Support Vector Machine

一. 线性可分支持向量机

Linear support vector machine in linear separable case

$$T = \{(x_1, y_1), (x_2, y_2) \cdots, (x_N, y_N)\}$$

$$x_i \in R^n \quad y_i \in \{-1, +1\}$$

间隔最大化 ⟶ 唯一解（区别于感知机）

$\Downarrow$

超平面: $wx + b = 0$

w ↓ 法向量　b ↓ 截距

分类决策函数: $f(x) = \text{sign}(wx + b)$

函数间隔: function margin

超平面 (w, b), 样本点 (x_i, y_i), 训练集 T

$$\hat{\gamma}_i = y_i |w \cdot x_i + b| \quad \text{（样本点的函数间距）}$$

$$\hat{\gamma} = \min_{i=1,2\cdots,N} \hat{\gamma}_i \quad \text{（训练集的函数间距）}$$

几何间隔: geometric margin

$$\gamma_i = y_i \left| \frac{w}{||w||} \cdot x_i + \frac{b}{||w||} \right|$$

（实例点到超平面的 signed distance）. 有向距离.

$$\gamma = \min_{i=1,2\cdots,N} \gamma_i$$

w与超平面正交:

$w^T x_1 + b = 0$ 　　(x_1, x_2 位于分离超平面上)

$w^T x_2 + b = 0$ 　　hyperplane.

$w^T (x_2 - x_1) = 0$

求几何间隔:

已知样本(x_i, y_i) 超平面 $w^T x + b = 0$，求 γ_i

作该样本在超平面上的投影

法向的单位向量: $\frac{w}{\|w\|}$

$x_i' = x_i - y_i \frac{w}{\|w\|} \cdot \gamma_i$

代入，$w^T x_i' + b = 0$

$\gamma_i = y_i \left(\frac{w}{\|w\|} x + \frac{b}{\|w\|} \right)$

$$\gamma_i = \frac{\hat{\gamma}_i}{\|w\|} \qquad \gamma = \frac{\hat{\gamma}}{\|w\|}$$

最大间隔分离 ⟺ 约束最优化: optim margin classifier

$$\max_{w,b} \ \gamma$$

$$\text{s.t.} \quad y_i\left(\frac{w}{\|w\|} \cdot x_i + \frac{b}{\|w\|}\right) \geq \gamma$$

$$\Updownarrow \qquad i = 1, 2 \cdots, N$$

$$\max_{w,b} \ \frac{\hat{\gamma}}{\|w\|}$$

$$\text{s.t.} \quad y_i(w \cdot x_i + b) \geq \hat{\gamma}$$

注意函数间隔的取值并不影响结果,

取 $\hat{\gamma} = 1$, 等价于 $\min_{w,b} \frac{1}{2}\|w\|^2$

$$\text{s.t.} \quad y_i(wx + b) \geq 1$$

(凸二次规划 convex quadratic programming.)

最优解的存在.唯一性证明:

存在性:

由于T线性可分, $y_i(wx+b)$必有可行解, 取最小值即为最优解.

注意 $w \neq 0$. (否则将不存在超平面)

唯一性：

① 设 $\|w_1\| = \|w_2\| = c$.

$$y_i(w_1 x_i + b) \geq 1$$
$$y_i(w_2 \cdot x_i + b) \geq 1$$
$$\rightarrow \quad y_i\left(\left(\frac{w_1+w_2}{2}\right) x_i + b\right) \geq 1$$

故 $\frac{w_1+w_2}{2}$ 也为可行解. 记为 w, $\|w\| \geq c$

（因为不一定为最优解）

$$c \leq \|w\| \leq \frac{\|w_1\|+\|w_2\|}{2} = c \quad \text{(三角不等式)}.$$

故 $\|w\| = \frac{\|w_1\|+\|w_2\|}{2} = \frac{\|w_1+w_2\|}{2}$

则 $w_1 = \lambda w_2$, $|\lambda| = 1$

若 $\lambda = -1$： $w = w_1 + w_2 = 0$, w 不是可行解.

故 $\lambda = 1$， $w_1 = w_2$. 设此时两个截距分别为 b_1, b_2.

