Линейный классификатор

$X \subset \mathbb{R}^d$ – пр-во объектов

$Y = \{-1, +1\}$

$X^\ell = (x_i, y_i)_{i=1}^{\ell}$

$a(\vec{x}, \vec{w}) = sign(\langle \vec{w}, \vec{x} \rangle + b)$

$w \in \mathbb{R}^d$; $b \in \mathbb{R}$ – сдвиг

Обучение линейного классификатора заключается в поиске вектора весов, на котором достигается минимум

$$w = \arg\min_{w \in \mathbb{R}^d} Q(w, X^\ell)$$

Наиболее простыми являются число неверно классиф.

$$Q(w, X^\ell) = \sum_i^\ell [y_i(\langle \vec{w}, \vec{x}_i \rangle + b) < 0] \to \min_w$$

$$\sum_i^\ell [y_i(\langle \vec{w}, \vec{x}_i \rangle + b) < 0] \le \sum_i^\ell L(y_i(\langle \vec{w}, \vec{x}_i \rangle + b)) \to \min$$

$L(M) = \log(1 + e^{-M})$ – логистическая функ. потерь

$$\frac{\partial f}{\partial v} = \frac{d}{dt} f(x_{0,1} + t v_1, \dots, x_{0,d} + t v_d)\Big|_{t=0}$$

$$\frac{\partial f}{\partial v} = \sum_1^d \frac{\partial f}{\partial x_j} \frac{d}{dt}(x_{0j} + t v_j) = \sum_{j=1}^d \frac{\partial f}{\partial x_j} v_j = \langle \nabla f, v \rangle$$

$$\langle \nabla f, v \rangle = \|\nabla f\| \|v\| \cos\varphi = \|\nabla f\| \cos\varphi$$

$S(x_0) = \{x \in \mathbb{R}^d \mid f(x) = f(x_0)\}$

Разложим ф. в ряд Тейлора на этой линии уровня

$f(x_0 + \varepsilon) = f(x_0) + \langle \nabla f, \varepsilon \rangle + o(\|\varepsilon\|)$

, где $x_0 + \varepsilon \in S(x_0)$. Поскольку $f(x_0 + \varepsilon) = f(x_0)$

$\langle \nabla f, \varepsilon \rangle = o(\|\varepsilon\|)$

$\left\langle \nabla f, \frac{\varepsilon}{\|\varepsilon\|} \right\rangle = o(1)$

Устремим $\varepsilon \to 0 \Rightarrow \frac{\vec{\varepsilon}}{\|\varepsilon\|} \to$ касательная в точке x_0. В пределе получим, что градиент ортогонален этой касательной

Задача 1

$\nabla_x \langle \vec{a}, \vec{x} \rangle = a$

$\frac{\partial}{\partial x_j} \langle a, x \rangle = \frac{\partial}{\partial x_j} \sum_1^d a_k x_k = a_j$

Задача 2

$\nabla_x \|x\|_2^2 = 2x$

$\frac{\partial}{\partial x_j} \sum_1^d x_k^2 = 2x_j$

Задача 3

$\nabla_x \langle Ax, x \rangle = (A + A^T) x$

$\nabla_x \langle A\vec{x}, \vec{x} \rangle = ?$

$$\langle Ax, x \rangle = \sum_{j=1}^{n} (Ax)_j x_j = \sum_{j=1}^{n} \left(\sum_{k=1}^{n} a_{jk} x_k \right) x_j =$$

$$\sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} x_k x_j = \sum_{j=1}^{n} a_{jj} x_j^2 + \sum_{j \neq k} a_{jk} x_j x_k ;$$

$$\frac{\partial}{\partial x_i} \sum_{j=1}^{n} a_{jj} x_j^2 + \frac{\partial}{\partial x_i} \left(\sum_{j \neq k} a_{ij} x_i x_j + \sum_{j \neq i} a_{ji} x_i x_j \right) =$$

$$= 2a_{ii} x_i + \sum_{j \neq i} a_{ij} x_j + \sum_{j \neq i} a_{ji} x_j = \sum_{j=1}^{n} a_{ij} x_j + \sum_{j=1}^{n} a_{ji} x_j =$$

