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

正确答案: A 分类不同于聚类。

解析：

A：Kmeans 是聚类方法，典型的无监督学习方法。分类是监督学习方法，BCD 都是常见的分类方法。

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

正确答案: BCD

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

正确答案: ABD

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

正确答案: BC

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

正确答案: AB ABCD

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

正确答案：（B）

解析：

二分类：每个分类器只能把样本分为两类。监狱里的样本分别为狱警、小偷、送餐员、其他。二分类肯 定行不通。瓦普尼克 95 年提出来基础的支持向量机就是个二分类的分类器，这个分类器学习过 程就是解一个基于正负二分类推导而来的一个最优规划问题（对偶问题），要解决多分类问题 就要用决策树把二分类的分类器级联，VC 维的概念就是说的这事的复杂度。

层次聚类： 创建一个层次等级以分解给定的数据集。监狱里的对象分别是狱警、小偷、送餐员、或者其 他，他们等级应该是平等的，所以不行。此方法分为自上而下（分解）和自下而上（合并）两种操作方式。

K-中心点聚类：挑选实际对象来代表簇，每个簇使用一个代表对象。它是围绕中心点划分的一种规则，所以这里并不合适。

回归分析：处理变量之间具有相关性的一种统计方法，这里的狱警、小偷、送餐员、其他之间并没有什 么直接关系。

结构分析： 结构分析法是在统计分组的基础上，计算各组成部分所占比重，进而分析某一总体现象的内部结构特征、总体的性质、总体内部结构依时间推移而表现出的变化规律性的统计方法。结构分析法的基本表现形式，就是计算结构指标。这里也行不通。

多分类问题： 针对不同的属性训练几个不同的弱分类器，然后将它们集成为一个强分类器。这里狱警、 小偷、送餐员 以及他某某，分别根据他们的特点设定依据，然后进行区分识别。

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

正确答案：（ABC）

解析：

这道题应该是很简单的，D 之所以不正确，是因为聚类是对无类别的数据进行聚类，不使用已经标记好的数据。

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

答案：（B）

解析：

我们可以评估无监督学习方法通过无监督学习的指标，如：我们可以评估聚类模型通过调整兰德系数

（adjusted rand score）。

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

答案: （B）

解析：

我们可以为每个聚类构建不同的模型， 提高预测准确率；“类别 id”作为一个特征项去训练， 可以有效地总结了数据特征。所以 B 是正确的。

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

答案: （C） 考的是过拟合和欠拟合的问题。

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

答案（C）

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

答案（B）

13. 如何优化 Kmeans？

解析：

使用 Kd 树或者 Ball Tree

将所有的观测实例构建成一颗 kd 树，之前每个聚类中心都是需要和每个观测点做依次距离计算，现在这些聚类中心根据 kd 树只需要计算附近的一个局部区域即可。

14. 描述下 KMeans 初始类簇中心点的选取。

解析：

K-means 算法选择初始 seeds 的基本思想就是：初始的聚类中心之间的相互距离要尽可能的远。

1.从输入的数据点集合中随机选择一个点作为第一个聚类中心

2.对于数据集中的每一个点 x，计算它与最近聚类中心(指已选择的聚类中心)的距离 D(x)

3.选择一个新的数据点作为新的聚类中心，选择的原则是：D(x)较大的点，被选取作为聚类中心的概率较大

4.重复 2 和 3 直到 k 个聚类中心被选出来

5.利用这 k 个初始的聚类中心来运行标准的 k-means 算法

15. 常用的聚类划分方式有哪些？列举代表算法。

解析：

1.基于划分的聚类:K-means，k-medoids，CLARANS。

2.基于层次的聚类：AGNES（自底向上），DIANA（自上向下）。

3.基于密度的聚类：DBSACN，OPTICS，BIRCH(CF-Tree)，CURE。

4.基于网格的方法：STING，WaveCluster。

5.基于模型的聚类：EM,SOM，COBWEB。

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

答案：A

解析：

K 均值聚类（KMeansClustering）是最基础和最常用的聚类算法。

它的基本思想是，通过迭代方式寻找 K 个簇（Cluster）的一种划分方案，使得聚类结果对应的代价函数最小。特别地，代价函数可以定义为各个样本距离所属簇中心点的误差平方和

