \vdots & \vdots \\ \sum_{r1} & \sum_{r2} \\ \hline \sum_{1r} & \sum_{2r} \\ \vdots & \vdots \\ \sum_{sr} & \sum_{ss} \end{pmatrix} \in \mathbb{R}^{r+s \times r+s}$$

$$p(x_1) = \int p(x_1, x_2) dx_2 = \int \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} \left(\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^\top \begin{pmatrix} \sum_{11} & \sum_{12} \\ \vdots & \vdots \\ \sum_{r1} & \sum_{r2} \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right) \right\} dx_2$$

$$\Rightarrow x_1 \sim N(\mu_1, \sum_{11})$$

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)} \Rightarrow x_1|x_2 \sim N(\mu_{1|2}, \sum_{1|2})$$

$$\mu_{1|2} = \mu_1 + \sum_{12} \sum_{22}^{-1} (x_2 - \mu_2)$$

$$\sum_{1|2} = \sum_{11} - \sum_{12} \sum_{22}^{-1} \sum_{21}$$

5) EM для FA

Определение $p(x, z)$

$$z \sim N(0, I)$$

$$\varepsilon \sim N(0, \psi)$$

$$x = \mu + Lz + \varepsilon$$

$$x|z \sim N(\mu + Lz, \psi)$$

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_z \\ \mu_x \end{bmatrix}, \Sigma\right)$$

$$\mu_{zx} = \begin{bmatrix} \mu_z \\ \mu_x \end{bmatrix}; \quad \mu_z = E[z] = 0$$

$$\mu_x = E[x] = E[\mu + Lz + \varepsilon] = \mu$$

$$\mu_{zx} = \begin{bmatrix} 0 \\ \mu \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \Sigma_{zz} & \Sigma_{zx} \\ \Sigma_{xz} & \Sigma_{xx} \end{bmatrix} = E\left[\begin{pmatrix} z - E_z \\ x - E_x \end{pmatrix} \begin{pmatrix} z - E_z \\ x - E_x \end{pmatrix}^T\right] = \begin{bmatrix} E\{(z - E_z)(z - E_z)^T\} \\ E\{(x - E_x)(z - E_z)^T\} \end{bmatrix}$$

$$\begin{bmatrix} E\{(z - E_z)(z - E_z)^T\} \\ E\{(x - E_x)(z - E_z)^T\} \end{bmatrix}$$

т.к. $z \sim N(0, I)$, то $\Sigma_{zz} = I$

$$\begin{aligned} \Sigma_{zx} &= E\{(z - E_z)(x - E_x)^T\} = E\left[\left(\cancel{\mu} + Lz + \varepsilon - \cancel{\mu}\right)\left(\cancel{\mu} + Lz + \varepsilon - \cancel{\mu}\right)^T\right] = \\ &= E\left[Lz z^T L^T + Lz \varepsilon^T + \varepsilon z^T + \varepsilon \varepsilon^T\right] = L \underbrace{E[z z^T]}_I L^T + E[\varepsilon \varepsilon^T] = LL^T + \psi \end{aligned}$$

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} I & L^T \\ L & LL^T + \psi \end{bmatrix}\right)$$

E-мод.

$$Q_i(z_i) = p(x^{(i)} | z^{(i)}, \Theta)$$

Параметры:

$$\Theta = (L, \mu, \psi)$$

$$z^{(i)} | x^{(i)} \sim \mathcal{N}(\mu_{z^{(i)}|x^{(i)}}, \Sigma_{z^{(i)}|x^{(i)}})$$

$$\mu_{z^{(i)}|x^{(i)}} = L^T (L L^T + \Psi)^{-1} (x^{(i)} - \mu)$$

$$\Sigma_{z^{(i)}|x^{(i)}} = I - L^T (L L^T + \Psi)^{-1} L$$

$$\text{M-max: } \underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^n \text{ELBO}(x^{(i)}, \Theta)$$

$$\sum_{i=1}^n \text{ELBO}(x^{(i)}, \Theta) = \sum_{i=1}^n E \left[\log \frac{p(x^{(i)}, z^{(i)}; \Theta)}{Q_i(z^{(i)})} \right] = \sum_{i=1}^n \left[\log p(x^{(i)}|z^{(i)}; \Theta) + \log p(z^{(i)}; \Theta) - \log Q_i(z^{(i)}) \right]$$

