PART6 非监督式特征提取与低维可视化

2023-06

掌握：

1. 理解什么是特征提取？什么是非监督式特征提取？什么是线性/非线性特征提取？
2. 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？
3. **PCA**的全称？
4. **掌握利用PCA进行特征提取或特征降维的基本实现过程。**

   如何根据累积方差解释比确定特征提取的数目？

   如何用PCA实现样本数据的低维可视化？
5. 能使用**t-SNE**进行数据的低维可视化。

1 给定观测样本集 $\boldsymbol{D}=\left\{\boldsymbol{x}_i, \mathrm{i}=1,\cdots,N\right\}$，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.

解：

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

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

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

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

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

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

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

注意：

观测$\boldsymbol{x}$第1主成分：$\xi_1=\boldsymbol{a}_1{}^{\boldsymbol{T}}\left(\boldsymbol{x}-\hat{\mu}\right)$

观测$\boldsymbol{x}$第2主成分：$\xi_2=\boldsymbol{a}_2{}^{\boldsymbol{T}}\left(\boldsymbol{x}-\hat{\mu}\right)$

其中：第1主成分

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，设计一个基于主成分分析法的特征降维算法，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.

解：

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

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

step2. 确定$\hat{\Sigma}$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 $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_{\mathrm{r}}$

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

step3. 确定$\mathrm{d}\times\boldsymbol{r}$的变换矩阵$\boldsymbol{A}_{\boldsymbol{r}}=\begin{bmatrix}\boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_{\boldsymbol{r}}\end{bmatrix}$

step4. 对于任意观测$\boldsymbol{x}$，提取该样本r维新的特征向量： $\xi_{\mathrm{r}}=\boldsymbol{A}_{\boldsymbol{r}}^{\boldsymbol{T}}\left(\boldsymbol{x}-\hat{\mu}\right)$

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法的特征降维算法，并满足累积方差解释比不低于0.9，请确定新的特征空间特征维数r. 请详细给出有关步骤和必要表达式.