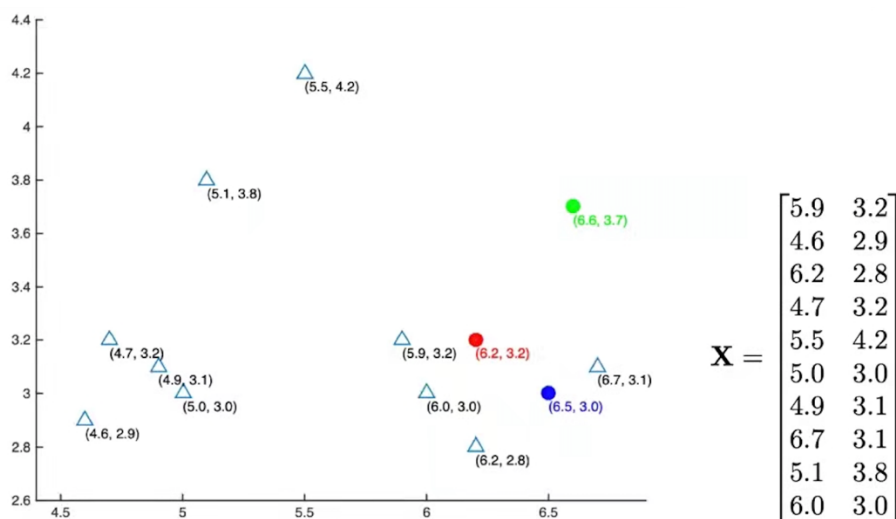


无监督学习

[简答题]

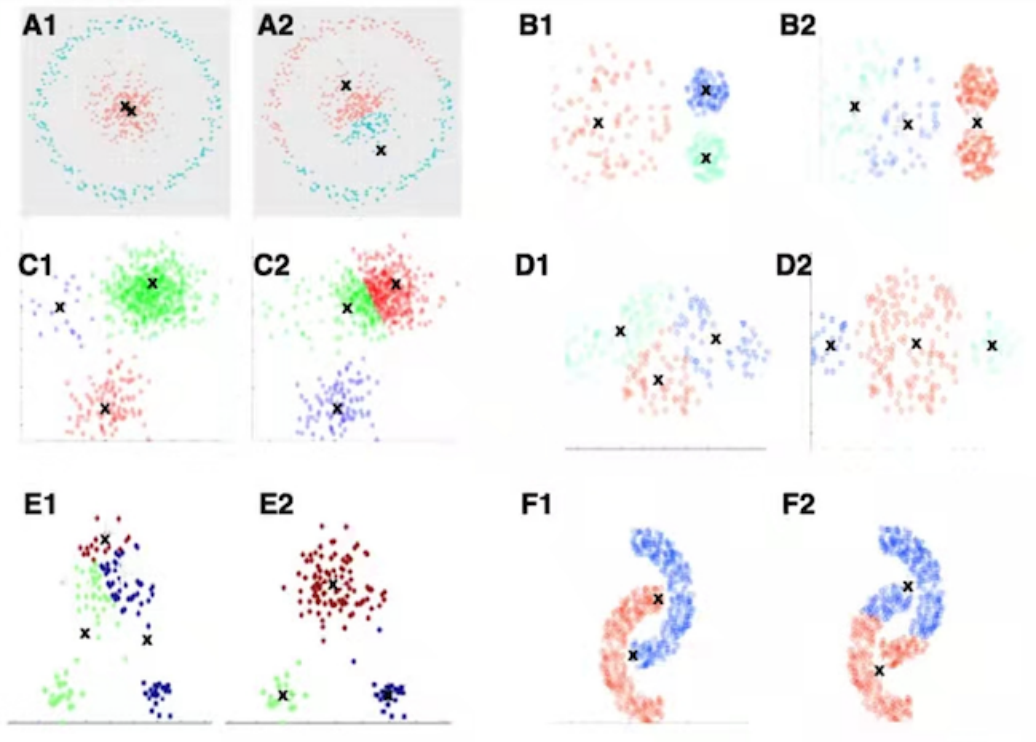
1	K 均值聚类算法隐含了对簇的什么假设？
2	K 均值迭代一定会收敛吗？
3	K 均值算法的初始化参数怎么选择？
4	K 均值算法的超参 K 怎么选择？
5	K 均值算法有什么局限性？可能用什么方法改进？
6	EM 算法为什么要引入隐变量 z？
7	EM 算法在 GMM 上的应用，是用哪些参数来体现不同簇的重要性和簇的大小密度等特性的？
8	K 均值和 EM 算法有什么共通之处吗？
9	如何定义簇的相似性？
10	层次聚类的限制有哪些？
11	KNN 算法和 Kmeans 算法的异同点是什么？
12	请简述 PCA 流程。