$$= \sum_{i=1}^n E \left[\log \frac{\frac{1}{2} \cdot \frac{1}{(2\pi)^{\frac{D}{2}} |\Psi|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} \cdot (x^{(i)} - \mu - L z^{(i)})^T \cdot \Psi^{-1} (x^{(i)} - \mu) \right\} \right] = \text{ne job. or } \Theta \text{ ne job. or } \Theta$$

$$= \sum_{i=1}^n E \left[-\frac{D}{2} \cdot \log(2\pi) - \frac{1}{2} \log |\Psi| - \frac{1}{2} (x^{(i)} - \mu - L z^{(i)})^T \cdot \Psi^{-1} (x^{(i)} - \mu - L z^{(i)}) \right] = J(\Theta) \quad \text{ne job. or } \Theta$$

$$\nabla_L J = 0 \Rightarrow$$

$$\nabla_\mu J = 0 \Rightarrow$$

$$\nabla_\Psi J = 0 \Rightarrow$$

$$\mu = \frac{1}{n} \sum_{i=1}^n x^{(i)}$$

$$L = \left(\sum_{i=1}^n (x^{(i)} - \mu) \cdot \mu_{z^{(i)}|x^{(i)}} \right) \cdot \left(\sum_{i=1}^n \left[\mu_{z^{(i)}|x^{(i)}} \cdot \mu_{z^{(i)}|x^{(i)}}^T + \Sigma_{z^{(i)}|x^{(i)}} \right] \right)^{-1}$$

$$\begin{aligned} \Psi &= \frac{1}{n} \sum_{i=1}^n \left[x^{(i)} \cdot x^{(i)T} - x^{(i)} \cdot \mu_{z^{(i)}|x^{(i)}}^T \cdot L^T - L \cdot \mu_{z^{(i)}|x^{(i)}} \cdot x^{(i)T} + L \left(\mu_{z^{(i)}|x^{(i)}} \cdot \mu_{z^{(i)}|x^{(i)}}^T + \right. \right. \\ &\quad \left. \left. + \Sigma_{z^{(i)}|x^{(i)}} \right) L^T \right] \end{aligned}$$

$$\Psi_{ii} = \Psi_{ri}$$

PCA. ICA

1) Метод главных компонент (Principal Component Analysis)

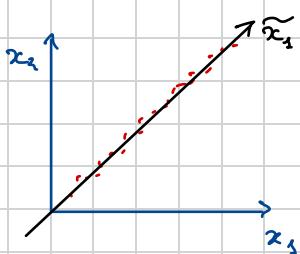
1. Мотивация

Задача: $\{x^{(1)}, \dots, x^{(n)}\} \in \mathbb{R}^d$

Требуется: понизить разнородность данных: $d \rightarrow k$, где $k \ll d$

т.е. это задача пониж. разнородности нр-ва признаков

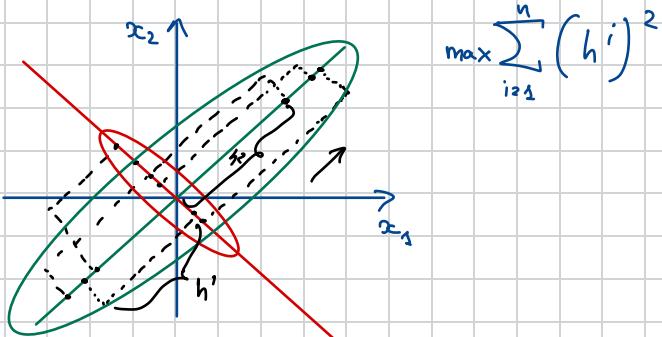
Пример:



