HW8

PB21111653 李宇哲

T1

Question：

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。

Answer：

用递归的方式构建一课与训练集一致的决策树

首先，选择一个划分属性，将数据集根据该属性划分为若干个子集，然后再每个子集中递归构建决策树，直到所有子集中的数据都属于同一类别或者无法再次选择划分属性为止

由于数据集不含冲突数据，每个子集的数据都属于同一类别，递归构建决策树时，每个叶子结点都属于一个数据的类别，且每个叶子结点的数据集都只包含属于该类别的数据

每个叶子节点对应于一个数据的类别，而且每个非叶子节点对应于一个划分属性，将数据集划分为若干个子集。由于数据集不含冲突数据，因此在每个子集中，所有的数据都属于同一类别。这样的决策树可以将每个特征向量映射到其正确的类别，训练误差为 0。

T2

Quesiton：

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 $w^*=(X^TX)^{-1}X^Ty$ （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 λw^TDw ,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：

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

(1).请说明选择规范化项 w^TDw 而非 $L2$ 规范化项 w^Tw 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。

Answer：

(1)

选择规范化项w^TDw的原因是，可以对不同特征的权重进行不同程度的约束，从而避免某些特征对预测结果的贡献过大

取值越大表示该特征的权重受到的约束越强，即该特征对预测结果的贡献更小

(2)

$\min_w \left(|Xw-y|^2+\lambda w^T Dw\right)$ 为了求解上述优化问题，我们可以先对w进行展开：
$w=[w_1 w_2 \ \cdots \ w_d]$ 其中，d为特征的数量。则有：

$$\begin{aligned}\lambda w^T Dw &= \lambda \sum_{i=1}^{d} w_i^2 D_{ii} \\ &= w^T \begin{bmatrix} \lambda D_{11} & 0 & \cdots & 0 \\ 0 & \lambda D_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda D_{dd} \end{bmatrix} w \\ &= w^T \Lambda w\end{aligned}$$

其中，Λ即为第二步中的对角矩阵，其对角线元素为λD_{ii}。

此时问题转化为求解：

$$\min_w \left(|Xw-y|^2+w^T\Lambda w\right)$$

对w求导，令导数为0，可得： $X^TXw-X^Ty+\Lambda w=0 \ w^*=(X^TX+\Lambda)^{-1}X^Ty$

T3

Question：

3.假设有 n 个数据点 $x_1,\ldots,x_n$ 以及一个映射 $\varphi: x \to \varphi(x)$,以此定义核函数 $K(x,x')=\varphi(x)\cdot\varphi(x')$ 。试证明由该核函数所决定的核矩阵 K： $K_{i,j}=K(x_i,x_j)$ 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 $\forall z\in R^n, z^TKz\geq 0$ 。

Answer：

(1)

证明：

对于 $\forall i,j\in[n]$，对于矩阵K我们有：

$$\begin{aligned}K_{i,j} &= \varphi(x_i)\cdot\varphi(x_j) \\ &= \varphi(x_j)\cdot\varphi(x_i) \\ &= K_{j,i}\end{aligned}$$

因此K是对称矩阵

(2)

将K做矩阵展开，并把二次型写为和式

$$
\begin{aligned}
z^TKz &= \sum_{i=1}^{n}\sum_{j=1}^{n} z_iK_{i,j}z_j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} z_i\varphi(x_i)^T\varphi(x_j)z_j \\
&= \sum_{i=1}^{n} z_i\varphi(x_i)^T\sum_{j=1}^{n} z_j\varphi(x_j) \\
&= ||\sum_{i=1}^{n} z_i\varphi(x_i)||^2 \\
&\geq 0
\end{aligned}
$$

T4

Question：

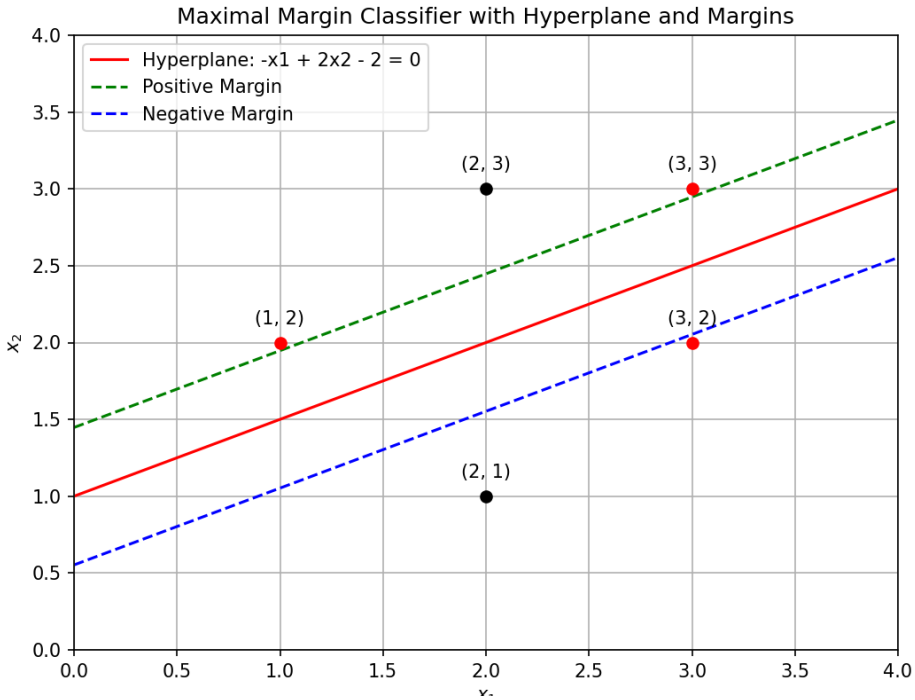
4.已知正例点 $x_1=(1,2)^T, x_2=(2,3)^T, x_3=(3,3)^T$ ，负例点 $x_4=(2,1)^T, x_5=(3,2)^T$ ，试求Hard Margin SVM的最大间隔分离超平面和分类决策函数，并在图上画出分离超平面、间隔边界及支持向量。

