《模式识别与机器学习》第3次习题课

助教：许修为、黄原辉

2024年01月07日

第三次习题课：第十章

选择题

- C：自底向上和自顶向下都可以
- D：kmeans算法的复杂度为O(Nkm)

1. （**多选题**）以下关于聚类的说法，正确的是（ABD ）

   A. 聚类是一种无监督学习方法

   B. EM 算法的 M 步可由极大似然估计推导得到

   C. 层次聚类只能使用自底向上的聚合策略

   D. K 均值聚类算法的时间复杂度与聚类样本的数量成线性关系

第三次习题课：第十章

选择题

- A：层次化聚类不需要
- C：聚类目标是类内距离尽可能小，类间尽可能大

2. （**多选题**）关于数据聚类方法，以下说法正确的是（BD ）

A. K 均值聚类、高斯混合聚类和层次化聚类方法都需要预先给定类簇个数

B. K 均值聚类易受初始均值向量选取的影响

C. 设S_w是类内离散度矩阵，S_B是类间离散度矩阵，则聚类目标可以是最大化$|S_wS_B^{-1}|$

D. 高斯混合聚类问题可以使用 EM 算法来求解

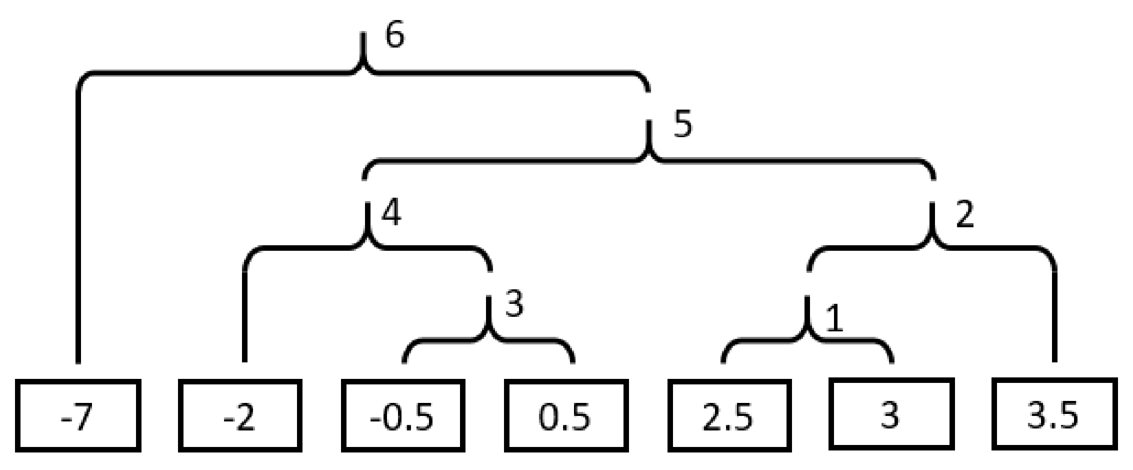
第三次习题课：第十章

计算题

- **注意：**1、2两簇可以合并

1. **（层次聚类）**现有 7 个一维数据点$\{-7, -2, -0.5, 0.5, 2.5, 3, 3.5\}$，采用自底向上的层次聚类方法对其进行聚类，其中类簇D_i和D_j之间的距离定义为下式，并画出聚类树。

