数据建模与分析 2020-2021 春夏 回忆卷

作者：@KaiKai123 @晕乎乎D

参考：@3180104673 https://www.cc98.org/topic/5116266/

小标题中标明的例题/习题，代表与考试题相似/相同的例题

1. 平衡 kd 树 - 例 3.2

原题数据忘记了，对应的例题：已知数据集 $T=\{(2,3)',(5,4)',(9,6)',(4,7)',(8,1)',(7,2)'\}$

1. 画出对应的平衡 kd 树。
2. 已知点 $(2,5)'$，求在欧式距离下，数据集中离它最近的点。

2. 熵

设离散型随机变量 X，概率分布为 $P(X=a_i)=p_i, i=1,2,\cdots,n$。

1. 写出 X 的熵 $H(p)$ 的定义，它的意义是什么。
2. 证明 $0\le H(p)\le \log(n)$。

   （Jensen 不等式，$f(\sum\lambda_i x_i)\ge\sum\lambda_i f(x_i)$，取 $f(x)=\log x, \lambda_i=p_i, x_i=1/p_i$ 即得证）

3. 朴素贝叶斯 - 例 4.1

已知训练数据集如表所示，使用训练数据集学习一个朴素贝叶斯分类器，并确定 $(1,R)'$ 的类标记。表中 $X^{(1)}$，$X^{(2)}$ 为特征，取值的集合分别为 $A_1=\{1,2,3\}$，$A_2=\{R,S,T\}$。Y 为类标记，$Y\in\{1,-1\}$。

$X^{(1)}$	1	1	1	1	2	2	2	3	3	3
$X^{(2)}$	R	R	S	T	S	R	S	R	T	S
Y	1	-1	-1	-1	1	1	-1	1	-1	-1

4. SVM

对于一个线性可分数据集，使用线性可分支持向量机的学习的最优化问题为

$$\begin{aligned}\min_{\omega,b}\quad & \frac{1}{2}||\omega||^2\\ \text{s.t.}\quad & y_i(\omega\cdot x_i+b)\ge 1, i=1,2,\cdots N\end{aligned}$$

1. 什么是线性支持向量机，与感知机的区别。
2. 写出题目所述最优化问题的对偶问题。
3. 通过对偶问题的解，写出原问题的解。

5. 层次聚类 - 例14.1

给定5个样本的集合，样本之间的欧氏距离如下矩阵 D 所示：

$$D=[d_{ij}]_{5\times5}=\begin{bmatrix}0&7&2&9&3\\7&0&5&4&6\\2&5&0&8&1\\9&4&8&0&5\\3&6&1&5&0\end{bmatrix}$$

其中 d_{ij} 表示第 i 个样本和第 j 个样本之间的欧式距离。

使用聚合层次聚类法对这5个样本进行聚类。

6. Markov 链 - 例19.3

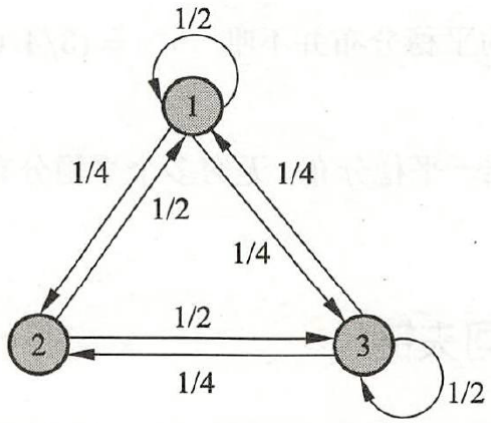

图 19.4　马尔可夫链例

对如图所示的 Markov 链

1. 求转移概率矩阵
2. 求平稳分布

7. SVD

设 $m \times n$ 的矩阵 A（$m > n$）。

1. 是否对于任意矩阵 A，均存在奇异值分解?
2. 写出对 A 进行奇异值分解的过程
3. 奇异值分解是否唯一?

8. 主成分分析 - 习题 16.1

对于以下样本数据，进行主成分分析。

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

9. 决策树

1. 对于特征 A，它对应的信息增益 $g(D, A)$ 对于构建决策树有哪些用处?
2. 为什么需要剪枝?

10. 感受与建议

对于这门课程的学习感受、建议