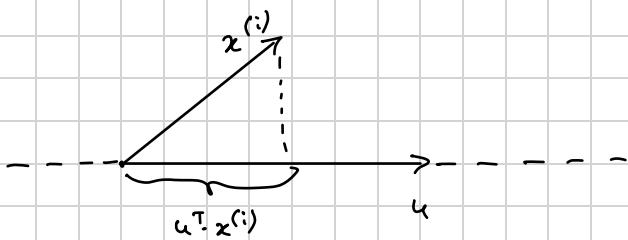
2. Предварительная обработка данных

$$\left. \begin{array}{l} a) \mu = \frac{1}{n} \sum_{i=1}^n x^{(i)} \\ b) \sigma_j^2 = \frac{1}{n} \sum_{i=1}^n (x_j^{(i)} - \mu_j)^2 \\ c) \tilde{x}_j^{(i)} = \frac{x_j^{(i)} - \mu_j}{\sqrt{\sigma_j^2}} \end{array} \right\} \begin{array}{l} \text{центрирование и} \\ \text{стандартизация} \\ \text{дисперсии} \\ \text{данных} \\ (\text{нормализация} \\ \text{данных}) \end{array}$$

3. Реализация PCA



Проекция вектора \bar{u} , $\|\bar{u}\| \geq 1$



$$\max_{u: \|u\|=1} \frac{1}{n} \sum_{i=1}^n (u^T x^{(i)})^2 = \underbrace{\frac{1}{n} \sum_{i=1}^n u^T x^{(i)} x^{(i)T} u}_{\sum} = u^T \left(\frac{1}{n} \sum_{i=1}^n x^{(i)} x^{(i)T} \right) u \Rightarrow \max_{u: \|u\|=1} u^T \sum u$$

$$\mathcal{L}(u, \lambda) = u^T \sum u - \lambda(u^T u - 1)$$

$$\nabla_u \mathcal{L} = \sum u - \lambda u = 0 \Rightarrow \boxed{\sum u = \lambda u}, \text{ т.е.}$$

u — собств. вектор

λ — собств. число

\sum — вад. квад. матрица
(при усн. квадратного среднего). Она
симметрична, поэтому телесное
популярное явление

Т.о. подпр-я штк. сортировка базируется на основе, где есть цел. числовые собств.

Векторное кв. матрицы основе

Алгоритм:

- 1) Нак. собств. числа/вектора для кв. матрицы основе
- 2) Сортируем λ по убыванию
- 3) берём верхнее (первое) λ к векторам u_1, \dots, u_k
- 4) проинициализируем выборку в новое k -мерное подпр-е:

$$x^{(i)} \rightarrow \underbrace{(u_1^T x^{(i)}, u_2^T x^{(i)}, \dots, u_k^T x^{(i)})}_\text{Симметрическая координаты в новом базисе} = g^{(i)}$$

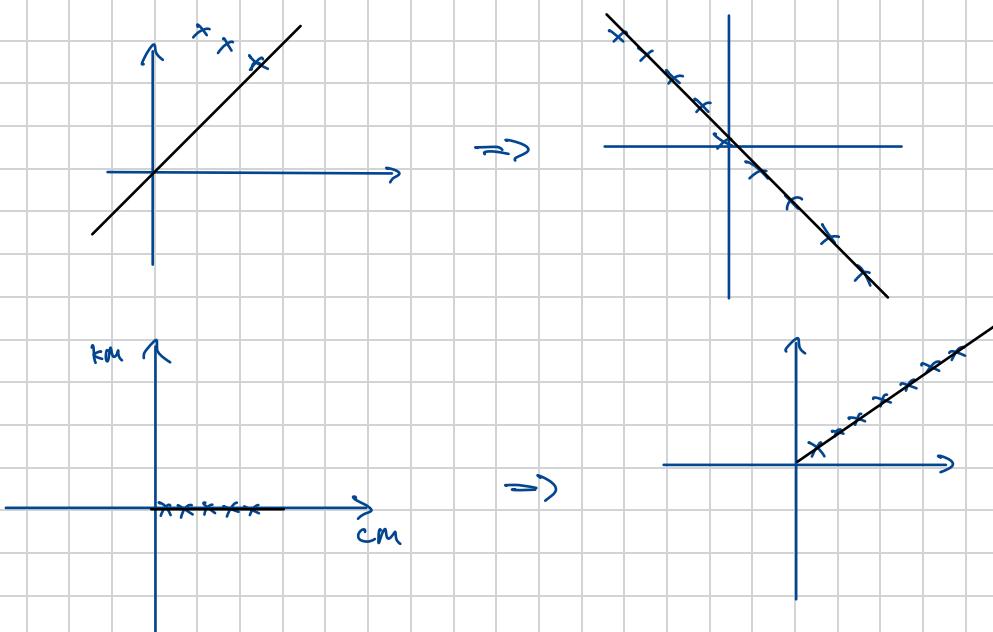
Симметрическая
координаты
в новом базисе

