第五次作业

1、 在短文本分析和图像匹配应用中，使用的典型特征提取方法分别是什么？

2、简述 PCA 和 LDA 的异同

3、给定均值为 d 维 0 向量的样本集合 $X=[x_1,x_2,\cdots,x_N]\in R^{d\times N}$，其中 d 是样本特征的个数，N 是样本个数，即，$\sum_{i=1}^{N}x_i=\mathbf{0}$，$x_i$ 是第 i 个样本。假定使用 PCA 算法对其进行特征变换，需要保留 k 个主成分。

(1) 写出特征变换的流程

(2) 如果特征的维度很高（例如 10000 维），并且 d>>N，应该如何处理？

(3) 写出 KPCA 对样本集合进行特征变换的流程。

4、处理模式识别高维数据的两种基本方法是什么？它们有什么不同之处？

5、根据特征选择与分类器的结合程度，特征选择方法可以分为哪三类？各有什么特性？

6、ID3、C4.5 和 CART 是三种典型的决策树分类算法，请分别说明这三种算法具体使用的决策特征选择方法（注：CART 分别说明分类和回归两种情况）。

7、简述最优特征选择方法的基本思想。

8、简述决策树剪枝的目的，以及两种常用的判断决策树是否需要剪枝的准则。

9、决策树的剪枝策略一般有哪些？各有什么优缺点？

10、给出信息增益、信息增益率、Gini 指数的计算公式，并对如下数据，

（1）计算属性“色泽”的信息增益；

（2）计算属性“触感”的信息增益率；

（3）计算属性“脐部”的 Gini 指数。

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否