hw8

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。

   假设不存在与训练集一致的决策树,那么,必然存在这样一个结点,有无法划分的多个数据,然而,这与不含冲突数据集产生了矛盾.因此,必然有与训练集一致的决策树.

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?
      1. 引入这一规范化项的原因是要对误差大的特征权重进行调整,使得这些误差很大的特征对我们的训练结果影响减小.而如果仅仅使用$L2$规范化项的话,对特征在模型训练中的权重没有影响.
      2. D_{ii}代表第i个特征的权重,取值越大,代表这个特征在模型中的权重越大
   2. 请对以上问题进行求解。

      对$f(w)=min_w(Xw-y)^2+\lambda w^TDw$ 求导可得

      $\frac{\partial f}{\partial w}=2(Xw-y)^TX+2\lambda Dw$

      令导数为0可得

      $(Xw-y)^TX+\lambda Dw=0$

      展开为

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

      整理可得

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

      移项即得

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

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;

      $K_{ij}=K(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)$

      $K_{ji}=K(x_j,x_i)=\varphi(x_j)\cdot\varphi(x_i)$

      因此$K_{ij}=K_{ji}$

      所以是对称矩阵

   2. K 是一个半正定矩阵，即 $\forall z\in R^n,z^TKz\geq 0$ 。

      计算上式可得

      $\sum_{i=1}\sum_{j=1}z_iz_jK_{ji}$

      由上一问的对称矩阵性质发现

      即证

      $\sum_i\sum_j z_iz_j\varphi(x_i)\varphi(x_j)\geq 0$

      定义辅助向量$F=(z_1\varphi(x_1),\ldots,z_n\varphi(x_n))$

      上式即等价于

$$|F|^2 \geq 0$$

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

一定会

1. goal function显示

$$\blacktriangleright \ \text{Goal: } \min_{\mu} \min_{C} \sum_{j=1}^{n} \left\| \mu_{C(j)} - \mathbf{x}_j \right\|^2$$

此式必存在为0的下界

2. K-means算法迭代时评价函数不会递增

因此,对于一个单调有界的数列(把算法迭代的序列作为一个数列),其一定会收敛.