第八章 特征选择与学习

一、填空题

1 数据的均值、方差、直方图等特征属于（通用）特征；工业生产中的温度、压力、速度等特征属于（结构化）特征；

2 比较常见的线性子空间特征学习算法包括（主成分分析/主元分析）、（独立成分分析/独立元分析）和（线性判别分析）。（请用中文回答）

3 比较常见的特征搜索策略包括（前向搜索）和（后向搜索）

4 MDS 和 IsoMap 的相同之处是（降维前后样本间距离保持不变），不同之处是前者使用（欧式）距离，后者使用（测地线）距离

5 10 维空间中的 5 个样本，最多需要（4）维的子空间可以保证投影损失为 0。

6 为了保证编码的稀疏性，标准的稀疏编码算法使用（编码向量的 L1 范数）作为约束条件。

二、判断题

7 表情特征可以用于身份识别。（×）

8 稀疏编码方法中的编码数值是通过将原始数据向字典上投影得到的。（×）

9 IsoMap 流形学习算法无法直接实现训练集以外的新样本的特征提取。（√）

10 稀疏编码算法中的重构损失和稀疏性均可作为目标函数或约束条件。（√）

11 稀疏编码算法的字典中的特征向量必须正交。（×）

12 PCA 算法中的基向量必须正交。（√）

三、简答题

13 请简要列举 3 种可以用于身份识别的生物特征，并分析其特点

答：

指纹特征：准确性高、识别成本低、采集成本中等

人脸特征：准确性较高、识别成本中等、采集成本低

基因特征：准确性极高、识别成本高、采集成本高

14 给出 PCA 中方差最大化和投影损失最大化两种思路在第一个基向量上求取最优解的等价性证明。

投影方差的目标函数

$$J_1(\boldsymbol{b}_1) = E\{y_1^2\} = E\{(\boldsymbol{b}_1^T\boldsymbol{x})^2\} = \boldsymbol{b}_1^T E\{\boldsymbol{x}\boldsymbol{x}^T\}\boldsymbol{b}_1 \approx \boldsymbol{b}_1^T C_X \boldsymbol{b}_1$$

投影损失的目标函数

$$\begin{aligned} J_2(\boldsymbol{b}_1) &= E\{\|\boldsymbol{x} - y_1\boldsymbol{b}_1\|^2\} = E\{(\boldsymbol{x} - (\boldsymbol{b}_1^T\boldsymbol{x})\boldsymbol{b}_1)^T(\boldsymbol{x} - (\boldsymbol{b}_1^T\boldsymbol{x})\boldsymbol{b}_1)\} \\ &= E\{\boldsymbol{x}^T\boldsymbol{x}\} - E\{2(\boldsymbol{b}_1^T\boldsymbol{x})^T(\boldsymbol{b}_1^T\boldsymbol{x})\} + E\{\boldsymbol{b}_1^T(\boldsymbol{b}_1^T\boldsymbol{x})^T(\boldsymbol{b}_1^T\boldsymbol{x})\boldsymbol{b}_1\} \\ &= E\{\boldsymbol{x}^T\boldsymbol{x}\} - E\{(\boldsymbol{b}_1^T\boldsymbol{x})^2\} \approx C_X - \boldsymbol{b}_1^T C_X \boldsymbol{b}_1 \end{aligned}$$

由于C_X是一个常数，因此:

$$\boldsymbol{b}_1^* = \underset{\boldsymbol{b}_1}{\operatorname{argmax}} J_1(\boldsymbol{b}_1) = \underset{\boldsymbol{b}_1}{\operatorname{argmin}} J_2(\boldsymbol{b}_1)\ s.t.\ \|\boldsymbol{b}_1\| = 1$$

说明两者在具有完全相同的第一个基向量$\boldsymbol{b}_1$的最优解。