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 $\boldsymbol{D}=\{\boldsymbol{x}_i, i=1,\cdots,N\}$，其中 $\boldsymbol{x}_i \in R^3$.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 $\boldsymbol{x} \in R^3$ 的第1、第2主成分.

参考答案：

step1. 基于样本集$\boldsymbol{D}$，估计**样本中心**μ及**协方差矩阵**Σ.

$$\mu^* = \frac{1}{\boldsymbol{N}}\sum_{i=1}^{N}\boldsymbol{x}_i \qquad \Sigma^* = \frac{1}{\boldsymbol{N}}\sum_{i=1}^{N}\left(\boldsymbol{x}_i-\mu^*\right)\left(\boldsymbol{x}_i-\mu^*\right)^T$$

step2. 确定Σ的$\boldsymbol{p}=3$个**本征值**及**本征向量**.

得 $\boldsymbol{p}$个本征值 $\lambda_1 \ge \lambda_2 \ge \lambda_3 \ge \boldsymbol{0}$

对应本征向量 $\boldsymbol{a}_i, \boldsymbol{i}=1,2,3$

step3. 确定3×2的变换矩阵$\boldsymbol{A}_2=[\boldsymbol{a}_1 \quad \boldsymbol{a}_2]$

step4. 对于任意观测$\boldsymbol{x}$，提取该样本的前两个主成分： $[\xi_1 \quad \xi_2]^T = \boldsymbol{A}_2^T\left(\boldsymbol{x}-\mu^*\right)$

2.给定数据集$\boldsymbol{D}=\{\boldsymbol{x_i},i=1,\dots,m\}$，其中$\boldsymbol{x_i}\in\boldsymbol{R^d}$。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$\boldsymbol{x}\in\boldsymbol{R^d}$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$\boldsymbol{D}$，估计**样本中心**μ及**协方差矩阵**Σ.

$$\mu^*=\frac{1}{N}\sum_{i=1}^{N}x_i \qquad \Sigma^*=\frac{1}{N}\sum_{i=1}^{N}\left(x_i-\mu^*\right)\left(x_i-\mu^*\right)^T$$

step2. 确定Σ^*的前$\boldsymbol{r}\left(\boldsymbol{r}<\boldsymbol{d}\right)$个最大**本征值**及**本征向量**.

得前$\boldsymbol{r}$个本征值 $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_r$

对应本征向量 $\boldsymbol{a}_i,\boldsymbol{i}=1,\cdots,\boldsymbol{r}$

step3. 确定$\mathrm{d}\times\boldsymbol{r}$的变换矩阵$\boldsymbol{A}_r=\left[\boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \cdots \quad \boldsymbol{a}_r\right]$

step4. 对于任意观测$\boldsymbol{x}$，其在新的r维空间的映射位置： $\xi_r=\boldsymbol{A}_r{}^T\left(\boldsymbol{x}-\mu^*\right)$

3.给定数据集$\boldsymbol{D}=\{\boldsymbol{x_i},i=1,\dots,m\}$，其中$\boldsymbol{x_i}\in\boldsymbol{R^d}$。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$\alpha=0.9$，请确定新的特征空间特征维数r，并将任意观测$\boldsymbol{x}\in\boldsymbol{R^d}$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$\boldsymbol{D}=\{\boldsymbol{x_i},i=1,\dots,m\}$，其中$\boldsymbol{x_i}\in\boldsymbol{R^4}$。**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.