$$d_{min}(D_i, D_j) = \min_{\substack{x \in D_i \\ x' \in D_j}} \|x - x'\|_2$$



清華大學

第十一章作业

一、选择题

1. （单选题）以下方法不属于子空间学习的数据降维方法是（ C ）↵

   A. 主成分分析 PCA↵

   B. 线性判别分析 LDA↵

   C. 拉普拉斯特征映射 LE↵ 流形学习

   D. 局部保持投影 LPP↵

2. （多选题）以下关于降维的说法，正确的是（ AB ）↵

   A. 主成分分析得到的子空间同时满足最近重构性和最大可分性↵

   B. 主成分分析既可通过特征值分解求解，也可以通过奇异值分解求解↵

   C. 多维尺度变换 MDS 需要输入原始数据和其距离矩阵↵ 只需要距离矩阵

   D. 等度量映射 Isomap 试图保持样本在局部的线性关系↵ 保持样本之间的测地线距离

第十一章作业

二、计算题

1. （主成分分析）现有 7 个二维数据点：

$$\{(-7,3.4),(-2,3.4),(-0.5,2),(0.5,2),(2.5,3.4),(3,3.4),(3.5,3.4)\}$$

使用主成分分析（PCA）将上述 7 个二维数据点降维至一维。

1）计算均值并中心化

均值为(0,3)，中心化之后的样本为

$$\{(-7,0.4),(-2,0.4),(-0.5,-1),(0.5,-1),(2.5,0.4),(3,0.4),(3.5,0.4)\}$$

2）计算协方差矩阵

$$\begin{bmatrix} -7 & -2 & -0.5 & 0.5 & 2.5 & 3 & 3.5 \\ 0.4 & 0.4 & -1 & -1 & 0.4 & 0.4 & 0.4 \end{bmatrix} \begin{bmatrix} -7 & 0.4 \\ -2 & 0.4 \\ -0.5 & -1 \\ 0.5 & -1 \\ 2.5 & 0.4 \\ 3 & 0.4 \\ 3.5 & 0.4 \end{bmatrix} = \begin{bmatrix} 81 & 0 \\ 0 & 2.8 \end{bmatrix}$$

3）对协方差矩阵做特征值分解

$$\begin{bmatrix} 81 & 0 \\ 0 & 2.8 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 81 & 0 \\ 0 & 2.8 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^T$$

4）取 81 对应的特征向量构成投影矩阵

$$W = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

输入: 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
　　低维空间维数 d'.

过程:

1: 对所有样本进行中心化: $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i$;
2: 计算样本的协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$;
3: 对协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 做特征值分解;
4: 取最大的 d' 个特征值所对应的特征向量 $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'}$.

输出: 投影矩阵 $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'})$.

第三次习题课：第十二章

选择题

- C：由于在特征选择过程中需多次训练学习器，因此包裹式选择的计算开销通常比过滤式选择大得多

1. **（多选题）**关于特征选择方法，以下说法正确的是（ABD）

   A. 数据降维和特征选择都可以应对维度爆炸问题

   B. 前向搜索、后向搜索和双向搜索都是贪心策略

   C. 过滤式特征选择的计算开销通常比包裹式特征选择更大

   D. 信息熵、类内类间距离、随机变量间的相关性都可以作为特征子集评价指标

第三次习题课：第十二章

选择题

- A：贪心算法，容易陷入局部最优
- B：过于绝对，特征相关性难以完全消除

2. **（多选题）**以下关于特征选择的说法，正确的是（CD）

   A. 一般而言，双向搜索策略能得到最优特征子集

   B. 特征选择不仅可以降低特征空间的维度，还可以消除特征之间的相关性

   C. 可以将分类器错误率作为设定特征评价准则的依据

   D. 特征数量太多会影响模型参数估计的稳定性

第三次习题课：第十二章

计算题

- 注意类内类间距离的物理含义

1. **（特征子集搜索与评价）**现有 7 个二维特征向量：

$$\{(-7,3.4),(-2,3.4),(-0.5,2),(0.5,2),(2.5,3.4),(3,3.4),(3.5,3.4)\},$$

且对应的标签分别为{0,0,1,0,1,1,1}。记每个特征向量的两维特征分别为 x 特征和 y 特征，请使用**基于类内类间距离的判据**作为评价指标，判断 x 特征和 y 特征的优劣。

解答：对于 x 特征：$\mu_1=-2.83$，$\mu_2=2.13$，$S_1^2=29.17$，$S_2^2=9.69$

因此 $J_F(x)=\frac{(\mu_1-\mu_2)^2}{S_1^2+S_2^2}=0.63$

对于 y 特征：$\mu_1=2.93$，$\mu_2=3.05$，$S_1^2=1.307$，$S_2^2=1.47$

因此 $J_F(y)=\frac{(\mu_1-\mu_2)^2}{S_1^2+S_2^2}=0.0052$

可见 x 特征更加分散，更具有判别能力。

第十三章作业

一、选择题

1. （单选题）设X是一个非空集合，在下列哪种距离定义下，X不是一个度量空间（ C ）

   A. 欧氏距离

   B. 切比雪夫距离

   C. 余弦距离 非负性×、对称性、三角不等式×

   D. 曼哈顿距离

2. （多选题）以下关于度量学习的说法，正确的是（ CD ）

   A. 余弦距离具有平移不变性和尺度缩放不变性 平移不变性×

   B. 计算三元组损失函数时，为了提升训练效率，选择的三元组应该越难越好

   C. 度量学习也可以作为数据降维的方法

   D. 度量学习中，马氏距离的度量矩阵是半正定对称阵

- 三元组的选择策略对损失函数的优化性质有很大影响
- 选取的三元组本身满足margin约束，则对梯度没有贡献，收敛会很慢
- 选取的三元组太难，比如负样本距离anchor样本过近，则可能导致训练不稳定甚至发散

