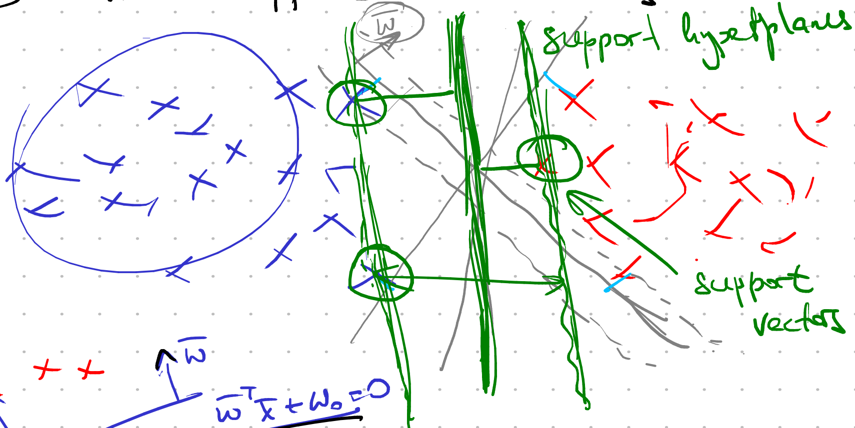
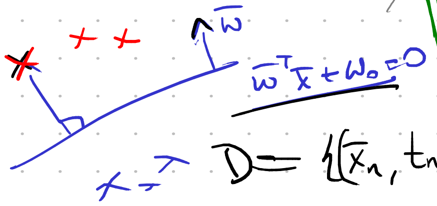
① SVM - support vector machines

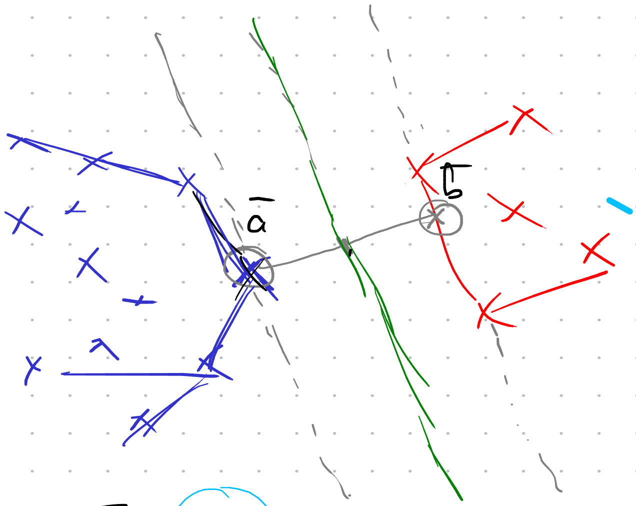


$$\boxed{\max_{\bar w} \min_{n} d(\bar x_n, A) = \max_{\bar w, w_0} \min_{n} \frac{t_n(\bar w^T \bar x_n + w_0)}{\|\bar w\|}}$$



$$D = \{(\bar x_n, t_n)\}_{n=1}^{N}, \quad t_n \in \{\pm 1\}$$

$$d(\bar x_n, A) = \frac{t_n(\bar w^T \bar x_n + w_0)}{\|\bar w\|}$$



$$\min \|\bar a - \bar b\|^2$$
$\bar a \in Conv(C_1)$
$\bar b \in Conv(C_2)$

$$\bar a = \sum_{n: t_n = 1} \alpha_n \bar x_n, \quad \alpha_n \geq 0, \quad \sum_{n: t_n = 1} \alpha_n = 1$$

$$\bar b = \sum_{n: t_n = -1} \alpha_n \bar x_n, \quad \alpha_n \geq 0, \quad \sum_{n: t_n = -1} \alpha_n = 1$$