$$J(c,\mu)=\sum_{i=1}^{M}\| x_i-\mu_{c_i}\|^2$$

其中 xi 代表第 i 个样本，ci 是 xi 所属于的簇，μci 代表簇对应的中心点，M 是样本总数。

17. 简述 K 均值算法的具体步骤

K均值聚类的核心目标是将给定的数据集划分成K个簇，并给出每个数据对应的簇中心点。算法的具体步骤描述如下：

（1）数据预处理，如归一化、离群点处理等。

（2）随机选取K个簇中心，记为$\mu_1^{(0)},\mu_2^{(0)},\cdots,\mu_K^{(0)}$。

（3）定义代价函数：$J(c,\mu)=\min_{\mu}\min_{c}\sum_{i=1}^{M}\|x_i-\mu_{c_i}\|^2$。

（4）令t=0,1,2,⋯ 为迭代步数，重复下面过程直到J收敛：

- 对于每一个样本x_i，将其分配到距离最近的簇

$$c_i^{(t)}\leftarrow\underset{k}{\operatorname{argmin}}\|x_i-\mu_k^{(t)}\|^2; \tag{5.2}$$

- 对于每一个类簇k，重新计算该类簇的中心

$$\mu_k^{(t+1)}\leftarrow\underset{\mu}{\operatorname{argmin}}\sum_{i:c_i^{(t)}=k}\|x_i-\mu\|^2. \tag{5.3}$$

K均值算法在迭代时，假设当前J没有达到最小值，那么首先固定簇中心$\{\mu_k\}$，调整每个样例x_i所属的类别c_i来让J函数减少；然后固定$\{c_i\}$，调整簇中心$\{\mu_k\}$使J减小。这两个过程交替循环，J单调递减：当J递减到最小值时，$\{\mu_k\}$和$\{c_i\}$也同时收敛。

18. K 均值算法的优缺点是什么？如何对其进行调优？

解析：

K 均值算法有一些缺点，例如受初值和离群点的影响每次的结果不稳定、结果通常不是全局最优而是局部最优解、无法很好地解决数据簇分布差别比较大的情况（比如一类是另一类样本数量的 100 倍）、不太适用于离散分类等。但是瑕不掩瑜，K 均值聚类的优点也是很明显和突出的，主要体现在：对于大数据集，K 均值聚类算法相对是可伸缩和高效的，它的计算复杂度是 O(NKt)接近于线性，其中 N 是数据对象的数目，K 是聚类的簇数，t 是迭代的轮数。尽管算法经常以局部最优结束，但一般情况下达到的局部最优已经可以满足聚类的需求。

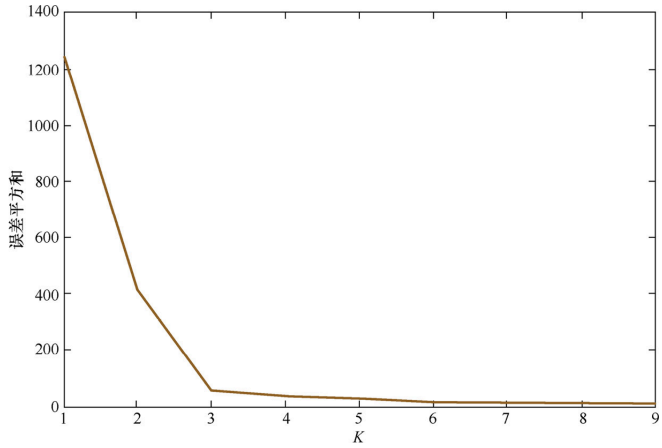
K 均值算法的调优一般可以从以下几个角度出发。

（1）数据归一化和离群点处理。

K 均值聚类本质上是一种基于欧式距离度量的数据划分方法，均值和方差大的维度将对数据的聚类结果产生决定性的影响，所以未做归一化处理和统一单位的数据是无法直接参与运算和比较的。同时，离群点或者少量的噪声数据就会对均值产生较大的影响，导致中心偏移，因此使用 K 均值聚类算法之前通常需要对数据做预处理。

