> 整理自同学的笔记

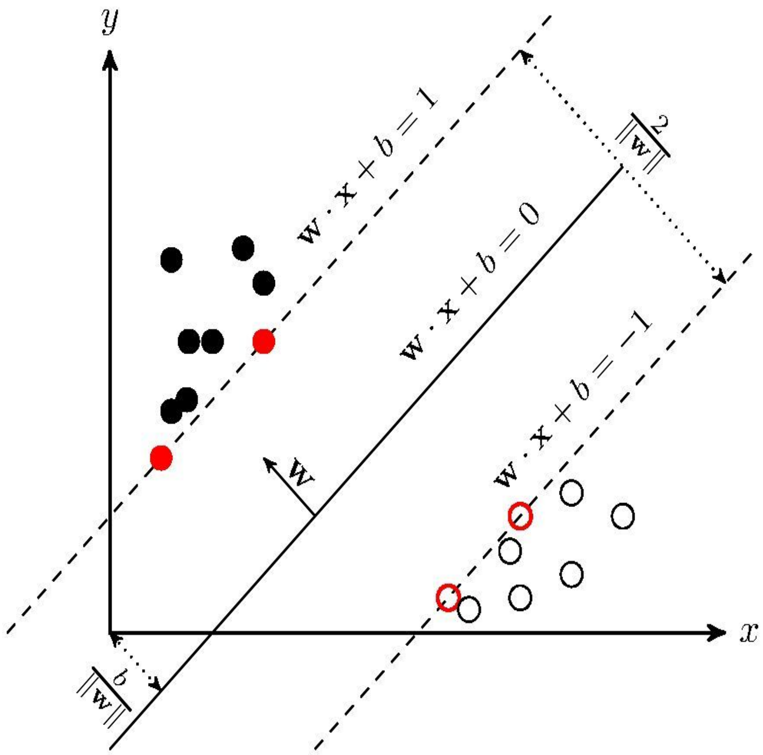
Support Vector Machine

Training data: $\{(\mathbf{x}_i, y)\}_{i=1}^n$, $y_i \in \mathcal{C} = \{-1, 1\}$.

Aim: $f(\mathbf{x}, \mathbf{w}, b) = b + \sum_{j=1}^d w_j x_j$, s.t. $y_i = \text{sign}(f(\mathbf{x}_i, w, b))$

SVM for linear separable data

Definition: A training sample is linear separate if there exists $(\hat{\mathbf{w}}, \hat{b})$, s.t. $y_i = \text{sign}(f(\mathbf{x}_i, \hat{\mathbf{w}}, \hat{b}))$, $\forall i \in [n] = \{1, 2, \cdots, n\}$, which is equivalent to $y_i f(\mathbf{x}_i, \hat{\mathbf{w}}, \hat{b}) > 0$, $\forall i \in [n]$.



点$\mathbf{x}_i$到线$\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$的距离$d(\mathbf{x}_i; \mathbf{w}, b) = \dfrac{y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)}{\|\mathbf{w}\|_2}$.

$$\max_{\mathbf{w}, b} \min_{\mathbf{x}_i \in D} margin(\mathbf{w}, b, D) = \max_{\mathbf{w}, b} \min_{\mathbf{x}_i \in D} d(\mathbf{x}_i) = \max_{\mathbf{w}, b} \min_{\mathbf{x}_i \in D} \frac{y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)}{\|\mathbf{w}\|_2}$$

Assumption 1: Training sample $D = \{(\mathbf{x}_i, y_i)\}$, is linear separable.

Definition:

The geometric margin $\gamma_f(\mathbf{z})$ of a linear classifier $f(\mathbf{x}, \mathbf{w}, b) = \langle \mathbf{w}, \mathbf{x} \rangle + b$ at a point $\mathbf{z}$ is its sigmoid Euclidean Distance to the hyperplane $\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$.

$$\gamma_f(\mathbf{z}) = \frac{y_i(\langle \mathbf{w}, \mathbf{z}_i \rangle + b)}{\|\mathbf{w}\|_2}$$

The geometric margin γ_f of a linear classifier f for sample $S = \{\mathbf{x}_1, \cdots, \mathbf{x_n}\}$ is the minimum margin over the points in the sample.

$$\gamma_f = \min_{i \in [n]} \gamma_f(\mathbf{x}_i)$$

Maximum Margin Classifier

$$\max_{\mathbf{w},b}\gamma_f=\max_{\mathbf{w},b}\left\{\frac{1}{\|\mathbf{w}\|}\min_{i\in[n]}y_i(\langle\mathbf{w},\mathbf{x}_i\rangle+b)\right\}$$