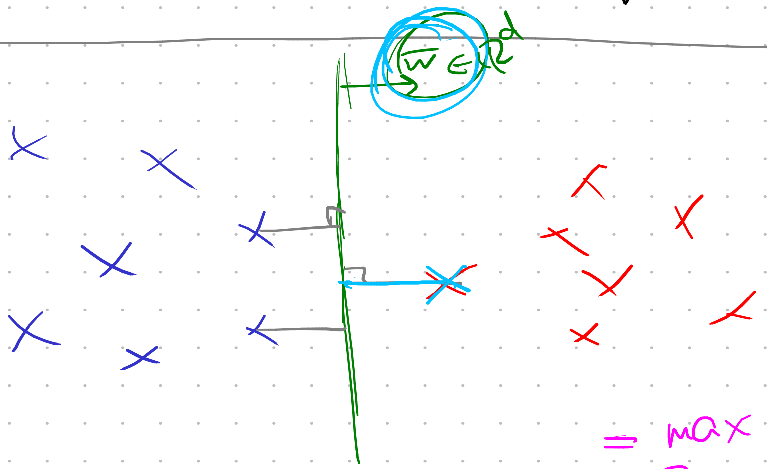
$$\min_{\bar\alpha} \left\| \sum_{t_n = 1} \alpha_n \bar x_n - \sum_{t_n = -1} \alpha_n \bar x_n \right\|^2 = \min_{\bar\alpha} \left\| \sum_{n} t_n \alpha_n \bar x_n \right\|^2$$

quadratic programming

при условии $\forall n \; \alpha_n \geq 0$

$$\sum_{t_n = 1} \alpha_n = \sum_{t_n = -1} \alpha_n = 1$$

N variables



$$\max_{\bar w, w_0} \min_{n} \frac{t_n(\bar w^T \bar x_n + w_0)}{\|\bar w\|} =$$

$$\forall n \; t_n(\bar w^T \bar x_n + w_0) \geq 0$$

$$= \max_{\bar w, w_0} \frac{1}{\|\bar w\|} \boxed{\min_{n} t_n(\bar w^T \bar x_n + w_0)}$$

$\|\bar w\| = 1$: $\cancel{\max_{\bar w, w_0} \min_{n}} \; t_n(\bar w^T \bar x_n + w_0)$

$\|\bar w\|^2 = 1$

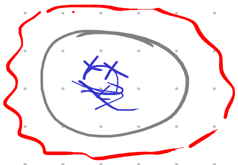
$\|\bar{w}\|: \quad \min_n t_n(\bar{w}^T \bar{x}_n + w_0) = 1$