② 设 x_1', x_2' 为集合 $\{x_i \mid y_i = +1\}$ 对应 (w, b_1), (w, b_2) 使不等式等号成立的点， x_1'', x_2'' 为 $\{x_i \mid y_i = -1\}$ … 成立的点.

$$w \cdot x_1' + b_1 = 1 \quad , \quad w \cdot x_2' + b_2 = 1$$
$$w \cdot x_1'' + b_1 = -1 \quad , \quad w \cdot x_2'' + b_2 = -1$$

$$b_1 = -\frac{w \cdot (x_1' + x_1'')}{2} \qquad b_2 = -\frac{w \cdot (x_2' + x_2'')}{2}$$

$$b_1 - b_2 = -\frac{1}{2}\left[w \cdot (x_1' + x_1'') - w \cdot (x_1' + x_2'')\right]$$

又 $w \cdot x_1' + b_2 \geq 1 = w \cdot x_2' + b_2$

$w \cdot x_2' + b_1 \geq 1 = w \cdot x_1' + b_1$

$$w \cdot x_1' = w \cdot x_2'$$

同理 $w \cdot x_1'' = w \cdot x_2''$

$b_1 - b_2 = 0$ $\quad b_1 = b_2$.

拉格朗日对偶性 Lagrange duality

$$\min f_w(w) \quad s.t. \quad h_i(w)=0 \quad i=1,2,\cdots,p$$

定义拉格朗日函数 Lagrangian:

β: Lagrange multipliers

$$\mathcal{L}(w,\beta) = f(w) + \sum_{i=1}^{p} \beta_i h_i(w)$$

$$\frac{\partial \mathcal{L}}{\partial w}=0, \quad \frac{\partial \mathcal{L}}{\partial \beta_i}=0$$

原始优化问题: (primal optimization problem)

$$\min_w f(w) \quad s.t. \quad \begin{array}{ll} g_i(w)\le 0 & i=1,2,\cdots,k \\ h_i(w)=0 & i=1,2,\cdots,p \end{array}$$

$$\mathcal{L}(w,\alpha,\beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{p} \beta_i h_i(w)$$

$(\alpha_i \ge 0)$

$$\theta_p(w) = \max_{\alpha,\beta} \mathcal{L}(w,\alpha,\beta) = \begin{cases} f(w) & g_i(w)\le 0 \quad i=1,2,\cdots,k \;\&\; h_i(w)=0 \quad i=1,2,\cdots,p \\ +\infty & \text{otherwise} \end{cases}$$

$$\Downarrow$$

$$\min_w f(w) \Leftrightarrow \min_w \theta_p(w) \Leftrightarrow \min_w \max_{\alpha,\beta} \mathcal{L}(w,\alpha,\beta)$$

Note: $p^* = \min \theta_p(w)$, value of primal problem.

$$\theta_D(\alpha,\beta) = \min_w \mathcal{L}(w,\alpha,\beta)$$

$$\max_{\alpha,\beta} \theta_D(\alpha,\beta) = \max_{\alpha,\beta} \min_w \mathcal{L}(w,\alpha,\beta)$$

Note: $d^* = \max_{\alpha,\beta} \theta_D(\alpha,\beta)$ value of dual problem.

$$d^* = \max_{\alpha,\beta} \min_w \mathcal{L}(w,\alpha,\beta) \le \min_w \max_{\alpha,\beta} \mathcal{L}(w,\alpha,\beta) = p^*$$

When $d^* = p^*$:

① f & g_i's are convex

② h_i's are affine (仿射函数: $\exists\, a_i, b_i$, 使 $h_i(w) = a_i^T w + b_i$)

③ $\exists\, w$, s.t. $g_i(w) < 0$

$\underbrace{\exists\, w^*, \alpha^*, \beta^*}_{\downarrow}$, s.t. $p^* = \mathcal{L}(w^*, \alpha^*, b^*) = d^*$

KKT conditions:

