Support Vector Machine (支持向量机)

1. 线性可分的支持向量机

线性可分问题：可以在特征空间中找到一个分离的超平面 $w^Tx+b=0$

将特征空间划分为正例负例，分类决策函数 $f(x)=\mathrm{sign}(w^Tx+b)$

1.1 函数间隔与几何间隔

复习：空间中任意一个点 x_i 到平面 $w^Tx_i+b=0$ 的距离为 $\frac{|w^Tx_i+b|}{\|w\|}$

对所有点 $\|w\|$ 与 i 无关，$|w^Tx_i+b|$ 表示点到平面的相对距离

Def（函数间隔）对于给定超平面 (w,b) 和训练数据

定义超平面关于样本点 (x_i,y_i) 的函数间隔 $\hat{r}_i=y_i(w^Tx_i+b)$

① 代表是否分类正确（正负）

② 分类的确信度（比较相对距离大小）

超平面关于训练集的函数间隔 $\hat{r}=\min\limits_i \hat{r}_i$

如果将 w 与 b 进行同比例变换，超平面不变，但函数间隔变化

对 w 加约束 $\|w\|=1$，此时对应几何间隔 $r_i=y_i\left(\frac{w^Tx_i+b}{\|w\|}\right)$

与样本数据集 $r=\min\limits_i r_i$

1.2 间隔最大化

对训练数据找到几何间隔最大的超平面，可以转化成以下带约束的最优化问题

$$\max_{w,b} \quad r$$

$$\text{s.t.} \quad y_i\left(\frac{w^Tx_i+b}{\|w\|}\right)\geq r \qquad i=1,2,\cdots,N$$

考虑几何间隔与函数间隔的等价性

$$\max_{w,b} \quad \frac{\hat{r}}{\|w\|}$$

$$\text{s.t.} \quad y_i\left(\frac{w^Tx_i+b}{\|w\|}\right)\geq \frac{\hat{r}}{\|w\|}$$

函数间隔的取值不影响以上问题的解

$$\max_{w,b} \frac{\hat{r}}{\|w\|} \quad \Longleftrightarrow \quad \min_{w,b} \|w\|_2^2$$

$$s.t.\ y_i(w^T x_i + b) \geq \hat{r} \qquad s.t.\ y_i(w^T x + b) \geq 1$$

(转化为二次规划问题)

得 w^*, b^* $\quad f(x) = \text{sign}(w^* x + b^*)$

支持向量 (support vector) 在超平面 $w^T x + b = 1$ 或 -1 上的点

2. 学习的对偶算法

2.1 拉格朗日的对偶性

1. 原问题

考虑一个约束最优化问题

$$\min_{x \in R^n} f(x)$$

$$s.t.\ c_i(x) \leq 0 \quad i = 1, 2, \cdots, k$$

$$h_j(x) = 0 \quad j = 1, 2, \cdots, l$$

引入拉格朗日函数

$$L(x, \alpha, \beta) = f(x) + \sum_{i=1}^{k} \alpha_i c_i(x) + \sum_{j=1}^{l} \beta_j h_j(x) \quad \text{要求 } \alpha_i \geq 0$$

考虑 x 的函数

$$\theta_p(x) = \max_{\alpha_i = 0, \beta} L(x, \alpha, \beta)$$

① if x 违反了约束条件, $\begin{cases} c_i(x) > 0 \\ h_j(x) \neq 0 \end{cases}$, 则 $\theta_p(x) = +\infty$

② if x 没有违反约束, $\theta_p(x) = f(x)$

因此考虑 $\min\limits_{\alpha} \theta_p(x) = \min\limits_{x} \max\limits_{\alpha_i > 0, \beta} L(x, \alpha, \beta)$ 与原问题等价

2. 对偶问题

$$\max_{\alpha \geq 0\, \beta} \min_{x} L(x, \alpha, \beta)$$

3. 原问题和对偶问题之间的关系

(1) 若原问题与对偶问题都有最优解，则对偶问题的最优解 ≤ 原问题的最优解

$$\max_{\alpha,\beta} \min_{x} L(x,\alpha,\beta) \leq \min_{x} \max_{\alpha,\beta} L(x,\alpha,\beta)$$

对于 $\forall \alpha,\beta$，$\theta_d(\alpha,\beta) = \min_{x} L(x,\alpha,\beta)$

$$\leq L(x,\alpha,\beta)$$

$$\leq \max_{\alpha,\beta} L(x,\alpha,\beta) = \theta_p(x)$$

则 $\max_{\alpha,\beta} \theta_d(\alpha,\beta) \leq \min_{x} \theta_p(x)$

(2) 一定条件下，原问题的解 = 对偶问题的解

假设 ① $f(x)$ 与 $c_i(x)$ 是凸函数

② $h_j(x)$ 是仿射函数 $(Ax+b)$

③ $c_i(x)<0$ 严格可行

KKT条件：$\nabla L(x^*,\alpha^*,\beta^*)=0$

$$\begin{cases} \alpha_i^* c_i(x^*)=0 \\ c_i(x^*)\leq 0 \\ \alpha_i^* \geq 0 \\ h_j(x^*)=0 \end{cases}$$

