第十三讲 降维初步：主成分分析

牟克典

2021年6月9日

概要

1 预备知识

主成分分析

- 如果数据的一些特征之间存在相关性, 处理起来不太方便;
- 如果数据维数过高, 影响算法性能.

我们希望能构造一组新的相互不相关的特征来表示数据:

- 通常用原来特征的线性组合来构造新特征.
- 希望特征变换的过程中损失的信息尽可能少.
- 构造出的新特征个数比原来的特征数少很多, 达到降维的目的.

标准线性组合

设$\mathbf{x} = (x_1, x_2, \cdots, x_m)^T$是$m$维随机向量, $\alpha \in \mathbf{R}^m$ 且$\alpha^T\alpha = 1$, 则称

$$y = \alpha^T \mathbf{x}$$

为标准线性组合.

本讲主要考虑标准线性组合.

设$\mathbf{x} = (x_1, x_2, \cdots, x_m)^T$是$m$维随机向量，其均值为$\mu$，协方差矩阵

$$\Sigma = E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T] = [\sigma_{ij}]_{m \times m},$$

则

- $E[(\mathbf{x} - \mu)] = 0.$
- $\mathrm{trace}(\Sigma) = \sum_{i=1}^{m} \mathrm{Var}(x_i) = \sum_{i=1}^{m} \sigma_{ii}.$
- Σ是半正定的.
- 不妨设Σ的特征值按照降序排列为

$$\lambda_1 \geq \lambda_2 \cdots \geq \lambda_m \geq 0$$

关于$\lambda_m \geq 0$的证明:

- 由Σ是半正定的即可推出. 或者也可采用下面的证明方式:
- 不妨设α为Σ属于λ_m的特征向量, 则$\alpha^T \alpha > 0$.

$$\begin{aligned} & \alpha^T \Sigma \alpha \\ & \alpha^T E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T]\alpha \\ = & E[\alpha^T (\mathbf{x}-\mu)(\mathbf{x}-\mu)^T \alpha] \\ = & E[\Big((\alpha^T(\mathbf{x}-\mu)\Big)^2] \geq 0 \end{aligned}$$

$$\alpha^T \Sigma \alpha = \lambda_m \alpha^T \alpha \geq 0$$

因此 $\lambda_{m \geq 0}$.

- Σ可对角化: 即

$$A^T\Sigma A = \Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_m),$$

这里$A = [\alpha_1, \alpha_2, \cdots, \alpha_m]$为正交矩阵,
$\alpha_i = (\alpha_{1i}, \alpha_{2i}, \cdots, \alpha_{mi})^T$ 且α_i是Σ的属于λ_i的特征向量,
对$1 \leq i, j \leq m$而言,
    - $\alpha_i^T \alpha_i = 1$,
    - $\alpha_i^T \cdot \alpha_j^T = 0$, $i \neq j$,
    - $\Sigma\alpha_i = \lambda_i\alpha_i$,
    - $\alpha_i^T \Sigma = \lambda_i \alpha_i^T$.
- $\alpha_1, \alpha_2, \cdots, \alpha_m$正好构成了$\mathbf{R}^m$的一组标准正交基, 即

对$\forall\alpha \in \mathbf{R}^m$, $\exists c_1, c_2, \cdots, c_m \in \mathbf{R}$ s.t. $\alpha = \sum_{i=1}^{m} c_i\alpha_i$.

概要

1. 预备知识
2. 总体主成分分析
   - 主成分变换
   - 标准化随机变量
   - 方差贡献率
   - 因子负荷量

主成分变换

设$\mathbf{x} = (x_1, x_2, \cdots, x_m)^T$是均值为$\mu$, 协方差矩阵为$\Sigma$的$m$维随机向量, 则如下线性变换被称为主成分变换:

$$\mathbf{y} = A^T(\mathbf{x} - \mu).$$

并称$\mathbf{y}$的第i个分量

$$y_i = \alpha_i^T(\mathbf{x} - \mu)$$

为$\mathbf{x}$的第i主成分, 这里α_i为A的第i个列向量.