13. [计算题] K 均值聚类的 3 个初始聚类中心如图红、绿、蓝三个点所示。给出第一次迭代后属于第一个簇的样本，以及更新后第一个簇的中心坐标（保留 2 位小数）。



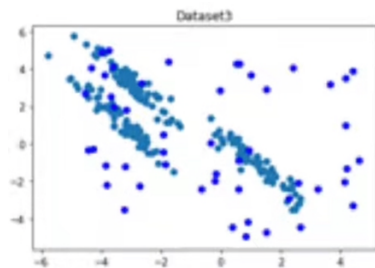
14. [填空题]以下 ABCDEF 六种数据分布，1 和 2 哪种更可能是用 K 均值聚类算法获得的结果？

A () , B () , C () , D () , E () , F ()



中国科学院大学

15 [单选题]以下数据分布（浅蓝色是待分类样本，深蓝色是噪声点），用哪种聚类算法最合适？



- A. K 均值
- B. GMM
- C. DBSCAN

16 [单选题]以下哪类降维方法不需要构造映射函数？

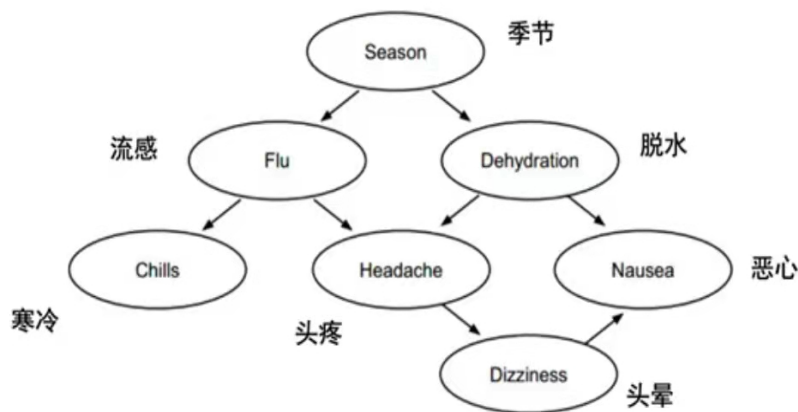
- A. PCA
- B. KPCA
- C. MDS

半监督学习

1. [单选题]TSVM 最直观地体现了什么假设?
 - A. 低密度分割假设
 - B. 流行假设
 - C. 平滑假设
2. [单选题]基于图的半监督学习最直观地体现了什么假设?
 - A. 低密度分割假设
 - B. 流行假设
 - C. 独立性假设
3. [单选题]多视角学习最直观地体现了什么假设?
 - A. 平滑假设
 - B. 流行假设
 - C. 独立性假设
4. [多选题]对一幅图像样本进行少量扰动（如旋转 5 度），不会改变该图像的类别标签。这与下列哪些假设相符合？
 - A. 聚类假设
 - B. 平滑假设
 - C. 流行假设
5. [多选题]KPCA 体现了（ ）与（ ）的辩证统一。
 - A. 升维、降维
 - B. 非线性、线性
 - C. 非监督、监督
6. [判断题]对比有限个标注样本的 SVM，考虑更多无标签样本的 TSVM 得到的判别间隔一定会更大。
7. [判断题]梯度下降有时会陷入局部极小值，EM 算法也会。
8. [判断题]增加非标注数据，一定会提高分类器的性能。
9. [判断题]增加有标注数据，一定会提高分类器的性能。
10. [判断题]不同训练样本的价值不同，即使是采用同样数量的标注样本，通过选取更有价值的训练样本，可能提高分类器的性能。
11. [判断题]表征样本的特征维度越高，样本之间的区分性越强。
12. [判断题]非线性划分的样本只有映射到更高维度特征空间才可能转为线性划分。

概率图模型

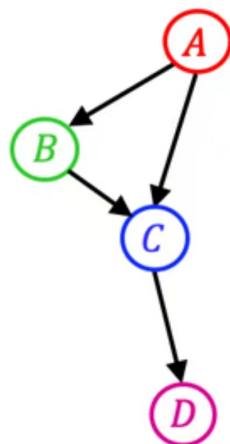
1. [单选题]现用一个 HMM 模型实现英文的词性标注系统，系统中共有 10 种词性，则状态转移矩阵的大小为 ()
A. 10
B. 20
C. 100
2. [单选题]下图所示的贝叶斯网包含 7 个变量，那么相互（条件）独立的变量是 ()



