PB21081601 张芷苒

> 1.试证明对于不含冲突数据集（即特征向量完全相同但标记不同）的训练集，必存在与训练集一致（即训练误差为 0) 的决策树。

对于一个不含冲突的数据集，任意两个样本的特征向量要么不同，要么它们具有相同的标记。

构造这样一个决策树：

1. 从根节点开始，如果当前节点的所有数据点具有相同的标记，则将当前节点标记为叶节点，叶节点的标记就是这些数据点的标记。
2. 否则，找到一个特征，使得该特征可以将数据集分成至少两个子集。将当前节点分裂为基于该特征值的若干个子节点。
3. 对每个子节点，递归地应用上述步骤。

由于数据集中没有冲突数据（即没有两个特征相同但标签不同的数据点），因此每次分裂都会将数据集分成更小的子集，直到每个叶节点的数据点都具有相同的标记。这样构造出的决策树能够完美地分类训练集中的所有数据点，因此训练误差为 0。

> 2.最小二乘学习方法在求解 $\min_w (Xw-y)^2$ 问题后得到闭式解 $w^* = \left(X^T X\right)^{-1} X^T y$（为简化问题，我们忽略偏差项 b)。如果我们知道数据中部分特征有较大的误差，在不修改损失函数的情况下，引入规范化项 $\lambda w^T D w$ ，其中 D 为对角矩阵，由我们取值。相应的最小二乘分类学习问题转换为以下形式的优化问题:
> $\min_w (Xw-y)^2 + \lambda w^T D w$
> (1).请说明选择规范化项 $w^T D w$ 而非 $L2$ 规范化项 $w^T w$ 的理由是什么。 D 的对角线元素 D_{ii} 有何意义，它的取值越大意味着什么?
> (2).请对以上问题进行求解。

(1) 选择规范化项 $w^T D w$ 的理由是：当我们知道数据中的某些特征具有较大的误差时，可以通过对这些特征引入更大的惩罚来减少它们对模型的影响。矩阵 D 的对角线元素D_{ii} 代表第 i 个特征的权重，取值越大意味着对第 i 个特征施加更大的惩罚，从而降低其对模型的影响。

(2) 最小化问题：

$$\min_w (Xw-y)^2 + \lambda w^T D w$$

将其转换为标准形式：

$$L(w) = (Xw-y)^T (Xw-y) + \lambda w^T D w$$

对 w求导并设导数为零：

$$
\frac{\partial L(w)}{\partial w}=2X^T(Xw-y)+2\lambda Dw=0
$$

得到：

$$
X^TXw-X^Ty+\lambda Dw=0
$$

解得：

$$
(X^TX+\lambda D)w=X^Ty
$$

最终得到 w的闭式解：

$$
w=(X^TX+\lambda D)^{-1}X^Ty
$$

> 3.假设有 n 个数据点 $x_1,...,x_n$ 以及一个映射 $\varphi:x\rightarrow\varphi(x)$, 以此定义核函数 $K\left(x,x^{\prime}\right)=\varphi(x)\cdot\varphi\left(x^{\prime}\right)$ 。试证明由该核函数所决定的核矩阵 $K:K_{i,j}=K\left(x_i,x_j\right)$ 有以下性质:
> (1). K 是一个对称矩阵;
> (2). K 是一个半正定矩阵，即 $\forall z\in R^n,z^TKz\geq 0$ 。

(1) 由于核函数 $K(x,x')$ 由内积定义，即 $K(x,x')=\varphi(x)\cdot\varphi(x')$，而内积具有对称性，因此：

$$
K(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)=\varphi(x_j)\cdot\varphi(x_i)=K(x_j,x_i)
$$

因此核矩阵 K 是对称的。

(2) 对于任意向量 $z\in\mathbb{R}^n$，有：

$$
z^TKz=\sum_{i=1}^{n}\sum_{j=1}^{n}z_iz_jK(x_i,x_j)
$$

由于 $K(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)$，可以写成：

$$
z^TKz=\sum_{i=1}^{n}\sum_{j=1}^{n}z_iz_j(\varphi(x_i)\cdot\varphi(x_j))=(\sum_{i=1}^{n}z_i\varphi(x_i))\cdot(\sum_{j=1}^{n}z_j\varphi(x_j))=\|\sum_{i=1}^{n}z_i\varphi(x_i)\|^2
$$

