Homework8

> PB20020480 王润泽

Q1

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。

考虑决策树的生成，算法生成叶节点，并递归返回条件有：

- 当前节点的所有样本属于同一类，叶节点类标签 -> 当前类；
- 当前节点的所有样本在所有属性上取值相同，即都具有相同的特征；那么此时选择为叶节点，叶节点类标签 -> 样本中最多类

由此可见，若两训练数据样本特征向量相同，那么它们会到达决策树的同一叶节点（只代表某一类），若二者数据标签不同（冲突数据），则会出现训练误差，决策树与训练集不一致。

如果没有冲突数据，到达某节点的样本会出现以下两种情况:

- 样本间特征向量相同且属于同一类，满足递归结束条件，该节点为叶节点，类标签正确（无训练误差）；
- 样本间特征向量不同时，递归结束条件不满足，数据会根据属性继续划分，直到上一条情况出现。

综上得证，当数据集不含冲突数据时，必存在与训练集一致（训练误差为0）的决策树。

Q2.

最小二乘学习方法在求解 $min_w(X\omega - y)^2$ 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 $\lambda\omega^TD\omega$,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：

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

(1).请说明选择规范化项 $\omega^TD\omega$ 而非 L_2 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?

添加了一个对角矩阵 D,从而对不同的特征进行不同程度的约束，更加精细地控制模型的复杂度。

λw^TDw 可以理解为对不同特征分量规范权重不同，λD_{ii} 权重越大，第 i 个特征在规范化项中的影响越大，需要更加谨慎地进行约束，以避免过拟合的出现，这会使得随机梯度下降时该特征被尽可能的忽略，以减少影响。所以对于那些有较大误差的特征，应当赋予较大的 D_{ii}

(2).请对以上问题进行求解

对于上述问题对于那些有较大误差的特征，应当赋予较大的 D_{ii}，在每次梯度下降时，迭代为

$$w_j := w_j(1 - \alpha\lambda D_{jj}) - \alpha\sum_{i=1}^{m}(x^{(i)} - y^{(i)})x^{(i)}$$

假设有 n 个数据点 $x_1, ..., x_n$ 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;

证明：

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

所以是一个对称矩阵

(2). K 是一个半正定矩阵，即 $\forall z \in \mathbb{R}^n, z^T K z \geq 0$

$$\begin{aligned} z^T K z &= \sum_{i=1}^{n}\sum_{j=1}^{n} z_i K_{ij} z_j \\ &= \sum_{i=1}^{n} z_i \varphi(x_i) \sum_{j=0}^{n} z_j \varphi(x_j) \\ &= (\sum_{i=1}^{n} z_i \varphi(x_i))^2 \geq 0 \end{aligned}$$

所以K是一个半正定矩阵

4. $K-means$ 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

首先对于待优化的问题是最小化损失函数

$$J(\mu, C) = \sum_{j=1}^{n} ||\mu_{C(j)} - x_j||^2 \geq 0$$

其中 C有K种分类取值，每个 x_j 都对应了唯一一个分类；$\mu_{c(j)}$ 是 C为某种取值时的平均值,共有k个。

在 $K-means$ 算法中，每次迭代

1. 固定 μ，优化 C

$$\min_C \sum_{i=1}^{K}\sum_{x\in C_i} |x-\mu_i|^2 = \min_C \sum_{j}^{n} |x_j - \mu_j|$$

2. 固定 C，优化 μ

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

$$\mu_i = \frac{1}{|C_i|}\sum_{x\in C_i} x$$

由于每次迭代都是寻找最小值，所以每次迭代 J 都是递减的，且而 J 有下界。由单调有界定理，迭代最终一定可以收敛。

$$\lim_{k\to\infty} J^{(k)} \to 0$$