即

$$\begin{aligned}&\max_{\mathbf{w},b}\frac{1}{\|\mathbf{w}\|},\\ \text{s.t. }&\min_{i\in[n]}y_i(\langle\mathbf{w},\mathbf{x}_i\rangle+b)=1\\ &\Rightarrow y_i(\langle\mathbf{w},\mathbf{x}_i\rangle+b)\geq 1\\ &\Rightarrow\min_{\mathbf{w},b}\frac{1}{2}\|\mathbf{w}\|^2\end{aligned}$$

用反证法可证等号可以取到。

Definition: Given a SVM classifier $\langle\mathbf{w},\mathbf{x}_i\rangle+b=0$, the marginal hyperplanes are determined by $|\langle\mathbf{w},\mathbf{x}_i\rangle+b|=1$. The support vectors are the data instance on the marginal hyperplanes. （i.e. $\{\mathbf{x}_i:|\langle\mathbf{w},\mathbf{x}_i\rangle+b|=1,\mathbf{x}_i\in S\}$）

Not separable

minimize $\frac{1}{2}\|\mathbf{w}\|^2+C(training\ errors)$

minimize $\frac{1}{2}\|\mathbf{w}\|^2+C(distance\ of\ the\ error\ points\ and\ its\ correct\ position)$

SVM for non-separate cases:

$$\begin{aligned}&\min_{\mathbf{w},b,\epsilon}\frac{1}{2}\|\mathbf{w}\|+C\sum_{i=1}^{n}\epsilon_i,\\ \text{s.t. }&y_i(\langle\mathbf{w},\mathbf{x}_i\rangle+b)\geq 1-\epsilon_i,i\in[n]\\ &\epsilon_i\geq 0,i\in[n]\end{aligned}$$

Lagrange Duality

Consider the problem:

$$\begin{aligned}&\min f(\mathbf{x})\\ \text{s.t. }&g_i(\mathbf{x})\leq 0,i=1,\cdots,m\\ &h_i(\mathbf{x})=0,i=1,\cdots,p\\ &\mathbf{x}\in X\end{aligned}\tag{1}$$

f, g_i, h_i are all continously differentiable.

$$g(\mathbf{x})=\begin{bmatrix}g_1(\mathbf{x})\\ \vdots\\ g_m(\mathbf{x})\end{bmatrix},h(\mathbf{x})=\begin{bmatrix}h_1(\mathbf{x})\\ \vdots\\ h_p(\mathbf{x})\end{bmatrix}$$

Feasible Set: $D=\{\mathbf{x}:g(\mathbf{x})\leq 0,h(\mathbf{x})=0,\mathbf{x}\in X\}$.

Each $\mathbf{x}\in D$ is called a feasible solution. The optimal function value is $f^*=\inf_{\mathbf{x}\in D}f(\mathbf{x})$.

Transition from the domain to the image $S=\{(g(\mathbf{x}),h(\mathbf{x}),f(\mathbf{x})):\mathbf{x}\in X\}$ $(\dim=m+p+1)$

Definition 1: Associated with the primal problem, we define the Lagrangian L:
$\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^p\to\mathbb{R}$.

$$L(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x})$$

Definition 2: A vector $(\lambda^*, \mu^*) = (\lambda_1^*, \cdots, \lambda_m^*, \mu_1^*, \cdots, \mu_p^*)$ is said to be a geometric multiplier vector（or simply geometric multiplier）for the primal problem if:

$$\lambda_i^* \geq 0, i = 1, \cdots, m \text{ and } f^* = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \mu^*)$$

Lemma（Visualization Lemma）:

1. The hyperplane with normal $(\lambda, \mu, 1)$ that pass through $(g(\mathbf{x}), h(\mathbf{x}), f(\mathbf{x}))$ intercepts the vertical axis $\{(\mathbf{0}, z), z \in \mathbb{R}\}$ at the level $L(\mathbf{x}, \lambda, \mu)$.
2. Among all hyperplanes with normal $(\lambda, \mu, 1)$ that contains in their positive half space the set S, the highest attained level of interception of the vertical axis is $\inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \mu)$.

Proposition: Let (λ^*, μ^*) be a geometric multiplier. Then $\mathbf{x}^*$ is a global minimum of the primal problem iff $\mathbf{x}^* \in \arg\min_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \mu^*)$, $\lambda_i^* g_i(\mathbf{x}^*) = 0$, $i = 1, \cdots, m$（complementary slackness）.

