第十次作业

阅读机器学习实战 8.1, 8.2, 8.3，9.1

Exercise 10.1

聚类属于哪种学习方式 ( )。

A. 无监督学习
B. 监督学习
C. 强化学习
D. 都不属于

Exercise 10.2

关于K均值和DBSCAN的比较，以下说法不正确的是( )。

A. K均值很难处理非球形的簇和不同大小的簇
B. DBSCAN使用基于密度的概念
C. DBSCAN可以处理不同大小和不同形状的簇。
D. K均值使用簇的基于层次的概念

Exercise 10.3

关于kmean算法的实现描述错误的是 （ ）

A. 收敛速度慢
B. 可以轻松发现非凸形状的簇
C. 原理简单，实现容易
D. 需要事先确定 k 的值

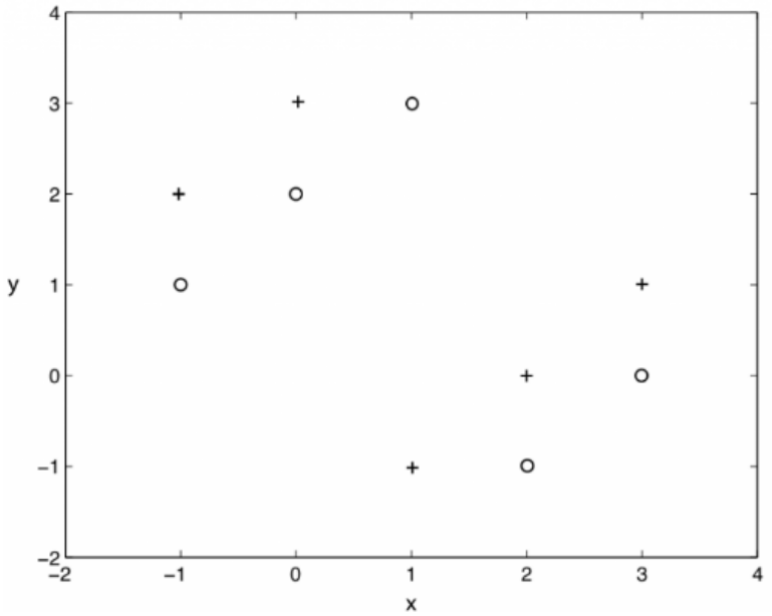
Exercise 10.4

下列关于Kmeans聚类算法的说法错误的是( )。

A. 对大数据集有较高的效率并且具有可伸缩性
B. 初始聚类中心的选择对聚类结果影响不大
C. 是一种无监督学习方法
D. K值无法自动获取，初始聚类中心随机选择

Exercise 10.5

假设我们使用 kNN 训练模型，其中训练数据具有较少的观测数据（下图是两个属性 x、y 和两个标记为 “+” 和 “o” 的训练数据）。现在令 $k = 1$，则图中的 Leave-One-Out 交叉验证错误率是多少?



A. 0%
B. 20%
C. 50%
D. 100%

解释：KNN 算法是标记类算法，取当前实例最近邻的 k 个样本，k 个样本中所属的最多类别即判定为该实例的类别。本题中 k = 1，则只需要看最近邻的那一个样本属于“+” 还是“o”即可。

Leave-One-Out 交叉验证是一种用来训练和测试分类器的方法，假定数据集有N 个样本，将这个样本分为两份，第一份 N-1 个样本用来训练分类器，另一份 1 个样本用来测试，如此迭代 N 次，所有的样本里所有对象都经历了测试和训练。

分别对这 10 个点进行观察可以发现，每个实例点最近邻的都不是当前实例所属的类别，因此每次验证都是错误的。整体的错误率即为 100%。

Exercise 10.6

k-means是一种迭代算法，在其内部循环中重复执行以下两个步骤，哪两个?

A、移动簇中心，更新簇中心u_k

B、分配簇，其中参数$c^{(i)}$被更新

C、移动簇中心u_k,将其设置为等于最近的训练示例$c^{(i)}$

D、簇中心分配步骤，其中每个簇质心u_i被分配（通过设置$c^{(i)}$）到最近的训练示例$x^{(i)}$

Exercise 10.7

给定含有5个样本的集合

$$X = \begin{bmatrix} 0 & 0 & 1 & 5 & 5 \\ 2 & 0 & 0 & 0 & 2 \end{bmatrix}$$

请用 k 均值聚类算法将样本聚到两个类中。

- 第一轮：

  随机选择两个点作为类中心：第一类（0，2）第二类（0，0）

  计算其余各个点到这两个点之间的距离并进行分类，分类结果为

  第一类：（0，0），（1，0），（5，0）

  第二类：（0，2），（5，2）
- 第二轮：

  更新类中心：第一类（2，0），第二类（2.5，2）

  根据新的类中心重新计算距离进行分类，分类结果与上一轮相同，聚类停止

Exercise 10.8

最常用的降维算法是PCA，以下哪项是关于PCA的?

1、PCA是一种无监督的方法
2、它搜索数据具有最大差异的方向
3、主成分的最大数量<= 特征的数量
4、所有主成分彼此正交

A、2，3和4
B、1，2和3
C、1，2和4
D，以上都有

Exercise 10.9

主成分分析（PCA）是一种重要的降维技术，以下对于PCA的描述不正确的是（）：

A、主成分分析是一种无监督方法
B、主成分数量一定小于等于特征的数量
C、各个主成分之间相互正交
D、原始数据在第一主成分上的投影方差最小

Exercise 10.10

应用PCA后，以下哪项可以是前两个主成分?

1、（0.5，0.5，0.5，0.5）和（0.71，0.71，0.0）

2、（0.5，0.5，0.5，0.5）和（0.0，-0.71，0.71）

3、（0.5，0.5，0.5，0.5）和（0.5，0.5，-0.5，-0.5）

4、（0.5，0.5，0.5，0.5）和（-0.5，-0.5，0.5，0.5）

A、1和2
B、1和3
C、2和4
D、3和4

Exercise 10.11

维度的诅咒是什么?

解析：维度的诅咒是指许多在低维空间中不存在的问题，在高维空间中发生。在机器学习领域，一个常见的现象是随机抽样的高维向量通常非常稀疏，提升了过拟合的风险，同时也使得在没有充足训练数据的情况 下，要识别数据中的模式非常困难

Exercise 10.12

一旦降低了数据集的维度，是否可以逆操作？如果可以，怎么 做？如果不能，为什么？

解析：一旦使用我们讨论的任意算法减少了数据集的维度，就几乎不可能再将操作完美地逆转，因为在降维过程中必然丢失了一部分信息。此外，虽然有一些算法（例如PCA）拥有简单的逆转换过程，可以重建出与原始数据集相似的数据集，但是也有一些算法不能实现逆转（例如T-SNE）。

Exercise 10.13

如何定义聚类？你能列举几种聚类算法吗？ 聚类算法的主要应用有哪些？

解析：在机器学习中，聚类是将相似的实例组合在一起的无监督任务。 相似性的概念取决于你手头的任务：例如，在某些情况下，两个附近的实例将被认为是相似的，而在另一些情况下，只要它们属于同一密度组，则相似的实例可能相距甚远。流行的聚类算法包括K-Means、DBSCAN、聚集聚类、BIRCH、均值平移、亲和度传播和光谱聚类。