Answer：

最大间隔分离超平面也就是以下问题的条件极值问题：
$min_{\boldsymbol{w},b}\frac{1}{2}||\boldsymbol{w}||^2 \quad subject\ to\ \forall i, y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)\geq 1$ 将数据点代入约束条件，得到：
$y_1(\boldsymbol{w}^T\boldsymbol{x}_1+b)\geq 1 \qquad y_2(\boldsymbol{w}^T\boldsymbol{x}*2+b)\geq 1\ y_3(\boldsymbol{w}^T\boldsymbol{x}\ *3+b)\geq 1 \qquad y_4(\boldsymbol{w}^T\boldsymbol{x}\ *4+b)\geq 1\ y_5(\boldsymbol{w}^T\boldsymbol{x}$
显然这个问题是一个二维问题，因此可以将$\boldsymbol{w}$表示为$\boldsymbol{w}=(w_1,w_2)^T$。再将上面的约束转化为等式约束，令拉格朗日乘子$\alpha_i\geq 0$可以得到：
$L=\frac{1}{2}(w_1^2+w_2^2)+\sum\ *i=1^5\alpha_i(1-y_i(\boldsymbol{w}^T\boldsymbol{x}\ *i+b))$ 将上述等式代入原优化问题，得到：

$$
\begin{aligned}
\max\ *\boldsymbol{\alpha} \quad & \sum\ *i=1^n\alpha_i-\frac{1}{2}\sum\ *i=1^n\sum *j=1^n\alpha_i\alpha_jy_iy_j\boldsymbol{x}_i^T\boldsymbol{x}*j \ \text{s.t.} \\
& \sum *i=1^n\alpha_iy_i=0 \qquad\qquad \alpha_i\geq 0,\quad i=1,2
\end{aligned}
$$

经过求解，得到$w_1=-1,w_2=2,b=-2$，因此最大间隔分离超平面的解析式为：
$x^{(1)}-2x^{(2)}+2=0$ 分类决策函数为：$f(\boldsymbol{x})=\text{sign}(x^{(1)}-2x^{(2)}+2)$ 支持向量为：
$\boldsymbol{x}_1=(1,2)^T\ \boldsymbol{x}_3=(3,3)^T\ \boldsymbol{x}_5=(3,2)^T$ 图像如下



T5

Question：

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

$\frac{\partial}{\partial z}\sigma(z) = \frac{\partial}{\partial z}\frac{1}{1+e^{-z}} = -(\frac{1}{1+e^{-z}})^2 \times (-e^{-z})$

$= \sigma^2(z)(\frac{1-\sigma(z)}{\sigma(z)}) = \sigma(z)(1-\sigma(z))$

Answer：

根据链式法则，可以得到：$\frac{\partial}{\partial w_j}L_{CE}(w,b) = \frac{\partial}{\partial \sigma}L_{CE}(w,b)\cdot\frac{\partial \sigma}{\partial z}\cdot\frac{\partial z}{\partial w_j}$ 首先计算第一项：$\frac{\partial L_{CE}(w,b)}{\partial \sigma} = -\left(\frac{y}{\sigma(w\cdot x+b)} - \frac{1-y}{1-\sigma(w\cdot x+b)}\right)$ 然后计算第二项：$\frac{\partial \sigma}{\partial z} = \sigma(z)(1-\sigma(z))$ 然后计算第三项：$\frac{\partial z}{\partial w_j} = \frac{\partial}{\partial w_j}(w\cdot x+b) = x_j$ 代入原式可以得到：$\frac{\partial}{\partial w_j}L_{CE}(w,b) = -\left(\frac{y}{\sigma(w\cdot x+b)} - \frac{1-y}{1-\sigma(w\cdot x+b)}\right)\cdot\sigma(w\cdot x+b)(1-\sigma(w\cdot x+b))\cdot x_j$ 最后化简一下得到：$\frac{\partial}{\partial w_j}L_{CE}(w,b) = (\sigma(w\cdot x+b)-y)\cdot x_j$

T6

Question：

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

Answer：

该算法最后会收敛，证明需要针对算法的两个步骤作证明：

- 算法第一步会找到将所有的点定位到距离其最近的中心点，处理后会让其距离中心点的距离更小，这样更新的距离一定会比这个点上次的距离要短，因此所有点和中心点的距离在这一步必定减小。
- 算法第二步重新确定了每个聚类的中心点，我们需要证明：对于任意聚类A和其中的点$a_1, \ldots, a_n$，定义其新中心点$C(A) = \frac{1}{A}\sum_{i=1}^{n} a_n$，则对于任意的点$x$，$\sum i=1n|ai-C(A)|2 \leq \sum i=1n|ai-x|2$。

  下面来对右边的的式子进行变换：

$$
\begin{aligned}
\sum_i |a_i - x|^2 &= \sum_i |a_i - C(A) + C(A) - x|^2 \\
&= \sum_i |a_i - C(A)|^2 + \sum_i (a_i - C(A)) \cdot (C(A) - x) + \sum_i |C(A) - x|^2 \\
&= \sum_i |a_i - C(A)|^2 + (C(A) - x) \cdot \sum_i (a_i - C(A)) + |A||C(A) - x|^2 \\
&= \sum_i |a_i - C(A)|^2 + |A||C(A) - x|^2 \\
&\geq \sum_i |a_i - C(A)|^2
\end{aligned}
$$

  倒数第二步用到了$\sum_i (a_i - C(A)) = 0$，定理得证