《大数据与人工智能方向基础》课程复习

**PART6** 非监督式特征提取

**2020‐05‐25** 要点：

1. 特征工程；特征降维；特征选择、特征提取
2. 监督式、非监督式特征提取
3. 线性、非线性特征提取
4. **PCA**的全称？**PCA**有哪些应用？
5. 理解样本协方差矩阵的本征值与本征列向量的意义；各主成分的分布方差？
6. 掌握利用**PCA**进行特征提取、特征降维的基本实现过程。
   1. 针对特征空间任意观测样本，如何提取指定的主成分？第**1**主成分？第**2** 主成分？**…**
   2. 如何根据累积方差解释比确定主成分数目 **(**即：新的特征空间的特征维数**)**？

练习：

1 给定观测样本集***D x*** *i*,i 1,,*N*，其中***x****i* *R*3 .请结合该样本集，设计一个基于主成分分析的特征降维方法，以便基于该算法，提取原始空间任意观测样本 ***x****R*3的第1、第2主成分. 参考答案：

|  |
| --- |
| **step1.** 基于样本集***D***，估计**样本中心**及**协方差矩阵**.  ***N N T***  **\***  ***i***1 ***xi***  **\* *N***1 ***i***1 ***xi*** **\******xi*** **\*** 1  ***N***  **step2.** 确定 的 ***p***  3个**本征值**及**本征向量**. 得 个本征值***p***   1   2 3 **0**  对应本征向量 ***a ii*** , 1,2,3  **step3.** 确定32的变换矩阵***A*2** ***a***1 ***a***2  **step4.** 对于任意观测 ，提取该样本的前两个主成***x*** 分： **1** **2*****T* =*A***2***T*** ***x*** **\*** |

**1** / **2**

《大数据与人工智能方向基础》课程复习

**2.**给定数据集𝑫 𝒙𝒊, 𝑖 1, … , 𝑚 ，其中𝒙𝒊 ∈ 𝑹𝒅。请结合该样本集D，设计一个基于PCA的特征降维算法，以便基于该方法将任意观测𝒙 ∈ 𝑹𝒅的降至r维**.** 请详细给出有关步骤和必要表达式. 参考答案：

|  |
| --- |
| **step1.** 基于样本集 ，估计***D* 样本中心**及**协方差矩阵**.  ***N N T***  **\***  1 ***xi***  **\* *N***1 ***i***1 ***xi*** **\******xi*** **\*** 1  ***N i***  **step2.** 确定 的前**\* *r r***  ***d***个最大**本征值**及**本征向量**. 得前 个本征值***r***  1   2  **r**  对应本征向量 ***a ii*** , 1,,***r***  **step3.** 确定d 的变换矩阵***r Ar*** ***a***1 ***a***2  ***ar***   **step4.** 对于任意观测 ，其在新***x*** 的r维空间的映射位置： ***r* =*ArT*** ***x*** **\*** |

**3.**给定数据集𝑫 𝒙𝒊, 𝑖 1, … , 𝑚 ，其中𝒙𝒊 ∈ 𝑹𝒅。请结合该样本集D，设计一个基于主成分分析的特征降维方法，并且使用新的特征描述样本时，满足累积方差解释比不低于𝛼 0.9，请确定新的特征空间特征维数r，并将任意观测𝒙 ∈ 𝑹𝒅降至

r维。请详细给出有关步骤和必要表达式.

**4.**给定鸢尾花数据集𝑫 𝒙𝒊, 𝑖 1, … , 𝑚 ，其中𝒙𝒊 ∈ 𝑹𝟒。请结合该样本集D，基于主成分分析法实现上述数据集在新的二维空间可视化。请详细给出有关实现步骤和必要表达式.

**2** / **2**