① $$\frac{\partial \mathcal{L}(w^*, \alpha^*, \beta^*)}{\partial w_i} = 0$$

② $$\frac{\partial \mathcal{L}(w^*, \alpha^*, \beta^*)}{\partial \beta_i} = 0$$

③ $\alpha_i^* \cdot g_i(w^*) = 0$ dual complementarity condition
对偶 互补

④ $g_i(w^*) \le 0$

⑤ $\alpha^* \ge 0$

支持向量：support vector

使不等式等号成立的样本点

正例：H_1：$w\cdot x+b=1$

负例：H_2：$w\cdot x+b=-1$

H_1、H_2之间形成长带，超平面位于其中央并与其平行

H_1 H_2之间的距离为间隔 margin $=\frac{2}{\|w\|}$

H_1 H_2称为间隔边界.

在间隔边界外移动/去掉实例点，解不变.

对偶算法 dual algorithm.

（对应上式，令 $g_i(w)=-y_i(w^Tx_i+b)+1\le 0$）

构造 Lagrange 函数

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{N}\alpha_i y_i(w\cdot x_i+b)+\sum_{i=1}^{N}\alpha_i$$

α：$(\alpha_1,\alpha_2,\dots\alpha_N)^T$ 为 Lagrange multiplier vector ($\alpha_i\ge 0$)

根据 Lagrange 对偶性，原问题等价于 $\max\limits_{\alpha}\min\limits_{w,b} L(w,b,\alpha)$

① 求 $\min\limits_{w,b} L(w,b,\alpha)$

令 $\nabla_w L=0$ $\nabla_b L=0$

解得 $\boxed{w=\sum_{i=1}^{N}\alpha_i y_i x_i \qquad \sum_{i=1}^{N}\alpha_i y_i=0}$

代入上式

$$\min_{w,b} L(w,b,\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N}\left(\alpha_i y_i x_i \sum_{j=1}^{N}\alpha_j y_j x_j\right) + \sum_{i=1}^{N}\alpha_i$$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i x_j + \sum_{i=1}^{N}\alpha_i$$

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

即求 $\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i x_j - \sum_{i=1}^{N}\alpha_i$

s.t. $\sum_{i=1}^{N}\alpha_i y_i = 0$

故超平面方程为: $\sum_{i=1}^{N}\alpha_i y_i x_i \cdot x + b = 0$

b: $\exists\, \alpha_j > 0$ (∃支持向量)

(否则若 $\alpha = 0$, $w=0$, 不存在解)

$y_j(w \cdot x_j + b) - 1 = 0$.

又 $y_j^2 = 1$

$b = y_j - \sum_{i=1}^{N}\alpha_i y_i x_i x_j$

KKT条件:

$\nabla_w L(w,b,\alpha) = 0$.

$\nabla_b L(w,b,\alpha) = 0$

$\alpha_i(y_i(w \cdot x_i + b) - 1) = 0$

$y_i(w \cdot x_i + b) - 1 \geq 0$

$\alpha_i \geq 0$.

$(i = 1, 2, \ldots, N)$

对 w, b 有贡献的仅有样本点中对应 $\alpha_i > 0$ 的 (x_i, y_i)

将 $\alpha_i > 0$ 的实例点 $x_i \in R^n$ 称为支持向量

对于支持向量：

由KKT条件： $\alpha_i (y_i (w \cdot x_i + b) - 1) = 0$

又 $\alpha_i > 0$，故 $y_i (w \cdot x_i + b) - 1 = 0$ ， $y_i (w \cdot x_i + b) = 1$

样本点处在间隔边界上.

总结：对于线性可分的数据集完美.

predict： $w^T x + b = \left(\sum_{i=1}^{N} \alpha_i y_i x_i \right)^T x + b.$

$= \sum_{i=1}^{N} \alpha_i y_i \langle x_i, x \rangle + b.$

只需计算支持向量与x的内积即可.

二. 线性支持向量机

存在特异点 outlier 不满足 $y_i (w^T x_i + b) \geq 1$

引入松弛变量 ε_i

$y_i (w^T x_i + b) \geq 1 - \varepsilon_i$

目标函数：

$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \varepsilon_i$ $C > 0$ C：惩罚参数.