$$= (Ax)_i + (A^T x)_i = (A + A^T) x$$

$$\|A\vec{x}+\vec{b}\|^2 = \langle A\vec{x}+\vec{b}, A\vec{x}+\vec{b}\rangle =$$

$$\langle Ax, Ax\rangle + 2\langle Ax, b\rangle + \langle b, b\rangle =$$

$$= \langle A^TAx, x\rangle + 2\langle x, A^Tb\rangle + \langle b, b\rangle.$$

$$\nabla_x \|Ax+b\|_2^2 = \nabla_x \langle A^TAx, x\rangle + \nabla_x 2\langle x, A^Tb\rangle +$$

$$+ \nabla_x \langle b, b\rangle = (A^TA + x^TA)x + 2A^Tb =$$

$$= 2A^TAx + 2A^Tb = 2A^T(Ax+b)$$

Errors in Regression Tasks

1 Mean Squarred Error

$$\frac{1}{n}\sum_{i}^{n}(a-y)^2$$

2 Root Mean Squarred Error due to interpretation

$$\sqrt[2]{MSE}$$

3. RMSE = 80

if $a \in [0, 1]$, RMSE - bad

elif $a \in [10^4, 10^6]$, RMSE - good

3. Determination coefficient + R^2

$$R_2^2 = \frac{\sum_{i}^{n}(a-y(x_i))^2}{\sum_{i}^{n}(y(x_i)-\bar{y})^2} - ?$$

Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i}^{n} |a_i - y(x_i)|$$

Let is consider the following task

$$\frac{1}{n}\sum_{i}^{n} (a - y_i)^2 \to \min$$

$$a^*_{mse} = \frac{1}{n}\sum_{i=1} y_i$$ — it is optimal

$$\frac{1}{n}\sum_{i}^{n} |a - y_i| \to \min$$

$$a^*_{mas} = \text{median}\{y_i\}_1^n$$ ⇒ or mode?

MEAN SQUARED LOGARITHMIC ERROR

$$L(y,a) = (\log(a+1) - \log(y+1))^2$$

1) $a > 0$; $y > 0$
2) штраф за нарушение порядка, чем за значение
3) больше штраф от занижения значения чем завышение.

Mean Absolute Percentage Error

$$L(y,a) = \frac{|y-a|}{|y|}$$

Symmatric

$$L(y,a) = \frac{|y-a|}{\frac{(|y|+|a|)}{2}}$$

Learning of linear regression

Task: $\frac{1}{n}\sum_{i=1}^{n}(\langle w, x_i\rangle - y_i)^2 \to \min$

$$\frac{1}{n}\|Xw - y\|^2 \to \min_w$$

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

Problems: 1) $()^{-1}$ — $O(n^3)$

2) $X^T X$ — м.б. вырожденной

X - objects, $X \in R^n$

Y - answers, $Y \in R$ or $Y \subset R^n$

$y_i = y(x_i)$ $y: X \to Y$ unknown function;

$a(x) = f(x, \alpha)$ - model of dependency

$\alpha \in R^p$ - vector of model params

Least Squares method

$$Q(\alpha, X^\ell) = \sum_{i}^{\ell} w_i (f(x_i, \alpha) - y_i)^2 \to min$$

w_i - weight of importance of i object

$Q(\alpha^*, X^\ell)$ - residual sum of squares

Task of Least Square method is connected to Principle of maximum likelihood

The model of data with non-correlated gaussian noise

$$y(x_i) = f(x_i, \alpha) + \varepsilon_i \quad (1)$$

, where $\varepsilon_i \sim N(0, \sigma_i^2)$ $i = 1, \dots, \ell$

Method of Max Likelihood

$$(2)\ L(\varepsilon_1, \dots, \varepsilon_\ell \mid \alpha) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2} \varepsilon_i^2\right) \Rightarrow \max_\alpha$$

$$-\ln L(\varepsilon_1, \dots, \varepsilon_\ell \mid \alpha) = \text{const}(\alpha) + \frac{1}{2} \sum_1^\ell \frac{1}{\sigma_i^2} (f(x_i, \alpha) - y_i)^2 \to \min_\alpha ; \quad (3)$$