$$\max_{\bar{w}, w_0} \frac{1}{\|\bar{w}\|} = \min_{\bar{w}} \|\bar{w}\|^2$$

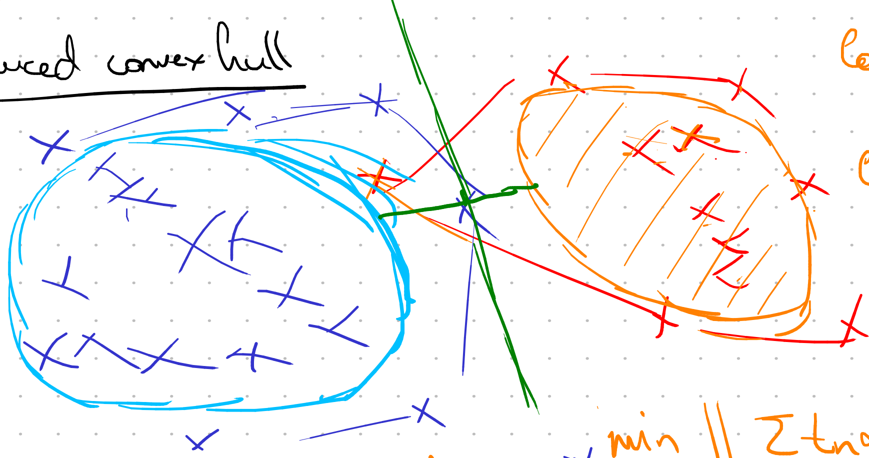
quadratic programming

$d+1$ variables

$$\forall n \quad t_n(\bar{w}^T \bar{x}_n + w_0) \geq 1$$



② Reduced convex hull

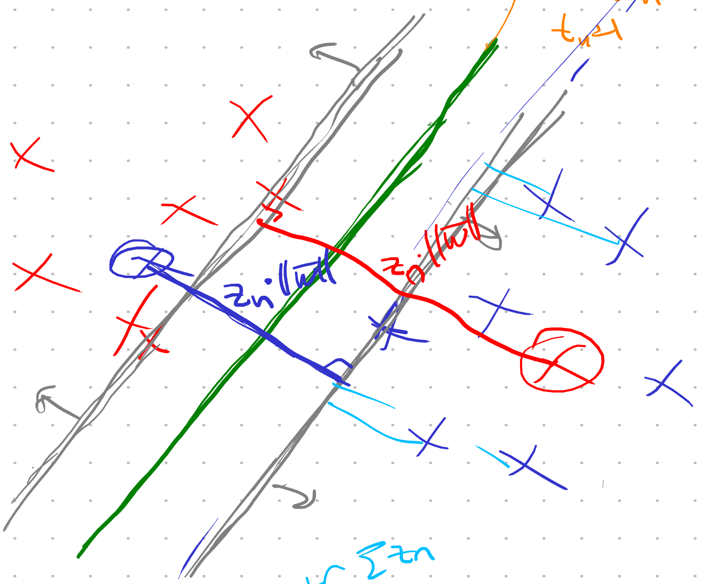


$$\mathrm{Conv}(X) = \left\{ \sum \alpha_n \bar{x}_n \;\middle|\; \begin{array}{l} \alpha_n \geq 0 \\ \sum_n \alpha_n = 1 \end{array} \right\}$$

$$\mathrm{Conv}_\gamma(X) = \left\{ \sum_n \alpha_n \bar{x}_n \;\middle|\; \begin{array}{l} \sum_n \alpha_n = 1 \\ 0 \leq \alpha_n \leq \gamma \end{array} \right\}$$

$$\min_{\alpha} \left\| \sum_n t_n \alpha_n \bar{x}_n \right\|^2,$$

$$\sum_{t_n = 1} \alpha_n = \sum_{t_n = -1} \alpha_n = 1, \quad \forall n \; 0 \leq \alpha_n \leq \gamma$$



slack

$$t_n(\bar{w}^T \bar{x}_n + w_0) + z_n \geq 1$$

$$\forall n \quad z_n \geq 0 \qquad N + d + 1 \text{ variables}$$

$$\min_{\bar{w}, w_0} \left\{ \|\bar{w}\|^2 + C \cdot \sum_n z_n \right\}$$

③

$$\min_{\bar{w}, w_0} \frac{1}{2}\|\bar{w}\|^2 + C \sum z_n \qquad \forall n \; t_n(\bar{w}^T \bar{x}_n + w_0) \geq 1$$

min

$$L(\bar{w}, w_0, \bar{\alpha}) = \frac{1}{2}\|\bar{w}\|^2 - \sum_n \alpha_n \left(t_n(\bar{w}^T \bar{x}_n + w_0) - 1 \right)$$

$\forall n \; \alpha_n \geq 0$

$$= \bar{w}^T \left(\sum_n \alpha_n t_n \bar{x}_n \right) + w_0 \cdot \sum_n \alpha_n t_n - \sum_n \alpha_n$$

$$\nabla_{\bar{w}} L = \bar{w} - \sum_n \alpha_n t_n \bar{x}_n = 0, \qquad \boxed{\bar{w} = \sum_n \alpha_n t_n \bar{x}_n}$$

$$\frac{\partial L}{\partial w_0} = -\sum_n \alpha_n t_n = 0 \qquad \boxed{\sum_n \alpha_n t_n = 0}$$