Одномерное представ:

$$x^{(i)} \approx \underbrace{y_1^{(i)} \cdot \bar{u}_1 + y_2^{(i)} \cdot \bar{u}_2 + \dots + y_k^{(i)} \cdot \bar{u}_k}_{\text{склер}} \Rightarrow \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix} \approx 2 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 3 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + 5 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

одномерный
вектор

Принцип:



4. Применение PCA

- Выделяющие факторы (просмотрование в 2-х или 3-х мерном пространстве)
- сжатие (изврн., разм.): где повышают фп. ML

Как выбрать k?

$$\frac{\lambda_1 + \lambda_2 + \dots + \lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_d} = 0,9 \Rightarrow k \text{ сопр. } 90\% \text{ вариации в данных}$$

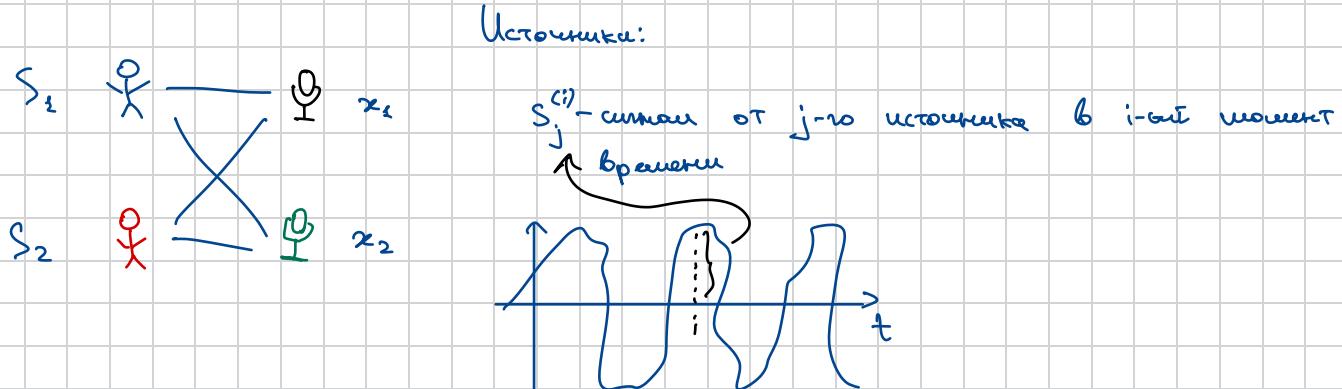
2. Сравнение методов обучения для классификации

	Вероятностное (моделирование P(x), детерминантное анализ)	Не вероятностное (сжатие / сущн. разнородности, выделение)
Данные "живут" в напространстве меньшей размерности	Рейторный анализ	PCA
Данные "живут" в кластерах	Модель смеси гауссеан (GMM)	K-средних

3) Метод независимых компонент

а) Модель

Проблема "мульти-корреляции". Пусть есть d источников сигналов, d микрофонов:



Модели

$$x^{(i)} = A \cdot s^{(i)}$$

$$S^{(i)} = (S_1^{(i)}, S_2^{(i)}, \dots, S_d^{(i)})$$

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

Модели:

$$x^{(i)} = A \cdot s^{(i)}$$

A — матрица смещивания

Но знаем только $x^{(i)}$

A и $S^{(i)}$ — неизвестны

Чтоб: найти $W = A^{-1}$, такое что $S^{(i)} = W \cdot x^{(i)}$

$$S^{(i)} = W \cdot x^{(i)}$$

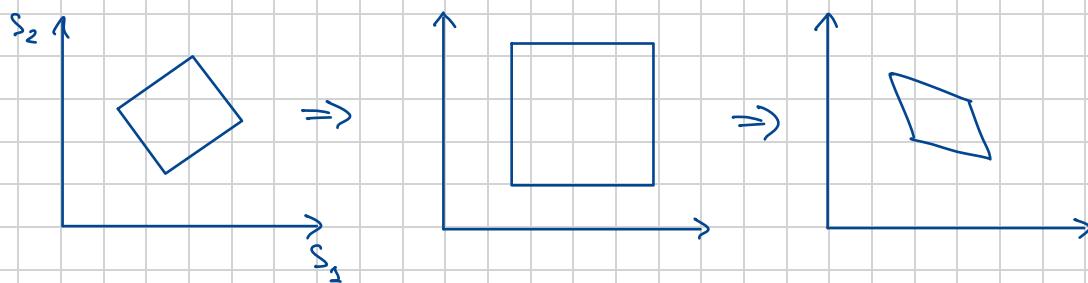
Предположения:

1) Кол-во источников \geq кол-во микрофонов

2) $x = As$

3) Источники авт. независимы

4) Источники не авт. гауссовские



3) Метод независимых компонент

2) Преобразование плотности

Пусть с. ве. \$S\$ имеет плотность \$p_S(s)\$

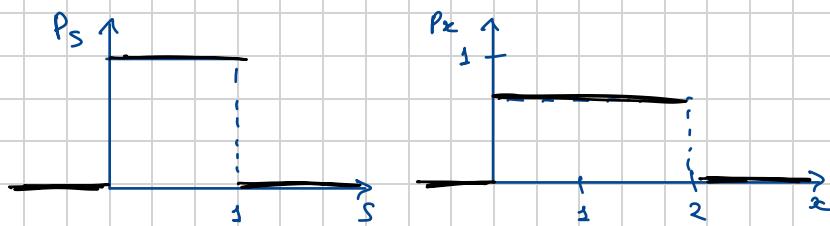
Пусть теперь \$x = A \cdot s\$. Чему равно \$p_x(x)\$?

Пусть \$w = A^{-1}\$ \$s = wx \Rightarrow p_S(s)

$$\boxed{p_x(x) \neq p_S(wx)}$$

$$S \sim u[0,1];$$

$$p_S(s) = \mathbb{1}_{\{0 \leq s \leq 1\}}. \text{ Пусть } A=2, x=2s$$



$$\boxed{\int p_x(x) dx = 1}$$

Корректная формула: \$p_x(x) = p_S(wx) \cdot |w|\$, где \$|w|\$ - определитель матрицы \$w\$

В общем случае здесь будет Якобиан

3. Виды ICH

Пусть \$S_j\$ является нек. расп. \$p_S\$

$$\text{Тогда } p(s) = \prod_{j=1}^d p_S(s_j)$$

$$\boxed{p_x(x) = p_S(wx) \cdot |w| = \prod_{j=1}^d p_S(w_j^T \cdot x) \cdot |w|}, \text{ где } w_j - j-\text{ая строка } w$$

Возможно \$p_S\$ - чисто линейное



\$\hookrightarrow F(x) = \Gamma(x)\$, где \$\Gamma(x)\$ - чисто линейная ф-я

$$p_s(x) = F'(x)$$

$$L(w) = \sum_{i=1}^n \left(\sum_{j=1}^d \log \sigma(w_j^T x^{(i)}) + \log |w| \right)$$

Вик. $\nabla_w L$ є нан. методом розмежування

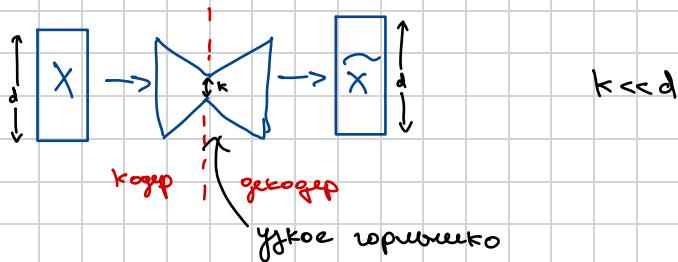
В методі згортки $x^{(i)}$: $w := w + d \begin{pmatrix} 1 - 2\sigma(w_1^T x^{(i)}) \\ 1 - 2\sigma(w_2^T x^{(i)}) \\ \vdots \\ 1 - 2\sigma(w_d^T x^{(i)}) \end{pmatrix} x^{(i)T} + (w^T)^{-1}$