The application of least square method suppose that data has ① normal distribution and ② independent ③ dispersions are the same

Weights of objects $w_i = \sigma_i^{-2}$

Nadaraya - Watson

The approx. with the help of const $a(x) = \alpha$ at the vicinity $x \in X$

$$Q(\alpha; X^\ell) = \sum_{i=1}^{\ell} w_i(x)(\alpha - y_i)^2 \quad (4)$$

$w_i(x) = K\left(\frac{\rho(x, x_i)}{h}\right)$ - веса объектов x_i relatively to x;

K - kernel, non-increasing, limited, smooth

h - width of window

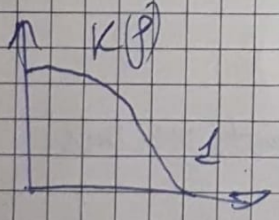
$$a_h(x, X^\ell) = \frac{\sum_{i=1}^{\ell} y_i w_i(x)}{\sum_{i=1}^{\ell} w_i(x)} = \frac{\sum_{i=1}^{\ell} y_i K\left(\frac{\rho(x_i, x)}{h}\right)}{\sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h}\right)}$$

$\sum w_i(x) = n$

Question 1: What is the solution of LSM if the solution is constant.

Average!

Example of kernel



Let the following conditions be satisfied:

1) Samples $X^l = (x_i, y_i)_{i=1}^{l}$ is simple & belongs to distribution $p(x, y)$;

2) Kernel $K(r)$ is limited;

$$\int_0^\infty K(r)\,dr < \infty, \quad \lim_{r \to \infty} rK(r) = 0;$$

3) Dependence $E(y|x)$ has no vertical asymptotes;

$$E(y^2|x) = \int_Y y^2 p(y|x)\,dy < \infty \text{ for any } x \in X$$

4) The sequence h_l decreases, but not too fast.

$$\lim_{l \to \infty} h_l = 0, \quad \lim_{l \to \infty} l h_l = \infty$$

Then there is convergence in probability

$$a_{h_l}(x, X^l) \xrightarrow{P} E(y|x) \text{ for any}$$

point $x \in X$, where

$E(y|x)$, $p(x)$, $P(x)$ is continuous

$p(x) > 0$

Kernel $K(r)$

- influence on the smoothes of $a_h(x)$ significantly affects smoothness
- slightly affects the quality of approx

Width of h

- significantly affects the quality of approx
- with an uneven grid $\{x_i\}$ there is variable width of window $h(x) = \rho(x, x^{(k+1)}), x^{(k+1)}$ - k-th neighbour of x
- Optimization of window width by sliding control

$$LOO(h, X^l) = \sum_{i=1}^{l} (a_h(x_i; X^l \setminus \{x_i\}) - y_i)^2 \to \min_h$$

How to handle outliers

Main idea:

Bigger the error $\varepsilon_i = |a_h(x_i; X^l \setminus \{x_i\}) - y_i|$ the more a precedent (x_i, y_i) is an outlier and less its weight should be.

Input: X^ℓ – learning set Algo

Output: coeficients γ_i $i = 1, \dots, \ell$

1 Init $\gamma_i = 1$ $i = 1 \dots \ell$

2 repeat

3 for all $i = 1 \dots \ell$

4 calc estimates of sliding control

$$a_i = a_h(x_i; X^\ell \setminus (x_i, y_i)) = \frac{\sum_{j=1, j \neq i}^{\ell} y_j \gamma_j K\left(\frac{f_j}{h}\right)}{\sum_{j=1, j \neq i}^{\ell} \gamma_j K\left(\frac{f_j}{h}\right)}$$

5 for all objects $i = 1, \dots, \ell$

$$\gamma_i = \tilde{K}(|a_i - y_i|);$$

6 $\|\gamma_i\| > \tilde{\gamma}$