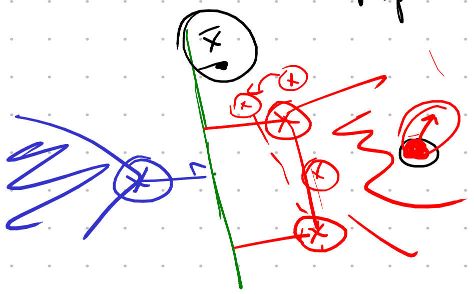
$$L(\bar{\alpha}) = \frac{1}{2}\left\| \sum_n \alpha_n t_n \bar{x}_n \right\|^2 - \left(\sum_n \alpha_n t_n \bar{x}_n\right)^T\left(\sum_n \alpha_n t_n \bar{x}_n\right) + \sum_n \alpha_n$$

$$L(\bar{\alpha}) = \sum_n \alpha_n - \frac{1}{2}\left(\sum_n \alpha_n t_n \bar{x}_n\right)^T\left(\sum_m \alpha_m t_m \bar{x}_m\right) =$$

N variables

$$= \sum_n \alpha_n - \frac{1}{2}\sum_n\sum_m \alpha_n\alpha_m t_n t_m \boxed{\bar{x}_n^T \bar{x}_m} \xrightarrow{\bar{\alpha}} \min$$

при усл. $\forall n \; \alpha_n \geq 0$, $\sum_n \alpha_n t_n = 0$



$$y(\bar{x}) = \bar{w}^T\bar{x} + w_0 =$$
$$= \sum_n \alpha_n t_n \boxed{(\bar{x}^T \bar{x}_n)} + w_0$$

Karush – Kuhn – Tucker:

$\forall n \quad \alpha_n \geq 0$ $\quad +\xi_n$

$\forall n \quad t_n y(\bar{x}_n) - 1 \geq 0$ $\quad +\xi_n$

$\forall n \quad \alpha_n \left(t_n y(\bar{x}_n) - 1\right) = 0$

$0 \qquad\qquad 0$

(4) Kernel trick

$$y(\bar{x}) = w_{11}x_1^2 + w_{12}x_1x_2 + w_{22}x_2^2 + $$
$$+ w_1x_1 + w_2x_2 + w_0$$

$$\underbrace{\bar{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}}_{\in \mathbb{R}^2} \overset{\varphi}{\longmapsto} \underbrace{\varphi(\bar{x}) = \begin{pmatrix} x_1^2 \\ x_1x_2 \\ x_2^2 \\ x_1 \\ x_2 \end{pmatrix}}_{\in \mathbb{R}^5}$$

$w_0 + \bar{w}^T\varphi(\bar{x}) = 0$

$$\bar{x} \in \mathbb{R}^d \longmapsto \varphi(\bar{x}) = \begin{pmatrix} x_1^2 \\ \vdots \\ x_d^2 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^{\frac{d(d+1)}{2}}$$

$$\longmapsto \begin{pmatrix} x_1^k \\ \vdots \\ x_d^k \\ \vdots \end{pmatrix} \in \mathbb{R}^{O(d^k)}$$

$$\begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2\end{pmatrix}^T \begin{pmatrix} y_1^2 \\ \sqrt{2}y_1y_2 \\ y_2^2 \end{pmatrix} = x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2 =$$

$$= (x_1y_1+x_2y_2)^2 = (\bar x^T \bar y)^2$$

$$\bar\varphi(\bar x)^T \bar\varphi(\bar y) = (\bar x^T\bar y)^2$$

$$k(\bar x,\bar y) = \bar\varphi(\bar x)^T\bar\varphi(\bar y), \quad k(\bar x,\bar y) - \text{kernel} \qquad \varphi: \mathbb{R}^d \to \mathbb{R}^D$$

$$L(\alpha) = \sum_n \alpha_n - \frac{1}{2} \sum_{n,m}\alpha_n\alpha_m t_n t_m k(\bar x_n,\bar x_m) \xrightarrow[\alpha]{} \min$$

$$\forall n\ \alpha_n \ge 0, \quad \sum_n \alpha_n t_n = 0$$

$$\bar w = \sum_n \alpha_n t_n \bar\varphi(\bar x_n) \in \mathbb{R}^D$$

$$y(\bar x) = \sum_n \alpha_n t_n k(\bar x, \bar x_n) + w_0$$

Теорема Мерсера:

$$\left.\begin{aligned} &1)\ k(\bar x,\bar y) = k(\bar y,\bar x) \\ &2)\ \iint k(\bar x,\bar y)\, g(\bar x) g(\bar y)\, d\bar x\, d\bar y \ge 0 \\ &\quad \forall g: \int g^2 < \infty \end{aligned}\right\} \Rightarrow k - \text{ядро}$$

$$\begin{pmatrix} x_1^2\\ \sqrt{2} x_1x_2 \\ x_2^2 \\ x_1 \\ x_2\end{pmatrix}^T \begin{pmatrix} \\ \bar\varphi(\bar y) \\ \\ \end{pmatrix} = (\bar x^T\bar y)^2 + \bar x^T \bar y$$

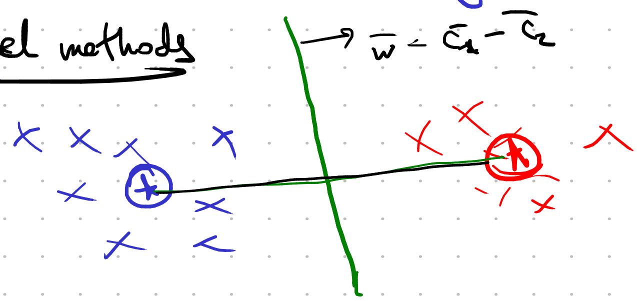
$$k(\bar x,\bar y) = e^{a\cdot\|\bar x-\bar y\|^2}$$

$$\begin{pmatrix} x_1^2 \\ \sqrt{2} \dots \\ x_d^2 \\ x_1 \\ \vdots \\ x_d\end{pmatrix}^T \begin{pmatrix} \\ \bar\varphi(\bar y) \\ \\ \end{pmatrix} = (\bar x^T\bar y)^2 + \bar x^T \bar y$$

$$k(\bar x,\bar y) = (\bar x^T\bar y)^k + (\bar x^T\bar y)^{k-1} + \dots + \bar x^T\bar y$$

$$k(\bar x,\bar y) = (\bar x^T\bar y + 1)^k - 1$$

5 Kernel methods



$\bar c_1 = \underbrace{\frac{1}{N_1}\sum_{n: t_n=1} \bar x_n}$ $\qquad$ $\bar c_2 = \underbrace{\frac{1}{N_2}\sum_{t_n=-1} \bar x_n}$ $\qquad$ $\frac{1}{N_1^2}\sum_{\substack{t_n=1\\ t_m=1}} \bar x_n^T \bar x_m$

$$y(\bar x) = \text{sign}\Big(\underbrace{\bar c_1^T \bar x - \bar c_2^T \bar x} - b\Big), \quad b = \frac{\|\bar c_1\|^2 - \|\bar c_2\|^2}{2}$$

$$= \text{sign}\left(\frac{1}{N_1}\sum_{t_n=1} \bar x_n^T \bar x - \frac{1}{N_2}\sum_{t_n=-1} \bar x_n^T \bar x - b\right)$$

$$y(\bar x) = \text{sign}\left(\frac{1}{N_1}\sum_{t_n=1} k(\bar x_n, \bar x) - \frac{1}{N_2}\sum_{t_n=-1} k(\bar x_n, \bar x) - b\right)$$

6 Equivalent kernel

$$p(y|\bar x, D) = \mathcal{N}(y \,|\, \bar\mu_N^T \bar x, \sigma_N^2(\bar x))$$

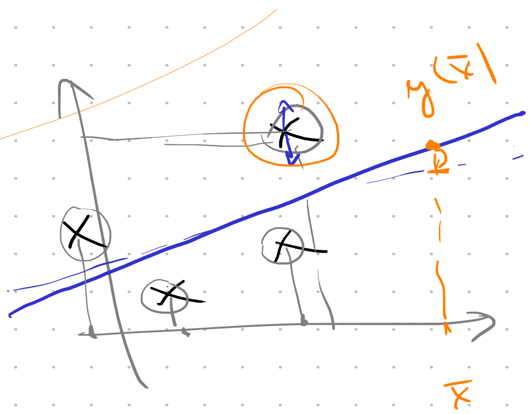
$$\bar\mu_N = \frac{1}{\sigma^2}\Sigma_N X^T \bar y, \quad \text{zge} \quad \Sigma_n^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2}\underbrace{X^T X}_{\bar x_n}$$