考虑$\mathbf{x} = (x_1, x_2)^T$, 其中

- $E[x_1] = E[x_2] = 0$,
- $\mathrm{Var}(x_1) = \mathrm{Var}(x_2) = 1$,
- $\mathrm{Cov}(x_1, x_2) = \rho > 0$,

协方差矩阵为

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

其特征值为

$$\lambda_1 = 1 + \rho, \quad \lambda_2 = 1 - \rho$$

相应的特征向量为

$$\alpha_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T, \quad \alpha_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)^T$$

由此得到**x**的第一、第二主成分分别为

$$y_1 = \frac{1}{\sqrt{2}}(x_1 + x_2);$$
$$y_2 = \frac{1}{\sqrt{2}}(x_1 - x_2)$$

相应的方差为

$$\mathrm{Var}(y_1) = 1 + \rho = \lambda_1;$$
$$\mathrm{Var}(y_2) = 1 - \rho = \lambda_2.$$
$$\mathrm{Var}(y_1) + \mathrm{Var}(y_2) = \mathrm{Var}(x_1) + \mathrm{Var}(x_2) = 2.$$

主成分的性质

TH1. 设$\mathbf{x} \sim (\mu, \Sigma)$， 则$\mathbf{y} = A^T(\mathbf{x} - \mu)$满足

(1) $E[\mathbf{y}] = \mathbf{0}$.

(2) $\mathrm{Var}(y_i) = \lambda_i$, $i = 1, 2, \cdots, m$.

(3) $\mathrm{Cov}(y_i, y_j) = 0$, $i \neq j$, $i, j = 1, 2, \cdots, m$..

(4) $\mathrm{Var}(y_1) \geq \mathrm{Var}(y_2) \geq \cdots \geq \mathrm{Var}(y_m) \geq 0$.

(5) $\sum_{i=1}^{m} \mathrm{Var}(y_i) = \mathrm{trace}(\Sigma) = \sum_{i=1}^{m} \mathrm{Var}(x_i)$.

(6) $\prod_{i=1}^{m} \mathrm{Var}(y_i) = |\Sigma|$.

(1) 的证明： $E[\mathbf{y}] = E[A^T(\mathbf{x}-\mu)] = A^T E[(\mathbf{x}-\mu)] = \mathbf{0}.$
(2)的证明：

$$\begin{aligned}\mathrm{Var}(y_i) = \mathrm{Var}(\alpha_i^T(\mathbf{x}-\mu)) &= \alpha_i^T E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T]\alpha_i \\ &= \alpha_i^T \Sigma \alpha_i = \lambda_i \alpha_i^T \alpha_i = \lambda_i\end{aligned}$$

(3)的证明：如果$i \neq j$,

$$\begin{aligned}\mathrm{Cov}(y_i, y_j) &= \mathrm{Cov}(\alpha_i^T(\mathbf{x}-\mu), \alpha_j^T(\mathbf{x}-\mu)) \\ &= \alpha_i^T E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T]\alpha_j \\ &= \alpha_i^T \Sigma \alpha_j = \lambda_j \alpha_i^T \alpha_j = 0.\end{aligned}$$

由(2)可知， (4), (5),(6)都成立. □

TH2. 不存在方差比λ_1更大的标准线性组合$y = \alpha^T \mathbf{x}$.

证明：考虑标准线性组合$y = \alpha^T \mathbf{x}$, 其中$\alpha \in \mathbf{R}^m$ 且$\alpha^T \alpha = 1$. 由于$\alpha_1, \alpha_2, \cdots, \alpha_m$正好构成了$\mathbf{R}^m$的一组标准正交基, 则

$$\exists c_1, c_2, \cdots, c_m \in \mathbf{R} \ s.t. \ \alpha = \sum_{i=1}^{m} c_i \alpha_i.$$

对此线性组合来说,

$$\begin{aligned} \mathrm{Var}(y) = \alpha^T \Sigma \alpha &= \left[\sum_{i=1}^{m} c_i \alpha_i^T\right] \Sigma \left[\sum_{i=1}^{m} c_i \alpha_i\right] \\ &= \sum_{i=1}^{m} c_i^2 \lambda_i \alpha_i^T \alpha_i = \sum_{i=1}^{m} c_i^2 \lambda_i. \end{aligned}$$

另一方面$\alpha^T\alpha = 1$意味着$\sum_{i=1}^{m} c_i^2 = 1$.

考虑如下最优化问题:

$$\max_{c_1,\cdots,c_m} \sum_{i=1}^{m} c_i^2 \lambda_i$$

$$s.t. \quad \sum_{i=1}^{m} c_i^2 = 1.$$

则此问题的最优解为

$$c_1 = 1, c_2 = \cdots = c_m = 0$$

此时

$$\lambda_1 = \max_{c_1,\cdots,c_m} \sum_{i=1}^{m} c_i^2 \lambda_i$$

对应的标准线性组合

$$y_1 = \alpha_1^T \mathbf{x}$$

正好是的第一主成分. □

TH3. 如果标准线性组合$y=\alpha^T\mathbf{x}$与$\mathbf{x}$的前k个主成分都不相关，则y的方差当y是第$k+1$主成分时达到最大.

