HW8

PB20111689 蓝俊玮

T1

> 试证明对于不含冲突数据（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

对于所有特征值都为离散的决策树来说，从根节点到叶节点上所有特征值就会构成一个特征向量。那么对于该决策树来说，因为不存在冲突数据，所以可以视为每一条路径上的特征向量都不同，那么训练集上的所有所有数据都会对应到一个叶节点。这样的话，对于每一个特征向量，最终都只有一条路径且指向其训练数据的标签上，所以必定存在与训练集一致的决策树。

对于特征值不全为离散的决策树来说，其整体原理和上述是一样的，不同的是中间节点将使用区间值表示。从决策树的基本学习算法看来，对于一个中间节点来说，在训练的过程中，如果存在两个数据的标签不同，那么这个中间节点的特征将会被进一步划分。因此和上述情况类似，当特征向量类似的两个数据将会被划分到同一个叶节点上。因此对于整个数据集来说，特征向量相同的数据只会对应到一个叶节点上。因此必定存在与训练集一致的决策树。

T2

> 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 $\lambda w^T Dw$,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，它的取值越大意味这什么？
> 2. 请对以上问题进行求解。

1. 选择规范化项 w^TDw 的理由是因为它可以针对具体特征的误差进行调节，从而减小有较大误差的那部分特征的影响。通过引入对角矩阵 D，我们可以对特定特征的权重进行缩放，从而抑制那些具有较大误差的特征对模型的影响。这种做法可以在某种程度上提高模型的鲁棒性，使得模型对于特征误差的影响降到最低。

   当 D_{ii} 的取值较大时，规范化项 λw^TDw 在优化问题中的作用也相应增强。优化算法会更倾向于降低对应特征的权重，以减小特征误差对模型的影响。这样做的效果是抑制那些具有较大误差的特征，比 $L2$ 规范化项更容易获得稀疏解，使模型更加健壮和稳定。

2. 目标优化函数为：

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

   对其进行求导可以得到：

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

   则可以求解出闭式解 w^*：

$$
\begin{aligned}
&X^TXw+\lambda Dw=X^Ty\\
&\Rightarrow (X^TX+\lambda D)w=X^Ty\\
&\Rightarrow w^*=(X^TX+\lambda D)^{-1}X^Ty
\end{aligned}
$$

T3

> 假设有 n 个数据点 $x_1,\ldots,x_n$ 以及一个映射 $\varphi: x\rightarrow\varphi(x)$，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 $\forall z\in R^n$ 有 $z^TKz\geq 0$

1. 根据核函数的定义，我们有 $K(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)$。由于内积运算满足交换律，即 $\varphi(x_i)\cdot\varphi(x_j)=\varphi(x_j)\cdot\varphi(x_i)$，因此有 $K(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)=\varphi(x_j)\cdot\varphi(x_i)=K(x_j,x_i)$。根据核矩阵的定义，我们有 $K_{i,j}=K(x_i,x_j)$，$K_{j,i}=K(x_j,x_i)$。根据上述推导，我们得到 $K_{i,j}=K(x_i,x_j)=K(x_j,x_i)=K_{j,i}$。因此，由核函数决定的核矩阵 K 是一个对称矩阵。

2. 首先，根据核矩阵的定义，我们有 $K_{i,j}=K(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)$。然后，考虑向量 z，我们可以将其表示为 $z=[z_1,z_2,\ldots,z_n]^T$，其中 z_i 是向量 z 的第 i 个元素。那么就有：

   $$z^TKz=\sum_{i=1}^n\sum_{j=1}^n z_iz_jK_{i,j}$$

   将 $K_{i,j}$ 替换为 $\varphi(x_i)\cdot\varphi(x_j)$：

   $$
   \begin{aligned}
   z^TKz&=\sum_{i=1}^n\sum_{j=1}^n z_iz_j\Big(\varphi(x_i)\cdot\varphi(x_j)\Big)\\
   &=\sum_{i=1}^n\sum_{j=1}^n z_iz_j\Big(\varphi(x_j)\cdot\varphi(x_i)\Big)\\
   &=\sum_{i=1}^n\left(\sum_{j=1}^n z_iz_j\Big(\varphi(x_j)\cdot\varphi(x_i)\Big)\right)\\
   &=\sum_{i=1}^n\left(z_i\varphi(x_i)\cdot\sum_{j=1}^n\Big(z_j\varphi(x_j)\Big)\right)\\
   &=\left(\sum_{i=1}^n z_i\varphi(x_i)\right)\cdot\left(\sum_{j=1}^n z_j\varphi(x_j)\right)
   \end{aligned}
   $$

   由于内积的性质，$\varphi(x_i)\cdot\varphi(x_i)=|\varphi(x_i)|^2\geq 0$，所以每个内积都是非负的。因此，我们有：

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

   因此对于 $\forall z\in R^n$ 都有 $z^TKz\geq 0$。所以 K 是一个半正定矩阵。

T4

> 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

K-means 的损失函数为 $E=\sum_{i=1}^{k}\sum_{\boldsymbol{x}\in C_i}||\boldsymbol{x}-\boldsymbol{\mu}_i||_2^2$，则有 $\frac{\partial E}{\partial \boldsymbol{\mu}_i}=2\sum_{\boldsymbol{x}\in C_i}(\boldsymbol{x}-\boldsymbol{\mu}_i)=0$ 得到 $\boldsymbol{\mu}_k=\frac{1}{|C_i|}\sum_{\boldsymbol{x}\in C_i}\boldsymbol{x}=\boldsymbol{\mu}_k'$，可以得知在更新之后的均值向量 $\boldsymbol{\mu}_k'$ 是使得损失函数最小的一个极值点。那么就说明了，在每次更新时以均值向量作为中心点时，都能让损失函数 E 变得更小。因此 K-means 算法的更新能够让损失函数 E 单调递减，同时又因为 $E\geq 0$ 是有界的，因此 K-means 算法具有收敛性。