- A. 寒冷 \perp 季节 | 流感
- B. 寒冷 \perp 季节 | 头痛
- C. 寒冷 \perp 季节

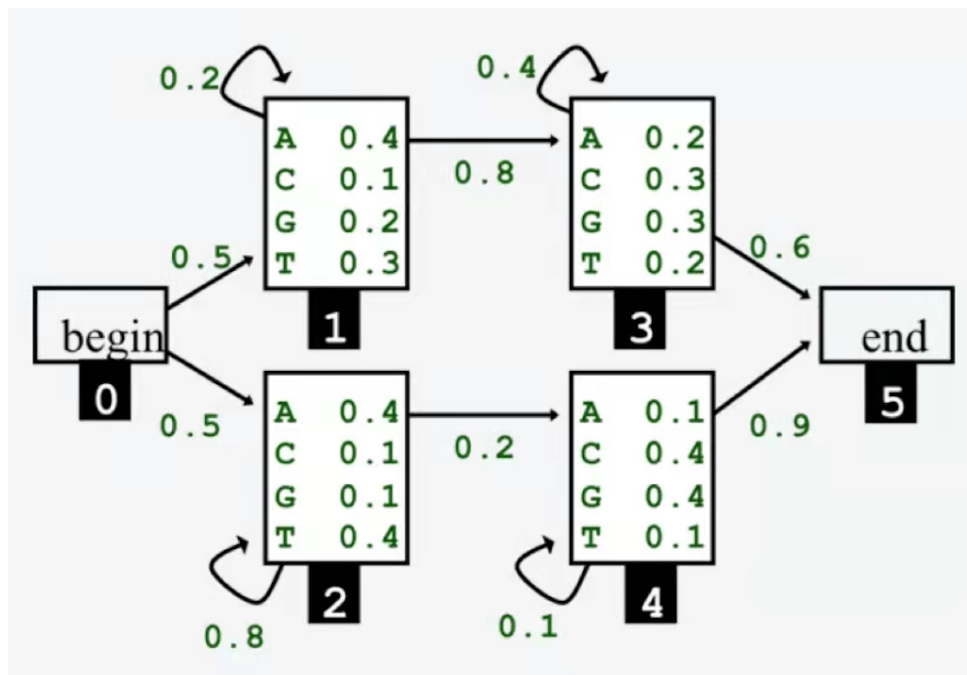
中国科学院大学
University of Chinese Academy of Sciences

3. [单选题]下图所示的贝叶斯网包含 4 个二值变量，那么该网络一共有 () 个参数。
A. 4
B. 8
C. 9





- D. 16

4. [计算题] 假设有 3 个盒子，分别装有不同数量的苹果 (A) 和桔子 (O)：
- 盒子一：2 个 A, 2 个 O；
 盒子二：3 个 A, 1 个 O；
 盒子三：1 个 A, 3 个 O；
- 每次随机选择一个盒子并从中抽取一个水果，观测并记录看到的水果是哪种。不幸的是，忘记去记录所选的盒子号码，只记录了每次看到的水果是 A 还是 O。
- (1) 请用 HMM 模型描述上述过程。
 (2) 假如观测到水果序列为 $x = \{A, A, O, O, O\}$ ，请给出最佳的盒子序列。
5. [计算题] 下图所示的 HMM 模型中：
- (1) 采用前向算法计算序列 “AGTT” 出现的概率。
 (2) 计算 “TATA” 最可能出现的状态序列。



6. [简答题] 简述 HMM 和 CRF 的区别。

集成学习

1. [单选题]随机森林中选择的决策树是 ()
 - A. 性能较好、深度较深的树
 - B. 性能较弱、深度较浅的树
2. [单选题]基于树的 Boosting 中选择的决策树是 ()
 - A. 性能较好、深度较深的树
 - B. 性能较弱、深度较浅的树
3. [单选题]Bagging 可降低模型的 ()
 - A. 偏差
 - B. 方差
4. [单选题]Boosting 可降低模型的 ()
 - A. 偏差
 - B. 方差
5. [单选题]假如对树采用 bagging 方式进行集成学习, 可采用 () 对决策树的超参数 (例如树的最大深度) 进行调优。
 - A. 交叉验证
 - B. 包外估计
6. [多选题]下列哪些操作可望减轻过拟合现象?

 - A. 增加正则项的权重
 - B. 如果是决策树模型, 减少树的最大深度;
 - C. 如果是线性模型, 换成更复杂的非线性模型;
 - D. 如果是多层神经网络模型, 加大模型的训练迭代次数
7. [简答题]简述 Adaboosting 流程。
8. [推算题]给定以下数据, 请写出 Adaboosting 流程。

序号	1	2	3	4	5	6	7	8	9	10
x	0	1	2	3	4	5	6	7	8	9
y	1	1	1	-1	-1	-1	1	1	1	-1