$\varepsilon_i > 0$

Primal Problem：

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \varepsilon_i$$

$$\text{s.t.} \quad y_i (w^T x_i + b) \geq 1 - \varepsilon_i \qquad i = 1, 2, \dots, N.$$

Lagrange:

$$\mathcal{L}(w, b, \varepsilon, \alpha, \mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\varepsilon_i - \sum_{i=1}^{N}\alpha_i\left(y_i(w^T \cdot x_i + b) + \varepsilon_i - 1\right) - \sum_{i=1}^{N}\mu_i\varepsilon_i$$

$$\alpha_i,\ \mu_i \geq 0.$$

$$\frac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N}\alpha_i x_i y_i$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \rightarrow \sum_{i=1}^{N}\alpha_i y_i = 0$$

$$\frac{\partial \mathcal{L}}{\partial \varepsilon_i} = 0 \rightarrow C - \alpha_i - \mu_i = 0.$$

$$\min_{w, b, \varepsilon}\mathcal{L}(w, b, \varepsilon, \alpha, \mu) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i x_j + \sum_{i=1}^{N}\alpha_i$$

$$\Updownarrow$$

$$\max_{\alpha}\mathcal{L}(w, b, \varepsilon, \alpha, \mu) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i x_j) + \sum_{i=1}^{N}\alpha_i$$

s.t. $\sum_{i=1}^{N}\alpha_i y_i = 0$

$$\left.\begin{array}{l} C - \alpha_i - \mu_i = 0 \\ \alpha_i \geq 0 \\ \mu_i \geq 0 \end{array}\right\} \rightarrow 0 \leq \alpha_i \leq C.$$

$$\Downarrow$$

$$\max_{\alpha}\mathcal{L}(w, b, \varepsilon, \alpha, \mu) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i$$

s.t. $\sum_{i=1}^{N}\alpha_i y_i = 0$ Dual Problem.

$0 \leq \alpha_i \leq C$

此时的 KKT条件:

$$\frac{\partial L}{\partial w} = w^* - \sum_{i=1}^{N} \alpha_i^* x_i y_i = 0$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \varepsilon_i} = C - \alpha_i^* - \mu_i^* = 0$$

极值要求.

$$\alpha_i^* \left(y_i (w^{*T} \cdot x_i + b) - 1 + \varepsilon_i^* \right) = 0$$

$$\mu_i^* \varepsilon_i^* = 0$$

对偶互补

$$y_i (w^{*T} \cdot x_i + b) - 1 + \varepsilon_i \geq 0$$

$$\varepsilon_i^* \geq 0$$

原始条件

$$\alpha_i^* \geq 0$$

$$\mu_i^* \geq 0$$

因子.

求解： $w^* = \sum_{i=1}^{N} \alpha_i^* x_i y_i$

$b^* = y_j - \sum_{i=1}^{N} \alpha_j^* x_i y_i \, x_j$ （若∃ α_j, $0 < \alpha_j < C$）

由 KKT，

$\alpha_i^* (y_i (w^* \cdot x_i + b) - 1 + \varepsilon_i) = 0$

$\mu_i^* \varepsilon_i^* = 0$

若 $0 < \alpha_j < C$，则 $\mu_j > 0$，$\varepsilon_j^* = 0$

又 $y_j (w^* \cdot x_j + b) - 1 + \varepsilon_j^* = 0$

$b = y_j - w^* x_j = y_j - \sum_{i=1}^{N} \alpha_i^* x_i y_i \cdot x_j$

总结： ① 选择 $C>0$，构造求解凸二次规划问题：

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$0 \le \alpha_i \le C \quad i = 1, 2, \cdots, N$$

② 选择1个 α_j^* 满足 $0 < \alpha_j^* < C$

计算 $b = y_j - \sum_{i=1}^{N} \alpha_j^* x_i y_i x_j$

$w^* = \sum_{i=1}^{N} \alpha_i^* x_i y_i$

③ 超平面： $w^{*T} x + b = 0$.