2.2. 支持向量机的求解

$$L(w,\alpha)=\frac{1}{2}\|w\|^2-\sum_i \alpha_i\{y_i(w^T x_i+b)-1\}$$

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

(1) 求偏导（关于 w, b）

$$\nabla_w L(w,b,\alpha)=w-\sum_{i=1}^{N}\alpha_i y_i x_i=0$$

$$\nabla_b L(w,b,\alpha)=\sum_{i=1}^{N}\alpha_i y_i=0$$

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

问题转化为 $\max\limits_{\alpha} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle + \sum_i \alpha_i$

$\quad$ s.t. $\sum_i \alpha_i y_i = 0$, $\alpha_i \geq 0$

则 $\nabla_w L(w^*, b^*, \alpha^*) = 0$, $w^* = \sum_{i=1}^{N}\alpha_i^* y_i x_i$

$\nabla_b L(w^*, b^*, \alpha^*) = 0$ $\alpha_i^*\{y_i(w^{*T}x_i + b) - 1\} = 0$, $\alpha_i \geq 0$, $y_i(w^{*T}x + b^*) - 1 \geq 0$

① if $\alpha_i \neq 0$ $y_i(w^* x_i + b^*) - 1 = 0$. $(x_i, y_i) \rightarrow$ 支持向量

② $w^* = \sum_{i=1}^{N}\alpha_i^* y_i x_i$, $\alpha_i^* > 0$ $\longrightarrow$ w^* 只与支持向量有关

③ 至少存在一个 $\alpha_j > 0$

3. 线性支持向量机与软间隔

3.1 线性支持向量机

引入松弛变量 $\xi_i \geq 0$, 则

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

s.t. $y_i(w^T x_i + b) \geq 1 - \xi_i$, $\xi_i \geq 0$ 软间隔最大化

3.2 对偶算法

$$L(w,b,\xi,\alpha,\mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}[\alpha_i\{y_i(w^T x_i + b) - 1 + \xi_i\} + \mu_i\xi_i]$$

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

$\nabla_b L = \sum_{i=1}^{N}\alpha_i y_i = 0$

$\nabla_{\xi_i} L = C - \alpha_i - \mu_i = 0$

$\min\limits_{w,b,\xi_i} L(w,b,\xi,\alpha,\mu) = -\frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle + \sum_{i=1}^{N}\alpha_j$

s.t. $\sum_i \alpha_i y_j = 0$, $C - \alpha_i - \mu_i = 0$, $0 \leq \alpha_i \leq C$, $0 \leq \mu_i \leq C$

① w 只与 α_i 有关

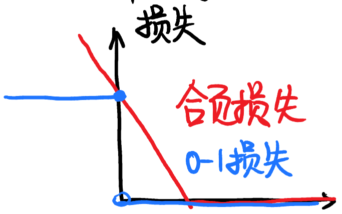
② $y_i(w^{*T}x_i + b) - 1 + \xi_i^* = 0$

支持向量 $\alpha_i^* \geq 0$,
(1) $\alpha_i^* < C$, $\mu_i^* > 0$, $\xi_i = 0$, 对应 (x_i, y_i) 在间隔线上
(2) $\alpha_i^* = C$, $\mu_i^* = 0$, $\xi_i > 0$
1° $\xi_i < 1$, 分类正确
2° $\xi_i = 1$, 落在分类平面
3° $\xi_i > 1$, 错误分类

3.3. 合页损失函数 (Hinge Loss)

线性支持向量机等价于最小化 $\sum_i[1-y_i(wx_i+b)]_+ + \lambda\|w\|^2$

其中 $[\cdot]_+ = \max(\cdot, 0)$



① 合页损失是0-1损失的上界，常作为一种代理损失

② 合页在确信足够高时才是0，对学习有更高效率

4. 非线性支持向量机与核函数

4.1 核技巧

1. 非线性分类问题

如果能用 R^n 中的超平面将正负例分开，则称此问题为线性可分问题

核技巧：通过非线性变换，将输入空间对应于一个特征空间，使在输入空间的超曲面模型对应于超平面模型

2. 核函数定义

设 $\mathcal{X}$ 是输入空间，设H是特征空间，如果存在 $x \to H$ 映射 $\Phi(x)$，使对于所有 $x, z \in \mathcal{X}$，函数 $K(x,z)=\Phi(x)\cdot\Phi(z)$，则称 $K(x,z)$ 为核函数

(1) 特征空间一般高维，甚至无穷维

(2) 以上映射不是唯一的

3. 核技巧的应用

目标函数：$W(\alpha)=\frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^N \alpha_i$

$f(x)=\text{sign}(\sum_i \alpha_i^* y_i K(x_i,x)+b^*)$，不需要显式定义 $\Phi(x)$

4.2 正定核

1. 正定核的充要条件

定理1：(正定核的充要条件) 设 $K: \mathcal{X}\times\mathcal{X}\to R$ 是对称函数，则 $K(x,z)$ 为正定核函数的充要条件是对任意 $x_i\in\mathcal{X}$ $(i=1,2,\cdots,m)$，$K(x,z)$ 对应的Gram矩阵：$K=[K(x_i,x_j)]_{m\times m}$ 是半正定矩阵

2. 常用核函数

· 多项式核函数(polynomial kernel function): $K(x, z) = (x \cdot z + 1)^p$

· 高斯核函数 (Gaussian kernel function): $K(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right)$