Proof:

（$\Rightarrow$）

Suppose $\mathbf{x}^*$ is a global minimum. Then $\mathbf{x}^*$ must be feasible, and thus

$$f(\mathbf{x}^*) \geq L(\mathbf{x}^*, \lambda^*, \mu^*) \geq f^* = f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}^*) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x}^*)$$

The definition of f^* leads to $f^* = f(\mathbf{x}^*)$, which implies that

$$f(\mathbf{x}^*) = L(\mathbf{x}^*) = f^* = \inf_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \lambda^*, \mu^*)$$

$$\Rightarrow \mathbf{x}^* = \arg\min_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \mu^*) \text{ and } f(\mathbf{x}^*) = L(\mathbf{x}^*) = f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}^*) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x}^*)$$

$$\Rightarrow \lambda_i^* g_i(\mathbf{x}^*) = 0$$

（$\Leftarrow$）

$$f(\mathbf{x}^*) = L(\mathbf{x}^*, \lambda^*, \mu^*) \leq L(\mathbf{x}, \lambda^*, \mu^*) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}^*) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x}^*) \leq f(\mathbf{x})$$

Lagrange Duality:

Lagrange Dual Function: $q(\lambda, \mu) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \mu)$.

Lagrange Dual Problem: $\max q(\lambda, \mu)$, s.t. $\lambda \geq 0$.

Dual optimal value: $q^* = \sup_{\{(\lambda, \mu): \lambda \geq 0\}} q(\lambda, \mu)$

$\text{dom } q = \{(\lambda, \mu) : q(\lambda, \mu) > -\infty\}$

convex:

1. $\text{dom } q \cap \{(\lambda, \mu) : \lambda \geq 0\}$ is convex.
2. $-q$ is convex.（$f(\mathbf{x}) = \sup_{y \in \mathcal{Y}} l(\mathbf{x}, y)$, $l(\mathbf{x}, y)$ is convex $\Rightarrow$ $f(\mathbf{x})$ is convex）

Theorem（Week Duality Theorem）: $q^* \leq f^*$

Proof: $\forall (\lambda, \mu), q(\lambda, \mu) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \mu) \leq \inf_{\mathbf{x} \in D} L(\mathbf{x}, \lambda, \mu) \leq f^*$

Definition: Consider $f: X \to Y$

1. The value $f(x) \in Y$ that it assumes at element $x \in X$ is called the image of x.
2. The image of a set $A \subset X$ under the mapping f is $f(A) = \{y \in Y : \exists x \in A, \text{s.t. } f(x) = y\}$.
3. The preimage of as set $B \subset Y$ is $f^{-1}(B) := \{x \in X : f(x) \in B\}$

   eg: $f(X) = \det(A)$, $f(x^2) = 2x$.

Definition: A hyperplane H in $\mathbb{R}^{d+1}$ is specified by a linear equation involving a nonzero vector $(\mathbf{u}, u_0)$ （called the normal vector of H）, where $\mathbf{u} \in \mathbb{R}^d$ and $u_0 \in \mathbb{R}$ and by a constraint C as follows:

$$H = \{(\mathbf{w}, z) : \mathbf{w} \in \mathbb{R}^d, z \in \mathbb{R}, u_0 z + \langle \mathbf{u}, \mathbf{w} \rangle = C\}$$

Hyperplane defines two half-spaces: the positive half-space $H^+ = \{(\mathbf{w}, z) : \mathbf{w} \in \mathbb{R}^d, z \in \mathbb{R}, u_0 z + \langle \mathbf{u}, \mathbf{w} \rangle \geq C\}$ and the negative half-space $H^+ = \{(\mathbf{w}, z) : \mathbf{w} \in \mathbb{R}^d, z \in \mathbb{R}, u_0 z + \langle \mathbf{u}, \mathbf{w} \rangle \leq C\}$.

$$l(\mathbf{w}, z) = u_0 z + \langle \mathbf{u}, \mathbf{w} \rangle$$

Definition: Duality gap is $f^* - q^*$.

Proposition:

1. If there is no duality gap, the set of geometric multipliers is equal to the set of optimal dual solution.
2. If there is duality gap, the set of geometric multipliers is empty.