证明：设$y=\alpha^T\mathbf{x}$, 其中$\alpha^T\alpha=1$ 且$\alpha=\sum\limits_{i=1}^{m}c_i\alpha_i$. 对此线性组合来说，

$$\mathrm{Var}(y)=\sum_{i=1}^{m}c_i^2\lambda_i.$$

对$1\le j<k$来说，

$$\begin{aligned}\mathrm{Cov}(y,y_j)&=\mathrm{Cov}(\alpha^T\mathbf{x},\alpha_j^T\mathbf{x})\\&=\left[\sum_{i=1}^{m}c_i\alpha_i^T\right]\Sigma\alpha_j\\&=c_j\lambda_j\alpha_j^T\alpha_j=c_j\lambda_j=0.\end{aligned}$$

这意味着对$1 \le j < k$来说，$c_j^2\lambda_j = 0$. 故

$$\mathrm{Var}(y) = \sum_{i=k+1}^{m} c_i^2\lambda_i.$$

和前面证明类似， 我们可得

$$\max_{c_1,\cdots,c_m} \mathrm{Var}(y) = \lambda_{k+1}.$$

对应的标准线性组合

$$y = \alpha_{k+1}^T \mathbf{x}$$

正好是的第$k+1$主成分. □

设$\mathbf{x} = (x_1, x_2, \cdots, x_m)^T$是均值为$\mu$, 协方差矩阵为$\Sigma$的$m$维随机向量, 对每个$x_i$定义

$$x'_i = \frac{x_i - \mu_i}{\sqrt{\sigma_{ii}}},$$

则

$$E[x'_i] = 0, \qquad \mathrm{Var}(x'_i) = 1.$$

对$\mathbf{x}' = (x'_1, x'_2, \cdots, x'_m)^T$来说,

$$\begin{aligned}\mathrm{Var}(x'_i, x'_j) &= E\left[(\frac{x_i - \mu_i}{\sqrt{\sigma_{ii}}})(\frac{x_j - \mu_j}{\sqrt{\sigma_{jj}}})\right] \\ &= \frac{E[(x_i - \mu_i)(x_j - \mu_j)]}{\sqrt{\sigma_{ii}\sigma_{jj}}} = \rho_{ij}\end{aligned}$$

即$\mathbf{x}'$的协方差矩阵Σ'为$\mathbf{x}$的相关矩阵: $\Sigma' = [\rho_{ij}]_{m\times m}$. 此时 $\mathrm{trace}(\Sigma') = m$.

$\mathbf{x}$的第 k主成分 y_k的方差贡献率 η_k定义为

$$\eta_k = \frac{\lambda_k}{\sum\limits_{i=0}^{m} \lambda_i}.$$

$\mathbf{x}$的前k个主成分 $y_1, \cdots, y_k$的累计方差贡献率 $\eta_{1\to k}$定义为

$$\eta_{1\to k} = \sum_{i=1}^{k} \eta_i = \frac{\sum\limits_{i=1}^{k} \lambda_i}{\sum\limits_{j=0}^{m} \lambda_j}.$$

累计方差贡献率 $\eta_{1\to k}$反映了$\mathbf{x}$的前k个主成分保留原有变量方差信息的比例，可以作为k的选择标准， 比如选择k使得 $\eta_{1\to k}$达到规定的百分比(如80%)以上.

- 主成分变换$\mathbf{y} = A^T\mathbf{x}$的逆变换为 $\mathbf{x} = A\mathbf{y}$.
- $x_i = \sum_{j=1}^{m} \alpha_{ij} y_j,\ i = 1, 2, \cdots, m.$
- 因子负荷量: 第k主成分与变量x_i的相关系数, 即y_k对x_i的贡献程度:

$$\rho(y_k, x_i) = \frac{\mathrm{Cov}(\sum_{j=1}^{m} \alpha_{ij} y_j, y_k)}{\sqrt{\lambda_k \sigma_{ii}}} = \frac{\alpha_{ik}\mathrm{Var}(y_k)}{\sqrt{\lambda_k \sigma_{ii}}} = \frac{\sqrt{\lambda_k}\alpha_{ik}}{\sqrt{\sigma_{ii}}}$$

- 因子负荷量满足如下性质:
  1. $\sum_{i=1}^{m} \sigma_{ii}\rho^2(y_k, x_i) = \lambda_k(\sum_{i=1}^{m} \alpha_{ik}^2) = \lambda_k.$
  2. $\rho^2(x_i, (y_1, \cdots, y_m)) = \sum_{k=1}^{m} \rho^2(y_k, x_i) = 1.$