Точне означення:

$$S^{(i)} = w \cdot x^{(i)}$$

Автоэнкодер

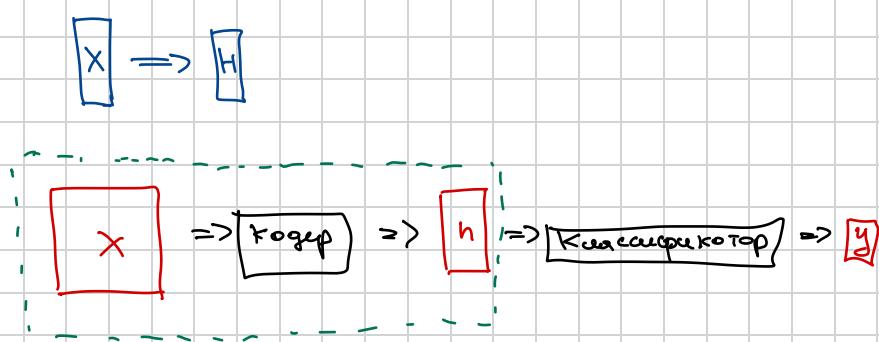
Нейронка для уменьшения



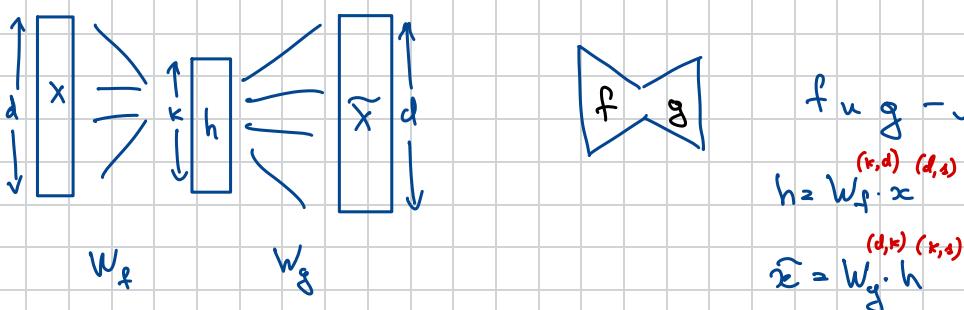
h -вектор скрытых признаков, скрытое представление

Задачи:

- сжатие (сформирование представления)
- уточнение изображ.
- преобразование представления данных
- генерация новых данных



Многократное автоэнкодер



$f \cup g$ — линейные функции

$$h = W_f \cdot x^{(r,d)} \cdot (d,s)$$

$$\tilde{x} = W_g \cdot h^{(d,r)} \cdot (s,s)$$

$$\text{Критерий: } \min_w \frac{1}{2} \sum_{i=1}^n \|w_g \cdot w_f \cdot x^{(i)} - x^{(i)}\|_2^2$$

$$\widehat{x} = g(f(x))$$

Несколько способов

$f \circ g$ - неинвертируемо

$$h = f(x; w_f)$$

$$\widehat{x} = g(x; w_g)$$

$$\text{Критерий: } \min_w \frac{1}{2} \sum_{i=1}^n \|g(f(x^{(i)}; w_f); w_g) - x^{(i)}\|_2^2$$

без. L_2

L_2 -норма

Разрешенное представление

$$\begin{array}{c|c|c|c} x & \equiv & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} & \equiv \tilde{x} \\ \hline h & & & \end{array}$$

$$\min_w \frac{1}{2} \sum_{i=1}^n \|g(f(x^{(i)}; w_f); w_g) - x^{(i)}\|_2^2 + \underbrace{\lambda \cdot \text{nnz}(f(x^{(i)}; w_f))}_{\begin{array}{l} \text{разрешенное} \\ \text{число ненулевых} \\ \text{элементов nnz(z)} \end{array}}$$

$$\min_w \frac{1}{2} \sum_{i=1}^n \|g(f(x^{(i)}; w_f); w_g) - x^{(i)}\|_2^2 + \lambda \cdot \|f(x^{(i)}; w_f)\|_1$$

Удешевление нулей

x - "чистое" значение

x' - "зашумленное" значение

$$\min_w \frac{1}{2} \sum_{i=1}^n \|g(f(\underline{x}^{(i)}; w_f); w_g) - x^{(i)}\|_2^2$$

