1. 以下哪些方法不可以直接来对文本分类？( )

A、Kmeans
B、决策树
C、支持向量机
D、KNN

2. 影响聚类算法结果的主要因素有（ ）

A.已知类别的样本质量；
B.分类准则；
C.特征选取；
D.模式相似性测度

3. 影响基本 K-均值算法的主要因素有（ ）

A. 样本输入顺序；
B. 模式相似性测度；
C. 聚类准则；
D. 初始类中心的选取

4. 如果以特征向量的相关系数作为模式相似性测度，则影响聚类算法结果的主要因素有（ ）

A. 已知类别样本质量；
B. 分类准则；
C. 特征选取；
D. 量纲

5. 欧式距离具有（ ）；马式距离具有（ ）。

A. 平移不变性；
B. 旋转不变性；
C. 尺度缩放不变性；
D. 不受量纲影响的特性

6. 一监狱人脸识别准入系统用来识别待进入人员的身份，此系统一共包括识别 4 种不同的人员：狱警，小偷，送餐员，其他。下面哪种学习方法最适合此种应用需求：( )。

A. 二分类问题
B. 多分类问题
C. 层次聚类问题
D. k-中心点聚类问题
E. 回归问题

7. 影响聚类算法效果的主要原因有：（ ）

A. 特征选取
B. 模式相似性测度
C. 分类准则
D. 已知类别的样本质量

8. 过拟合”只在监督学习中出现，在非监督学习中，没有”过拟合”，这是：（ ）

A. 对的

B. 错的

9. 在有监督学习中， 我们如何使用聚类方法？（ ）

1.我们可以先创建聚类类别， 然后在每个类别上用监督学习分别进行学习

2.我们可以使用聚类“类别 id”作为一个新的特征项， 然后再用监督学习分别进行学习

3.在进行监督学习之前， 我们不能新建聚类类别

4.我们不可以使用聚类“类别 id”作为一个新的特征项， 然后再用监督学习分别进行学习

A. 2 和 4

B. 1 和 2

C. 3 和 4

D. 1 和 3

10. 以下说法正确的是 ：（ ）

1.一个机器学习模型，如果有较高准确率，总是说明这个分类器是好的

2.如果增加模型复杂度， 那么模型的测试错误率总是会降低

3.如果增加模型复杂度， 那么模型的训练错误率总是会降低

A. 1

B. 2

C. 3

D. 1 and 3

11. 以下描述错误的是（ ）

A. SVM 是这样一个分类器，它寻找具有最小边缘的超平面，因此它也经常被称为最小边缘分类器

B. 在聚类分析当中，簇内的相似性越大，簇间的差别越大，聚类的效果就越差

C. 在决策树中，随着树中结点输变得太大，即使模型的训练误差还在继续降低，但是检验误差开始增大，这是出现了模型拟合不足的原因

D. 聚类分析可以看作是一种非监督的分类

12. 在以下不同的场景中,使用的分析方法不正确的有 （）

A. 根据商家最近一年的经营及服务数据,用聚类算法判断出天猫商家在各自主营类目下所属的商家层级

B. 根据商家近几年的成交数据,用聚类算法拟合出用户未来一个月可能的消费金额公式

C. 用关联规则算法分析出购买了汽车坐垫的买家,是否适合推荐汽车脚垫

D. 根据用户最近购买的商品信息,用决策树算法识别出淘宝买家可能是男还是女

13. 如何优化 Kmeans？

14. 描述下 KMeans 初始类簇中心点的选取。

15. 常用的聚类划分方式有哪些？列举代表算法。

16. （判断）聚类是在事先并不知道任何样本类别标签的情况下，通过数据之间的内在关系把样本划分为若干类别，使得同类别样本之间的相似度高，不同类别之间的样本相似度低。

A. 正确

B. 错误

17. 简述 K 均值算法的具体步骤

18. K 均值算法的优缺点是什么？如何对其进行调优？

19. 针对 K 均值算法的缺点，有哪些改进的模型？