（2）合理选择 K 值。

K 值的选择是 K 均值聚类最大的问题之一，这也是 K 均值聚类算法的主要缺点。实际上，我们希望能够找到一些可行的办法来弥补这一缺点，或者说找到 K 值的合理估计方法。但是，K 值的选择一般基于经验和多次实验结果。例如采用手肘法，我们可以尝试不同的 K 值，并将不同 K 值所对应的损失函数画成折线，横轴为 K 的取值，纵轴为误差平方和所定义的损失函数，如图所示。



由图可见，K 值越大，距离和越小；并且，当 K=3 时，存在一个拐点，就像人的肘部一样；当 K(1,3)时，曲线急速下降；当 K>3 时，曲线趋于平稳。手肘法认为拐点就是 K 的最佳值。

手肘法是一个经验方法，缺点就是不够自动化，因此研究员们又提出了一些更先进的方法，其中包括比较有名的 GapStatistic 方法

（3）采用核函数。

采用核函数是另一种可以尝试的改进方向。传统的欧式距离度量方式，使得 K 均值算法本质上假设了各个数据簇的数据具有一样的先验概率，并呈现球形或者高维球形分布，这种分布在实际生活中并不常见。面对非凸的数据分布形状时，可能需要引入核函数来优化，这时算法又称为核 K 均值算法，是核聚类方法的一种。核聚类方法的主要思想是通过一个非线性映射，将输入空间中的数据点映射到高位的特征空间中，并在新的特征空间中进行聚类。非线性映射增加了数据点线性可分的概率，从而在经典的聚类算法失效的情况下，通过引入核函数可以达到更为准确的聚类结果。

19. 针对 K 均值算法的缺点，有哪些改进的模型？

解析：

K 均值算法的主要缺点如下。

（1）需要人工预先确定初始 K 值，且该值和真实的数据分布未必吻合。

（2）K 均值只能收敛到局部最优，效果受到初始值很大。

（3）易受到噪点的影响。

（4）样本点只能被划分到单一的类中。

Kmeans++算法：

K 均值的改进算法中，对初始值选择的改进是很重要的一部分。而这类算法中，最具影响力的当属 Kmeans++算法。原始 K 均值算法最开始随机选取数据集中 K 个点作为聚类中心，而 Kmeans++按照如下的思想选取 K 个聚类中心。假设已经选取了 n 个初始聚类中心（0<n<K），则在选取第 n+1 个聚类中心时，距离当前 n 个聚类中心越远的点会有更高的概率被选为第 n+1 个聚类中心。在选取第一个聚类中心（n=1）时同样通过随机的方法。可以说这也符合我们的直觉，聚类中心当然是互相离得越远越好。当选择完初始点后，Kmeans++后续的执行和经典 K 均值算法相同，这也是对初始值选择进行改进的方法等共同点。

ISODATA 算法：

当 K 值的大小不确定时，可以使用 ISODATA 算法。ISODATA 的全称是迭代自组织数据分析法。在 K 均值算法中，聚类个数 K 的值需要预先人为地确定，并且在整个算法过程中无法更改。而当遇到高维度、海量的数据集时，人们往往很难准确地估计出 K 的大小。ISODATA 算法就是针对这个问题进行了改进，它的思想也很直观。当属于某个类别的样本数过少时，把该类别去除；当属于某个类别的样本数过

多、分散程度较大时，把该类别分为两个子类别。ISODATA 算法在 K 均值算法的基础之上增加了两个操作，一是分裂操作，对应着增加聚类中心数；二是合并操作，对应着减少聚类中心数。ISODATA 算法是一个比较常见的算法，其缺点是需要指定的参数比较多，不仅仅需要一个参考的聚类数量 Ko，还需要制定 3 个阈值。