Вариационный автоэнкодер

func 60%р. условные распр. вероятностей

$$f = p(h|x; W_f) \quad g = p(x|h; W_g)$$



$$\begin{array}{c}
 \boxed{x} = \boxed{} = \boxed{} = \boxed{\begin{matrix} x \\ x \\ \vdots \\ x \end{matrix}} \xrightarrow{\sim} \boxed{} = \boxed{} = \boxed{\begin{matrix} \phi \\ x \\ \vdots \\ 0 \end{matrix}} \\
 \phi \in (0,1) \\
 \text{бесконечн.}
 \end{array}$$

$$\begin{array}{c}
 \boxed{\begin{matrix} x \\ \vdots \\ x \end{matrix}} = \boxed{} = \boxed{} = \boxed{\begin{matrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{matrix}} \xrightarrow{2k} \boxed{} = \boxed{} = \boxed{\begin{matrix} \hat{x}_1 \\ \hat{x}_2 \\ \vdots \\ \hat{x}_d \end{matrix}} \\
 \mu \in \mathbb{R} \\
 \sigma^2 \in \mathbb{R}^+
 \end{array}$$

$$\begin{array}{c}
 \text{Нормальное} \\
 \hat{x} \\
 \boxed{\begin{matrix} \hat{x}_1 \\ \hat{x}_2 \\ \vdots \\ \hat{x}_d \end{matrix}}
 \end{array}$$

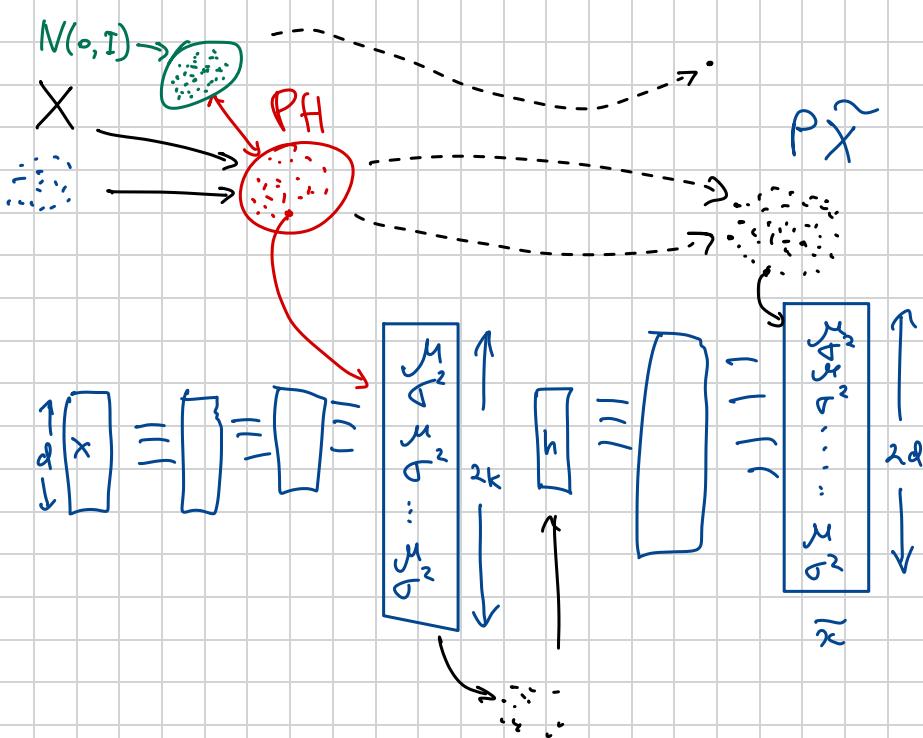
$$\begin{array}{c}
 N(x|h; \mu, \sigma^2) \\
 \text{беско } \Sigma = \begin{pmatrix} \sigma_1^2 & & & 0 \\ & \ddots & & \\ 0 & & \ddots & \sigma_d^2 \end{pmatrix}
 \end{array}$$

