Lecture 13-14作业

李星毅 U201712072 自实1701

6、已知两类样本，

。

，试用PCA变化做一维特征提取，给出新的一维特征值

解：（1）中心化：由于样本均值为



故无需中心化。

（2）计算协方差矩阵：



（3）求协方差矩阵的特征值和特征向量：



可以解得：，故有



（4） 选择对应的作为变换矩阵



从而有



综上所述：



7、特征选择的判据作用是什么？判据构造的原则是什么？请列举有哪些不同类型的判据，并简单说明至少三种特征选择方法。

解：

1. 特征选择的判据作用是用来衡量特征对分类的效果。
2. 判据构造的原则是：
3. 判据应该与错误率（或错误率的上界）有单调关系，这样才能较好地反映分类目标。
4. 当特征独立时，判据对特征应该具有可加性，即



这里是第类和第类的可分性准则函数，越大，两类的分离程度越大，是一系列特征变量。

1. 判据应该具有以下度量特性：



1. 理想的判据应该对特征具有单调性，即加入新的特征不会使判据减少，即



1. 可分性判据
   1. 基于类内类间距离的可分性判据



* 1. 基于概率分布的可分性判据



* 1. 基于熵的可分性判据



* 1. 利用统计检验作为可分性判据：最常用的比较两组样本差别的方法是t检验

1. 特征选择方法
   1. 最优搜索算法：
      1. 穷举法：计算每一可能的组合，逐一比较准则函数，适用于或很小的情况。
      2. 分支定界法（BAB）：按照一定的顺序将所有可能的组合排成一棵树，沿树进行搜索，避免一些不必要的计算，使找到最优解的机会最早。该方法是一种从包含所有候选特征开始，逐步去掉不被选中的特征的自顶向下的方法，具有回溯的过程，能够考虑到所有可能的组合。其基本思想是：设法将所有可能的特征组合构建成一个树状结构，按照特定的规律对树进行搜索，使得搜索过程尽可能早的达到最优解而不必遍历整棵树。要做到这一点必须要求准则判据对特征具有单调性，且当处理高维数据时，算法复杂度较高，所以很难广泛使用。
   2. 次优搜索算法（启发式搜索）：很多情况下，即使采取分支定界法，全局寻优的方法计算量可能仍然很大，因此提出了一些基于启发式搜索的次优选择算法：单独最优特征组合，序列前向选择（SFS）及广义的SFS（GSFS），序列后向选择(SBS)及广义的SBS(GSFS)，增l去r选择及广义的增l去r选择方法，浮动搜索方法及浮动的广义后向选择方法(FGSBS)等。
      1. 单独最优组合：对每一个特征单独计算类别可分性判据，根据单个特征的判据值从大到小排序，选前d个单独最佳的特征。
      2. SFS法
      3. GSFS法
      4. SBS法
      5. GSBS法
      6. L-R法
      7. 法
   3. 随机搜索算法——遗传算法：把优化问题比喻作在无数可能的重组和突变组合中发现适应性最强的组合的问题。该类方法在计算中将特征选择问题与遗传算法、模拟退火算法、粒子群优化算法，随机森林或一个随机重采样过程结合起来，以概率推理和采样过程作为算法基础，基于分类估计的有效性，在算法运行过程中对每个特征赋予一定的权重，再根据给定的或自适应的阈值对特征的重要性进行评价。例如，Relief及其扩展算法就是一种典型的根据权重选择特征的随机搜索方法，它能有效的去掉无关特征，但不能去除冗余特征，且只能用于两类分类问题。
   4. 以分类性能为准则的特征选择方法：Wrapper法。

8、假设一个抛硬币实验，假设正面向上设为U的概率为ρ，反面向上设为D概率为（1−ρ）. 我们进行了3次实验，得到两次正面，一次反面，即序列为′UUD′。

1. 使用最大似然估计求ρ 的估计值

（2）假设ρ的先验概率是服从 ，用最大后验概率估计ρ 的值.

（3）假设ρ的先验概率是服从 ，求贝叶斯估计的ρ的分布。

解：

1. 最大似然估计：



1. 最大后验概率：



1. 贝叶斯估计：

