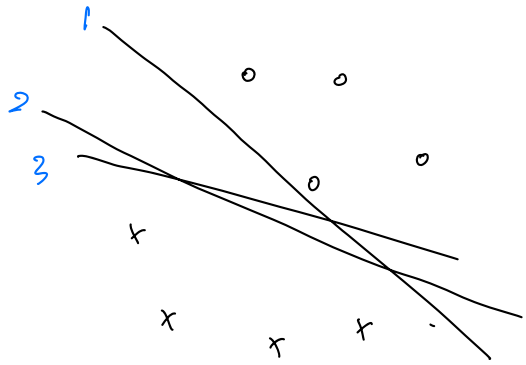
# SVM




perceptron

maximize correctness

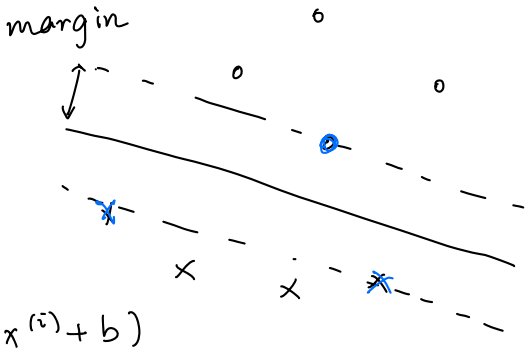
$$h_\theta(x) = \begin{cases} 1 & \theta^T x \geq 0 \\ 0 & \theta^T x < 0 \end{cases}$$

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$ ✓

⇓

SVM

maximize the margin




$$h_{w,b}(x) = \begin{cases} 1 & w^T x + b \geq 0 \\ -1 & w^T x + b < 0 \end{cases}$$

functional margin

$$\hat{\gamma} = \min_i \gamma^{(i)} = \min_i y^{(i)} (w^T x^{(i)} + b)$$

geometric margin

$$\gamma = \min_i \gamma^{(i)} = \min_i y^{(i)} \left(\frac{w^T}{\|w\|} x^{(i)} + \frac{b}{\|w\|}\right)$$

$$= \frac{1}{\|w\|} \hat{\gamma}$$

$\max \gamma$

* prime problem

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

$$\text{s.t.} \quad y^{(i)} (w^T x^{(i)} + b) \geq 1 \qquad i = 1, \ldots, m$$

dual problem

$$\max_\alpha \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \langle \alpha_i y^{(i)} x^{(i)}, \alpha_j y^{(j)} x^{(j)} \rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0 \qquad i = 1, \ldots, m$$

$$\sum_{i=1}^m \alpha_i y^{(i)} = 0$$

solution $w^* = \sum_{i=1}^m \alpha_i^* y^{(i)} x^{(i)}$ ✓

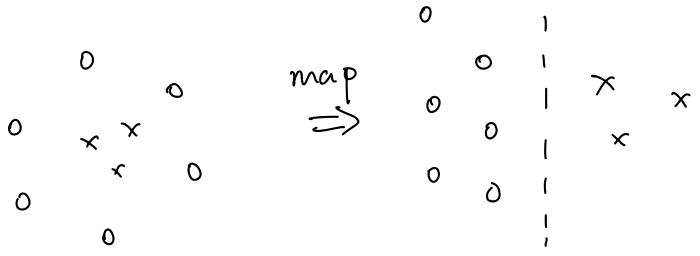
(after solving $\alpha^*$) $b^* = -\frac{1}{2} \left( \max_{i: y^{(i)} = -1} w^{*T} x^{(i)} + \min_{i: y^{(i)} = 1} w^{*T} x^{(i)} \right)$ ✓

prediction $\text{sign}\left[ w^{*T} z + b \right]$ for new sample $z$

kernel trick

$x \in \mathbb{R}^d \Rightarrow \phi(x) \in \mathbb{R}^D \quad D \gg d$

$K(x, x') \triangleq \phi^T(x)\,\phi(x') \in \mathbb{R}$




dual : $\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j K(x^{(i)}, x^{(j)})$

s.t $\alpha_i \geq 0 \quad i = 1, \cdots, m \quad \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$

prediction : $b^* = y^{(j)} - \sum_{i=1}^{m} \alpha_i^* y^{(i)} K(x^{(i)}, x^{(j)})$

$\text{sign}\left[ \sum_{i=1}^{m} \alpha_i^* y^{(i)} K(x^{(i)}, z) + b^* \right]$

Neural Network.

back propagation.