$\mathbf{x}$的前k个主成分 $y_1, \cdots, y_k$的对原有变量x_i的贡献率 $\nu_{1\to k}(i)$定义为

$$\nu_{1\to k}(i) = \sum_{j=1}^{k} \rho^2(y_j, x_i) = \sum_{j=1}^{k} \frac{\lambda_j \alpha_{ij}^2}{\sigma_{ii}}$$

概要

1. 预备知识
2. 总体主成分分析
   - 主成分变换
   - 标准化随机变量
   - 方差贡献率
   - 因子负荷量
3. 样本主成分分析

- 设$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$是对$m$维随机向量$\mathbf{x} = (x_1, x_2, \cdots, x_m)^T$进行$n$次独立观测的样本，其中$\mathbf{x}_j = (x_{1j}, x_{2j}, \cdots, x_{mj})^T$表示第$j$个观测样本，则观测数据矩阵

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \vdots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$$

- 样本均值向量为$\bar{\mathbf{x}} = \frac{1}{n}\sum\limits_{j=1}^{n} \mathbf{x}_j = (\bar{x}_1, \cdots, \bar{x}_m)^T$.
- 样本协方差矩阵为$S = [s_{ij}]_{m\times m}$，其中$s_{ij} = \frac{1}{n-1}\sum\limits_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$, $i, j = 1, 2, \cdots, m$.
- 样本相关矩阵为 $R = [r_{ij}]_{m\times m}$，其中 $r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$.

- 定义m维随机向量$\mathbf{x}=(x_1,x_2,\cdots,x_m)^T$ 到 $\mathbf{y}=(y_1,y_2,\cdots,y_m)^T$ 的线性变换

$$\mathbf{y}=A^T\mathbf{x},$$

其中 $A=[\alpha_1,\alpha_2,\cdots,\alpha_m]=\begin{bmatrix}\alpha_{11} & \cdots & \alpha_{1m}\\ \vdots & \vdots & \vdots\\ \alpha_{m1} & \cdots & \alpha_{mm}\end{bmatrix}.$

- 对每个分量来说, $y_i=\alpha_i^T\mathbf{x}$.
- 对每个观测数据$\mathbf{x}_j$来说, $\mathbf{y}_j=A^T\mathbf{x}_j$.
- y_i对应于$\mathbf{X}$的样本均值$\bar{y}_i=\frac{1}{n}\sum_{j=1}^{n}\alpha_i^T\mathbf{x}_j=\alpha_i^T\bar{\mathbf{x}}$.

- y_i对应于$\mathbf{X}$的样本方差$\mathrm{Var}(y_i) = \alpha_i^T S \alpha_i$.
- y_i, y_j对应于$\mathbf{X}$的样本协方差$\mathrm{Cov}(y_i, y_j) = \alpha_i^T S \alpha_j$.

- 我们首先对数据进行规范化:

$$x'_{ik} = \frac{x_{ik} - \bar{x}_i}{\sqrt{s_{ii}}}, \quad i = 1, 2, \cdots, m; k = 1, 2, \cdots, n.$$

- 我们仍以x_{ik}表示规范化的x'_{ik}, 并将规范的样本矩阵仍然记为$\mathbf{X}$, 此时样本协方差矩阵

$$S = R = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T$$

- 我们设R的特征值为$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, α_i 为R的属于λ_i的单位特征向量, 则样本主成分变换为

$$\mathbf{y} = A^T\mathbf{x},$$

这里$A = [\alpha_1, \alpha_2, \cdots, \alpha_m]$.

- n个样本$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$的第$j$主成分的行向量为

$$(y_{j1}, \cdots, y_{jn}) = \alpha_j^T \mathbf{X}$$

- 样本前k主成分矩阵为

$$\mathbf{Y}_{k \times n} = [\alpha_1, \cdots, \alpha_k]^T \mathbf{X}$$

基于以上的分析，我们可以

- 求出R的特征值为$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$;
- 确定使得累计方差贡献率达到预定值的主成分个数k;
- 求出前k个特征值 λ_i对应的单位特征向量 α_i;
- 求k个样本主成分
$$y_i = \alpha_i^T x$$
- 计算$\rho(y_j, x_i)$以及 $\nu_{1\to k}(i)$;
- 计算样本前k主成分矩阵.

小结

- 标准线性组合
- 总体主成分变换
- 总体主成分性质
- 方差贡献率
- 因子负荷量
- 规范化数据
- 样本主成分分析