支持向量：$\alpha_i^* > 0$

则有：$y_i(w^* \cdot x_i + b) - 1 + \varepsilon_i = 0$

若 $\alpha_i^* < C$，$\mu_i^* > 0$，$\varepsilon_i^* = 0$ 超平面上

若 $\alpha_i^* = C$，$0 < \varepsilon_i < 1$ 分类正确 间隔边界与超平面之间

若 $\alpha_i^* = C$，$\varepsilon_i > 1$ 分类错误 间隔边界与超平面之间

合页损失函数 hinge loss function

$$\min \sum_{i=1}^{N} [1 - y_i(w^T x_i + b)]_+ + \lambda \|w\|^2$$

$$[1 - y_i(w^T x_i + b)]_+ = \begin{cases} 1 - y_i(w^T x_i + b) & y_i(w^T x_i + b) < 1 \\ 0 & y_i(w^T x_i + b) \geq 1 \end{cases}$$

$$\min_{w,b,\varepsilon} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \varepsilon_i \quad \text{s.t.} \quad y_i(w^T x_i + b) \geq 1 - \varepsilon_i$$

$$\varepsilon_i \geq 0$$

令 $[1 - y_i(w^T x_i + b)]_+ = \varepsilon_i \geq 0$

则 $y_i(w^T x_i + b) \geq 1 - \varepsilon_i$

取 $\lambda = \frac{1}{2C}$，$HLF = \min \frac{1}{C}\left[\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \varepsilon_i\right]$

与原式等价.

Kernel trick

X：输入空间（欧氏空间 R^n 的子集或离散集合）

H：特征空间（希尔伯特空间）

若 $\exists\ \phi(x): X \to H$，s.t. $\forall\ x, z \in X$，$K(x,z) = \phi(x)^T\phi(z)$

应用：

对偶问题的目标函数：$W(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\ y_i y_j\ K(x_i, x_j) - \sum_{i=1}^{N}\alpha_i$

分类决策：$f(x) = \mathrm{sign}\left(\sum_{i=1}^{N}\alpha_i^* y_i\ K(x_i, x) + b^*\right)$

正定核：$K = [K(x_i, x_j)]_{m\times m}$，要求该 Gram 矩阵半正定

必要性：对 $\forall\ x = (c_1, c_2, \cdots, c_m) \in R^m$

$$xKx^T = \sum_{i,j=1}^{m} c_i c_j\ K(x_i, x_j)$$

$$= \sum_{i,j=1}^{m} c_i c_j\ \phi(x_i)\cdot\phi(x_j)$$

$$= \sum_{i=1}^{m} c_i\,\phi(x_i)\cdot\sum_{j=1}^{m} c_j\,\phi(x_j)$$

$$= \left[\sum_{i=1}^{m} c_i\,\phi(x_i)\right]^2 \geq 0$$

充分性：构造 $\phi: X \to H$：$x \to K(\cdot, x)$

$$\langle\phi(x), \phi(z)\rangle_H = \langle K(x,\cdot), K(z,\cdot)\rangle_H = K(x,z)$$

因此 K 为有效核函数.

常用核函数：

多次式 polynomial

$$K(x,z) = (x^T z + 1)^p$$

$$f(x) = \mathrm{sign}\left(\sum_{i=1}^{N} a_i^* y_i (x_i^T x + 1)^p + b^*\right)$$

高斯 Gausian

$$K(x,z) = \exp\left(-\frac{\|x-z\|^2}{2\delta^2}\right)$$

$$f(x) = \mathrm{sign}\left(\sum_{i=1}^{N} a_i^* y_i \exp\left(-\frac{\|x-z\|^2}{2\delta^2}\right) + b^*\right)$$

字符串 string

子串长度： $i_{|u|} - i_1 + 1$ 例：abcde 中 ae 为 5

$$K_n(s,t) = \sum_{u\in\Sigma^n} [\phi_n(s)]_u [\phi_n(t)]_u$$ 余弦相似度 cosine similarity.

$$= \sum\sum_{(i,j):\ s(i)=t(j)=u} \lambda^{p(i)} \lambda^{p(j)}$$

$$[\phi_n(s)]_u = \sum_{i:s(i)=u} \lambda^{p(i)} \quad (0 < \lambda \le 1)$$