Optimality conditions:

A pair $\mathbf{x}^*$ and (λ^*, μ^*) is an optimal solution and geometric multiplier iff

$$\begin{gathered}
\mathbf{x}^* \in X, g(\mathbf{x}^*) \leq 0, h(\mathbf{x}^*) = 0. (\text{Primal Feasibility}) \\
\lambda^* \geq 0 (\text{Dual Feasibility}) \\
\mathbf{x}^* \in \arg\min_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \mu^*) (\text{Lagrangian Optimality }) \\
\lambda_i^* g_i^*(\mathbf{x}) = 0, i = 1, \cdots, m (\text{Complementary Slackness})
\end{gathered}$$

Saddle Point Theorem:

A pair $\mathbf{x}^*$ and (λ^*, μ^*) is an optimal solution and geometric multiplier iff $\mathbf{x}^* \in X$, $\lambda^* \geq 0$ and $(\mathbf{x}^*, \lambda^*, \mu^*)$ is a saddle point of the Lagrangian. i.e.

$$L((\mathbf{x}^*, \lambda, \mu) \leq L(\mathbf{x}^*, \lambda^*, \mu^*) \leq (\mathbf{x}, \lambda^*, \mu^*)), \forall \mathbf{x} \in X, \lambda \geq 0$$

Strong Duality Theorem:

Consider the primal problem. Suppose that f is convex , X is a polyhedral, i.e. $X = \{\mathbf{x} : \langle \mathbf{a}_i, \mathbf{x} \rangle \leq b, i = 1, \cdots, r\}$, g_i and h_i are linear and f^* is finite. Then there is no duality gap and there exists at least one geometric multiplier （primal and dual problems have optimal solutions）.

SVM & SVM Dual

SVM:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\epsilon_i$$
$$\text{s.t. } y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \epsilon_i, i = 1, \cdots, n$$
$$\epsilon_i \geq 0, i = 1, \cdots, n$$

$$L(\mathbf{w}, b, \epsilon, \alpha, u) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\epsilon_i + \sum_{i=1}^{n}\alpha_i(1 - \epsilon_i - y_i(\langle \mathbf{w}, x_i\rangle + b)) - \sum_{i=1}^{n}u_i\epsilon_i, \alpha \geq 0, u \geq 0$$

$$\begin{aligned} q(a, u) &= \inf_{\mathbf{w},b,\epsilon} L(b, \epsilon, \alpha, u) \\ &= \inf_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n}\alpha_i y_i\langle \mathbf{w}, \mathbf{x}_i\rangle \\ &+ \inf_b b\sum_{i=1}^{n}\alpha_i y_i \\ &+ \inf_\epsilon \sum_{i=1}^{n}(C - \alpha_i - u_i)\epsilon_i \end{aligned}$$

$$\nabla_{\mathbf{w}} L(\mathbf{w}, b, \epsilon, \alpha, u)|_{\mathbf{w}=\hat{\mathbf{w}}} = 0 \Rightarrow \hat{\mathbf{w}} - \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i - 0$$
$$\nabla_b L(\mathbf{w}, b, \epsilon, \alpha, u)|_{b=\hat{b}} = 0 \Rightarrow -\sum_{i=1}^{n}\alpha_i y_i = 0$$
$$\nabla_\epsilon L(\mathbf{w}, b, \epsilon, \alpha, u)|_{\epsilon=\hat{\epsilon}} = 0 \Rightarrow C - \alpha_i - u_i = 0$$

$$\max q(\alpha, u) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\langle \mathbf{x}_i, \mathbf{x}_j\rangle + \sum_{i=1}^{n}\alpha_i$$
$$\text{s.t. } \sum_{i=1}^{n}\alpha_i y_i = 0, \alpha_i \geq 0$$
$$C - \alpha_i - u_i = 0, u_i \geq 0$$

SVM Dual:

$$\max q(\alpha) \text{s.t. } \sum_{i=1}^{n}\alpha_i y_i = 0$$
$$\alpha_i \in [0, C], i = 1, \cdots, n$$

Proposition:

Let α^* be one of the dual optimal solutions.

$$\mathbf{w}^* = \sum_{i=1}^{n}\alpha_i^* y_i \mathbf{x}_i$$
$$\alpha_i(1 - \epsilon_i - y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b)) = 0, \forall i (\text{Complementary Slackness})$$

α_k^* is one of the entries of α^* and $\alpha_k^* \in (0, C)$, then:

$$(1 - \epsilon_i - y_i(\langle \mathbf{w}, x_i\rangle + b)) = 0$$
$$\alpha_k^* \in (0, C) \Rightarrow u_k^* \in (0, C) \Rightarrow \epsilon_k^* = 0$$
$$b^* = y_k - \langle \mathbf{w}^*, \mathbf{x_k}\rangle$$