Bias & Variance

$\hat{h}(x) \qquad h(x)$

$\text{Bias}(\hat{h}) = E_D\left[\hat{h}(x) - h(x)\right]$

$\text{Var}(\hat{h}) = E_D\left[\hat{h}(x)^2\right] - E_D\left[\hat{h}(x)\right]^2$

$\sigma^2 = E_D\left[(h(x) - y)^2\right]$

D: all choices of training data sampled from $P_{XY}$

MSE Decomposition : $MSE = E_D\left[(\hat{h}(x) - y)^2\right]$

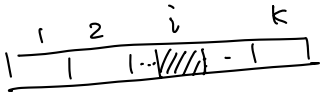
$= \text{Bias}(\hat{h}) + \text{Var}(\hat{h}) + \sigma^2$

Model Selection :

models + hyperparameters.

→ cross validation :

Hold-out : train, val, test split.

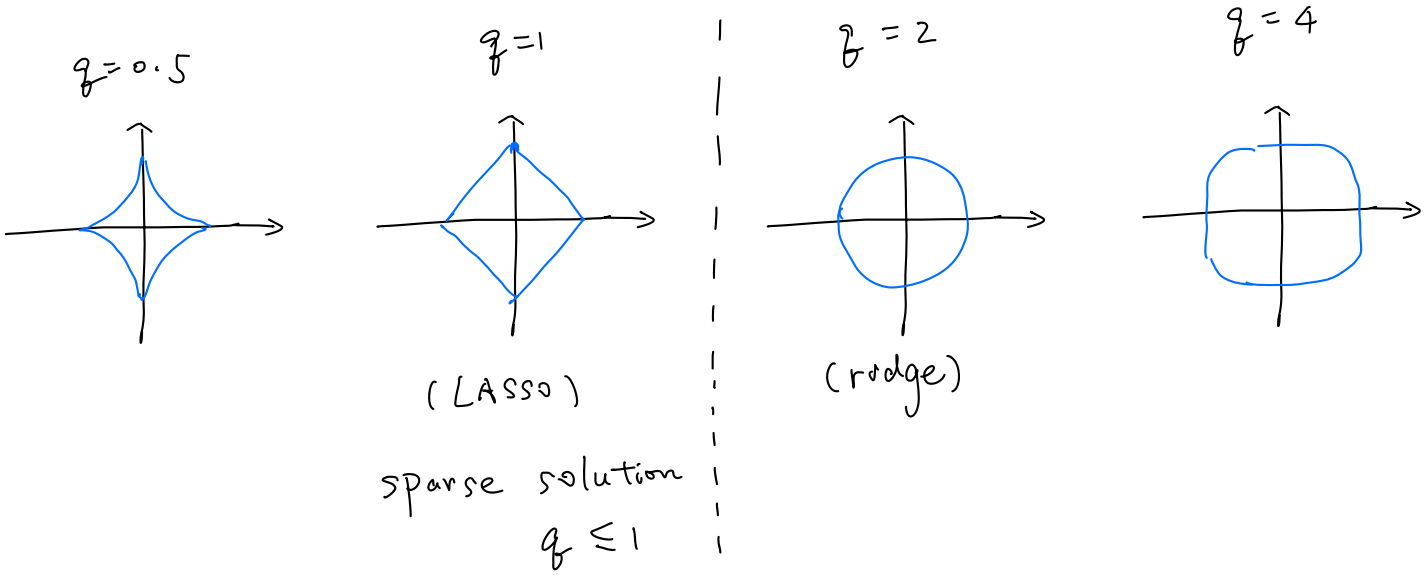
K-fold : 

take average.

Leave-One-Out : $k = m$

Regularization :

→ norm $L_q = \lambda \cdot \frac{1}{2} \|\theta\|_q^q + L(X, Y; \theta)$. ✓




→ Data Augmentation

→ Parameter Sharing

→ Drop out