由于内积的非负性，因此：

$$
z^TKz=\|\sum_{i=1}^{n}z_i\varphi(x_i)\|^2\geq 0
$$

因此核矩阵 K 是半正定的。

> 4.已知正例点 $x_1=(1,2)^T,x_2=(2,3)^T,x_3=(3,3)^T$ ，负例点 $x_4=(2,1)^T,x_5=(3,2)^T$ ，试求Hard Margin SVM的最大间隔分离超平面和分类决策函数，并在图上画出分离超平面、间隔边界及支持向量。

已知正例点 $x_1 = (1, 2)^T, x_2 = (2, 3)^T, x_3 = (3, 3)^T$），负例点$$x_4 = (2, 1)^T, x_5 = (3, 2)^T$。

需要求解以下二次规划问题：

$$\min_w \frac{1}{2}\|w\|^2$$

约束条件：

$$y_i(w \cdot x_i + b) \geq 1, \forall i$$

引入拉格朗日乘子 α_i，构造拉格朗日函数：

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

对 w 和 b 求导并设导数为零：

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

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

将 w 代入拉格朗日函数，得到对偶问题：

$$\max_\alpha \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i \cdot x_j$$

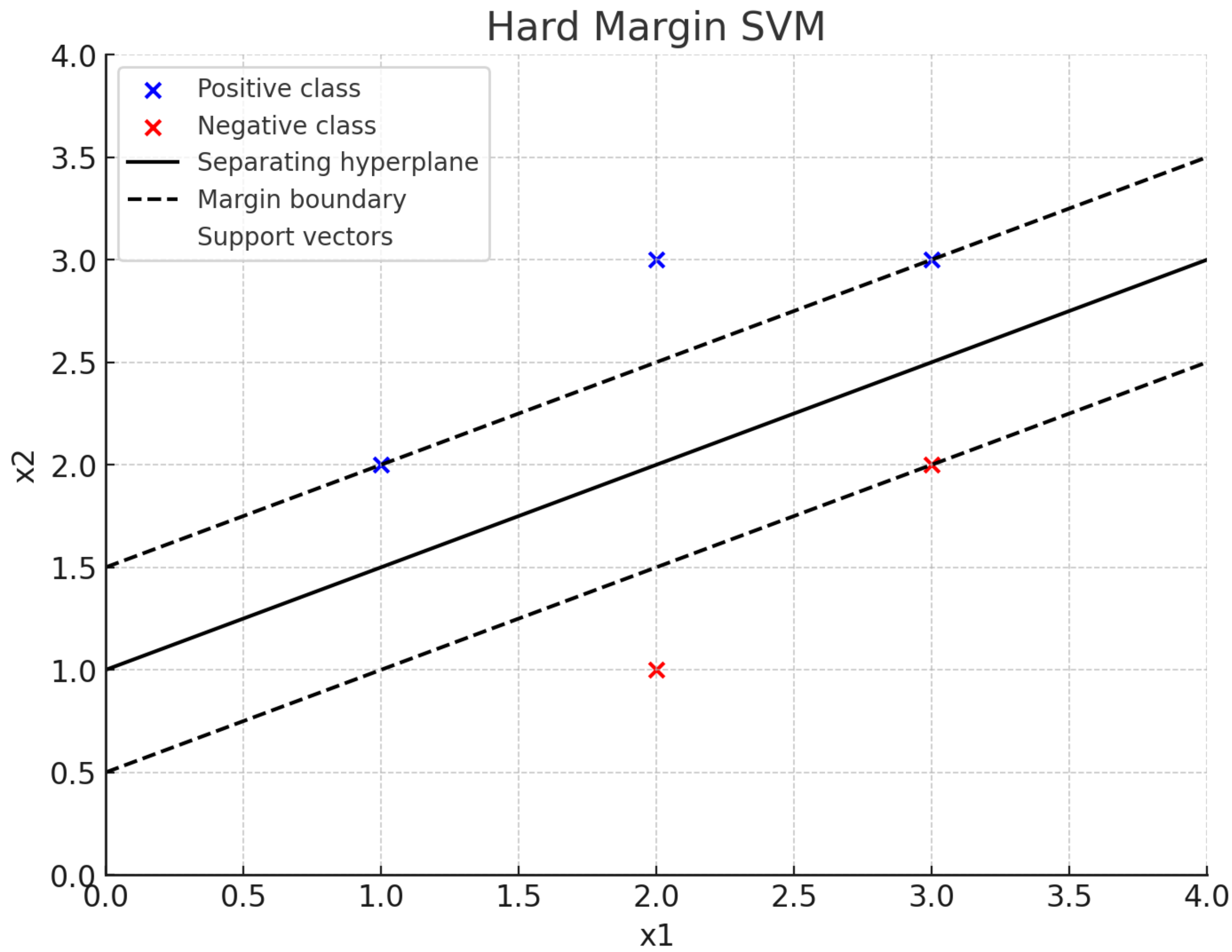
约束条件：

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\alpha_i \geq 0, \forall i$$

通过求解这个二次规划问题，可以得到拉格朗日乘子 (\alpha) 的值，进而得到w, b。然后分类决策函数为：

$$f(x) = \text{sign}(w \cdot x + b)$$



> 5.计算 $\frac{\partial}{\partial w_j} L_{CE}(w,b)$，其中
> $L_{CE}(w,b) = -[y \log \sigma(w \cdot x + b) + (1-y)\log(1-\sigma(w \cdot x + b))]$
> 为Logistic Regression的Loss Function。
> 已知
>
> $$\frac{\partial}{\partial z}\sigma(z) = \frac{\partial}{\partial z}\frac{1}{1+e^{-z}} = -\left(\frac{1}{1+e^{-z}}\right)^2 \times \left(-e^{-z}\right)$$