$$y(\bar x) = \bar\mu_N^T \bar x = \bar x^T\Big(\frac{1}{\sigma^2}\underbrace{\Sigma_N X^T \bar y}_{\in \mathbb{R}^N}\Big) = \frac{1}{\sigma^2}\bar x^T\left(\sum_{n=1}^{N} \Sigma_N \cdot \bar x_n y_n\right)$$

$\sum_n$

$$y(\bar x) = \sum_{n=1}^{N} \underbrace{\left(\frac{1}{\sigma^2}\bar x^T \Sigma_N \bar x_n\right)}_{k(\bar x, \bar x_n)} \cdot y_n$$

equivalent kernel



7 Relevance vector machines

$$p(y|\bar x, \bar w, \beta) = \mathcal{N}\left(y \,|\, \bar w^T \bar x, \sigma^2 = \frac{1}{\beta}\right)$$

$$\cancel{p(\bar w | \sigma_0^2) = \mathcal{N}}(\bar w | \bar 0, \underbrace{\sigma_0^2 \cdot I}_{\Sigma_0})$$

$$p(\bar w | \bar\alpha) = \prod_{i=1}^{d} \mathcal{N}(w_i | 0, \alpha_i^{-1}), \qquad \Sigma_0 = \begin{pmatrix} \alpha_1^{-1} & & 0 \\ & \alpha_2^{-1} & \\ & & \ddots \\ 0 & & & \alpha_d^{-1} \end{pmatrix}$$

$$D = (X, \bar y)$$

$$\begin{cases} \Sigma_N^{-1} = \Sigma_0^{-1} + \beta X^T X \\ \overline{\mu}_N = \beta \cdot \Sigma_N X^T \overline{y} \end{cases}$$

$$p(\overline{y} | X, \overline{\alpha}, \beta) = \int p(\overline{y} | X, \overline{w}, \beta)\, p(\overline{w} | \overline{\alpha})\, d\overline{w}$$

1) $\underbrace{\log p(\overline{y} / X, \alpha, \beta)}_{\overline{\alpha}, \beta \to \max} = \frac{N}{2} \log \beta + \frac{1}{2} \sum_i \log \alpha_i - \frac{1}{2} \overline{y}^T C^{-1} \overline{y} - \text{const}$

где $C = \frac{1}{\beta} I + X A^{-1} X^T$, где $A = \begin{pmatrix} \alpha_1 & & \\ & \ddots & \\ & & \alpha_d \end{pmatrix} = \Sigma_0^{-1}$

2) После $\frac{\partial \log p}{\partial \alpha_i}, \frac{\partial \log p}{\partial \beta}, \_\_\_ = 0$

$\gamma_i = 1 - \alpha_i \Sigma_{ii}$

$$\begin{cases} \alpha_i = \dfrac{\gamma_i}{\| \overline{\mu}_N \|^2} \\ \beta^{-1} = \dfrac{\| \overline{y} - X \overline{\mu}_N \|^2}{N - \sum_i \gamma_i} \end{cases}$$

3) К модель

$$\hat{y}(\overline{x}) = \sum_{n=1}^{N} w_n k(\overline{x}, \overline{x}_n) + b$$

$p(w_n) = \mathcal{N}(w_n | 0, \alpha_n^{-1})$ $\to 0$

$$p(\overline{y} | X, \overline{w}, \beta) = \mathcal{N}\left(y \,\Big|\, \sum_n w_n k(\overline{x}, \overline{x}_n) + b, \beta^{-1}\right)$$

Empirical Bayes

$$\begin{cases} \alpha_n = \gamma_n / \| \overline{\mu}_N \|^2 \longrightarrow \infty \\ \beta^{-1} = \_\_\_ \end{cases}$$