第十三章作业

二、简答题

1. 简述欧氏距离与余弦距离的区别，并举例说明各自的应用场景。↵ 参考课件第10，14页

余弦距离注重样本在向量方向上的差异，而非距离长度大小

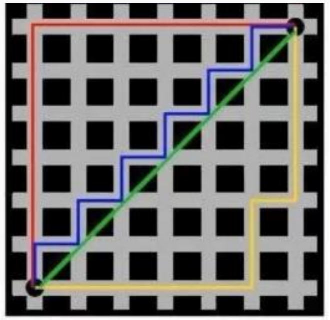
例如，统计两部剧的用户观看行为，用户A的观看向量为(0,1)，用户B为(1,0)；此时二者的余弦距很大，而欧氏距离很小；我们分析两个用户对于不同视频的偏好，更关注相对差异，显然应当使用余弦距离。说明两个用户的观看行为存在很大的差异

而当我们分析用户活跃度，以登陆次数(单位：次)和平均观看时长(单：分钟)作为特征时，余弦距离会认为(1,10)、(10,100)两个用户距离很近；但显然这两个用户活跃度是有着极大差异的，此时我们更关注数值绝对差异，应当使用欧氏距离

2. 分析为什么在 k-NN 算法中常采用欧氏距离，而不采用曼哈顿距离。↵ 参考课件第10，11页

欧式距离更关注于点在空间的相对距离，而曼哈顿距离择关注于点如何通过坐标轴上的方向到达另一点。

而在计算距离时我们更关注于样本之间的绝对差异，而不考虑样本之间的投影差异，所以采用欧式距离。



第三次习题课：第十四章

选择题

- B：参考主动学习定义，需要对未标记样本进行预测和筛选
- C：直推式半监督算法只能处理当前的无标签样本
- D：EM算法对初始值敏感

1. **（多选题）**下列关于半监督学习的说法正确的是（AC）

   A. 在半监督学习中，未标记样本和有标记样本需要是独立同分布的

   B. 主动学习只是依赖人工标注，没有对未标记样本的数据特点加以利用

   C. 直推学习可能无法对新样本进行预测

   D. 求解生成式半监督学习的 EM 算法的结果不依赖于初始值

第三次习题课：第十四章

选择题

- B：需要假设样本分布

2. **（单选题）**以下关于半监督学习的说法错误的是（B）

   A. 生成式半监督学习方法可基于高斯混合模型，并使用 EM 算法求解

   B. 生成式半监督学习方法对先验知识的要求较弱

   C. 聚类假设和流形假设是利用未标记样本的两种常见要求

   D. 协同训练假设数据拥有两个及以上充分且条件独立的视图

第三次习题课：第十四章

计算题

1. **（半监督图学习）**现有 3 个有标记样本$\{(\boldsymbol{x},y)\}$如下：

$$\{((-1,-1),-1),((-1,1),-1),((1,0),1)\}$$

其中$\boldsymbol{x}$为二维向量，$y\in\{-1,1\}$。试利用**半监督图学习**判断未标记样本$\{(-1,0),(2,0)\}$的标签，其中亲和矩阵基于 k 近邻（k 取 2）构建，即：

$$(\boldsymbol{W})_{ij}=\begin{cases}1, & \text{如果}\boldsymbol{x}_i\text{是}\boldsymbol{x}_j\text{的 k 近邻，或}\boldsymbol{x}_j\text{是}\boldsymbol{x}_i\text{的 k 近邻}\\ 0, & \text{否则}\end{cases}$$

解答：由条件知：

$$W=\begin{bmatrix}0&1&0&1&0\\1&0&0&1&0\\0&0&0&1&1\\1&1&1&0&1\\0&0&1&1&0\end{bmatrix}$$

根据 D 矩阵定义：$D=diag(2,2,2,4,2)$

有$D_{uu}=diag(4,2)$，$W_{uu}=\begin{bmatrix}0&1\\1&0\end{bmatrix}$，$W_{ul}=\begin{bmatrix}1&1&1\\0&0&1\end{bmatrix}$

可得$P_{uu}=D_{uu}^{-1}W_{uu}=\begin{bmatrix}0&\frac{1}{4}\\\frac{1}{2}&0\end{bmatrix}$，$P_{ul}=D_{uu}^{-1}W_{ul}=\begin{bmatrix}\frac{1}{4}&\frac{1}{4}&\frac{1}{4}\\0&0&\frac{1}{2}\end{bmatrix}$

由$f_l=\begin{bmatrix}-1\\-1\\1\end{bmatrix}$，知$f_u=(I-P_{uu})^{-1}P_{ul}f_l=\frac{1}{7}\begin{bmatrix}-1\\3\end{bmatrix}$，因此标签为-1 和 1