> $$= \sigma^2(z)\left(\frac{1-\sigma(z)}{\sigma(z)}\right) = \sigma(z)(1-\sigma(z))$$

$$\text{已知}$$

L_{CE}(w, b) = -[y \log \sigma(w \cdot x + b) + (1 - y) \log (1 - \sigma(w \cdot x + b))]

$$\text{其中\$}\sigma(z) = \frac{1}{1+e^{-z}}\text{\$。设\$}z = w \cdot x + b\text{\$\$，则}$$

\sigma(z) = \frac{1}{1 + e^{-z}}

$$\text{对\$}L_{CE}\text{\$进行求导：}$$

\frac{\partial L_{CE}}{\partial w_j} = - \left[y \frac{\partial}{\partial w_j} \log \sigma(z) + (1 - y) \frac{\partial}{\partial w_j} \log (1 - \sigma(z)) \right]

利用链式法则：

\frac{\partial}{\partial w_j} \log \sigma(z) = \frac{1}{\sigma(z)} \cdot \sigma(z) (1 - \sigma(z)) x_j = (1 - \sigma(z)) x_j

\frac{\partial}{\partial w_j} \log (1 - \sigma(z)) = \frac{1}{1 - \sigma(z)} \cdot (- \sigma(z) (1 - \sigma(z))) x_j = - \sigma(z) x_j

因此：

\frac{\partial L_{CE}}{\partial w_j} = - \left[y (1 - \sigma(z)) x_j - (1 - y) \sigma(z) x_j \right]

= - \left[y x_j - y \sigma(z) x_j

- \sigma(z) x_j + y \sigma(z) x_j \right]

  = - \left[y x_j - \sigma(z) x_j \right]

  = (\sigma(z) - y) x_j
  $$

> 6.K - means 算法是否一定会收敛? 如果是，给出证明过程；如果不是，给出说明.

是的.

证明:
K - means 算法通过在每次迭代中减少目标函数（簇内误差平方和）来进行聚类。目标函数定义为：

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2$$

其中 μ_i 是簇 C_i 的质心。

每次迭代包括：

1. 分配步骤： 将每个数据点分配到最近的簇。
2. 更新步骤： 重新计算每个簇的质心。

由于数据点的分配和质心的更新都不会增加目标函数 J，并且目标函数是有界的（非负），因此算法会在有限次迭代后收敛到一个局部最小值。