$$N(h|x; \mu, \sigma^2)$$

Генеративная модель

- генерируется $h \sim p(h)$
- генерируется $\hat{x} \sim p(x|h; W_g)$

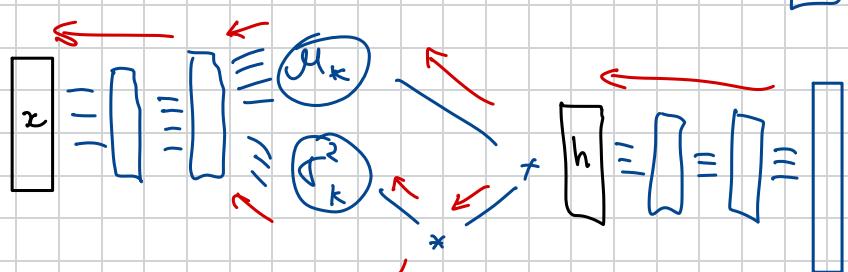
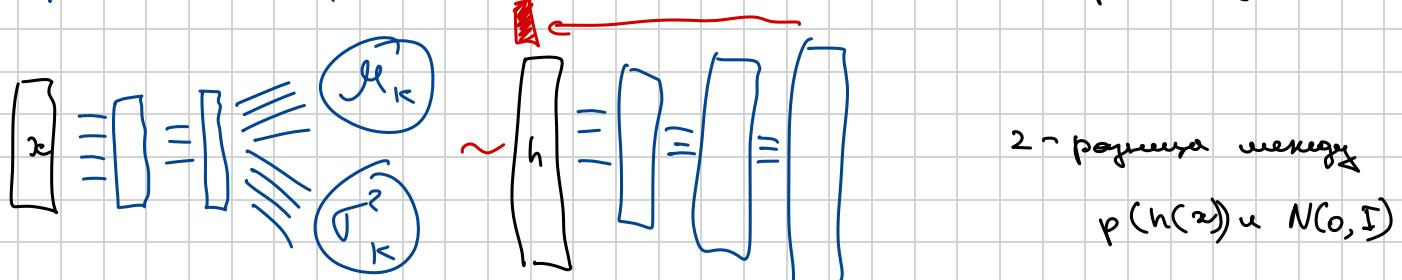
$$\text{Пусть } p(h) = N(0, I)$$



Число 6 оптимизационное -ование

$$\max_w \frac{1}{2} \sum_{i=1}^n \underbrace{\log p(x^{(i)}; w)}_1 - c \cdot \underbrace{KL(p(h|x^{(i)}; w_f) || N(0, I))}_2$$

Трик репрессоров



$$N(\mu, \sigma^2) = \mu + \sigma^2 \cdot N(0, I)$$

$$\Theta = w_g, \phi = w_f, z = h$$

$$p_\Theta(x) = \int p_\Theta(x, z) dz = \int p_\Theta(z|x) \cdot p(x) dz$$

$$p_\Theta(z) = \frac{p_\Theta(z|x) \cdot p(x)}{p(x|z)} \approx \frac{p_\Theta(z|x) \cdot p(x)}{q_\phi(z|x)}$$

$$p_{\theta}(x) = \frac{p_{\theta}(x|z) \cdot p(z)}{p(z|x)}$$

$$\log p_{\theta}(x) = \log \frac{p_{\theta}(x|z) \cdot p(z)}{p(z|x)} \cdot q_{\theta}(z|x) = \log p_{\theta}(x|z) - \log \frac{q_{\theta}(z|x)}{p(z)} + \log \frac{q_{\theta}(z|x)}{p(z|x)} \quad (\approx)$$

T.k. $E[\text{const}] = \text{const}$, т.о. $E[\log p_{\theta}(x)] = \log p_{\theta}(x)$, even $z \sim q_{\theta}(z|x)$

$$\approx E_z \left[\log p_{\theta}(x|z) \right] - E_z \left[\log \frac{q_{\theta}(z|x)}{p(z)} \right] + E_z \left[\cancel{\log \frac{q_{\theta}(z|x)}{p(z|x)}} \right]$$

Точность
 расчета $q_{\theta}(z|x)$
 $\underbrace{KL(q_{\theta}(z|x) || p(z))}_{\geq 0}$
 $\underbrace{KL(q_{\theta}(z|x) || p(z|x))}_{\geq 0}$

$\log p_{\theta}(x) \geq - \sqrt{\dots}$
ELBO
 минимум граничка
 не-правдивоподобия

$z, E[z]$:

$$\frac{z_1 + z_2 + \dots + z_m}{m}$$

