特征提取

特征提取（降维/特征约简）是指**将原始高维数据映射到低维空间**中。

不同问题下的特征提取**标准**：

无监督特征提取：减少信息损失

监督特征提取：实现正确分类

特征提取和特征选择的对比

特征提取**使用所有原始特征**，转换后的特征是**原始特征的线性组合**

特征选择仅使用**原始特征的子集**。

特征提取的原因

使数据可视化

数据压缩：高效存储和检索

移除噪声：帮助移除信息噪声，提高准确性

无监督PCA

- Two commonly used definitions of PCA
  - Maximum variance formulation最大化方差
    - The variance of the projected data is maximized.
  - Minimum-error formulation最小误差公式（减少信息损失）
    - Minimizes the average projection cost最小化平均投影成本

主成分分析的核心思想是降低由大量相关变量组成的数据集的维数，同时尽可能多地保留数据集中存在的变量。

这是通过转换为一组新的变量来实现的，即主成分（PC），它们是不相关的，并且按每个保留的总信息的分数排序，使得前几个保留了所有原始变量中存在的大部分变化。

协方差矩阵

协方差矩阵指示随机向量中的每对维度（特征）一起变化的趋势，即共同变化。

Covariance Matrix

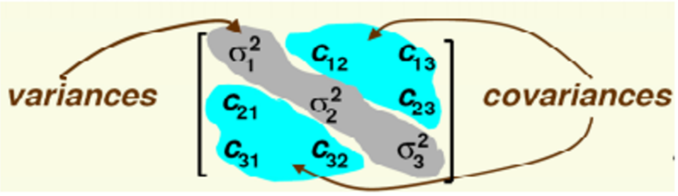
- Given random vector, $\vec{\mathbf{X}} = [x_1, x_2, ..., x_N]^T$, we define,

Mean vector $E[\boldsymbol{X}] = [E[X_1], E[X_2], ..., E[X_N]]^T = [\mu_1 \mu_2 ... \mu_N] = \boldsymbol{\mu}$

Covariance matrix

$$COV[\boldsymbol{X}] = \boldsymbol{\Sigma} = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T]$$

$$= \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] ... E[(X_1 - \mu_1)(X_N - \mu_N)] \\ \ddots \\ E[(X_N - \mu_N)(X_1 - \mu_1)] ... E[(X_N - \mu_N)(X_N - \mu_N)] \end{bmatrix} = \begin{bmatrix} \sigma_1{}^2 ... c_{1N} \\ ... \\ c_{N1} ... \sigma_N{}^2 \end{bmatrix}$$



协方差矩阵的性质

如果xi,xk正相关，则cik>0

如果xi,xk负相关，则cik<0

如果xi,xk无关,则cik=0

协方差矩阵是对称的

协方差矩阵是半正定矩阵

所有的特征值是非负的

行列式是非负的

- **Important Properties**
  - If x_i and x_k tend to increase together, then $c_{ik} > 0$
  - If x_i tends to decrease when x_k increases, then $c_{ik} < 0$
  - If x_i and x_k are uncorrelated, then $c_{ik} = 0$
  - $|c_{ik}| \le \sigma_i \sigma_k$, where σ_i is the standard deviation 标准差of x_i
  - $c_{ii} = \sigma_i{}^2 = VAR(x_i)$
  - Symmetric: $c_{ji} = c_{ij}$
  - Positive semi-definite:半正定矩阵
    - Eigenvalues are nonnegative所以特征值是非负的
    - Determinant is nonnegative, $|C| \ge 0$行列式是非负的

特征值与特征向量

Eigenvectors and Eigenvalues

- **Definition:** $\boldsymbol{v}$ is an eigenvector of matrix $\boldsymbol{A} \in \mathbb{R}^{m*m}$ if there exists a scalar λ, such that:

$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v} \quad \begin{cases} \boldsymbol{v}\text{: an eigenvector (nonzero vector)特征向量} \\ \lambda\text{: the corresponding eigenvalue特征值} \end{cases}$$

- **Computation**

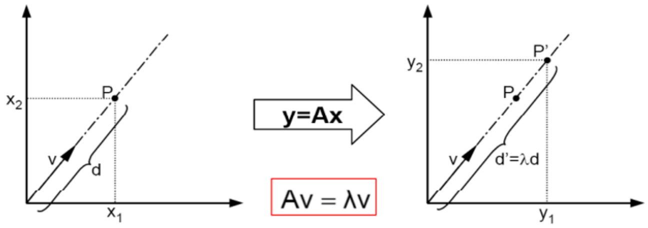
$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v} \qquad (\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{v} = \boldsymbol{0}$$

$$\boldsymbol{v} \neq \boldsymbol{0} \quad \Rightarrow \quad |\boldsymbol{A} - \lambda\boldsymbol{I}| = 0$$

- **Note**
  - $tr(\boldsymbol{A}) = \sum_i \lambda_i$
  - $|\boldsymbol{A}| = \prod_i \lambda_i$
  - If λ is an eigenvalue of the matrix $\boldsymbol{A}$, then λ^2 is an eigenvalue of $\boldsymbol{A}^2$. $(\boldsymbol{A}^2 = \boldsymbol{A}\boldsymbol{A})$
  - If λ is an eigenvalue of the matrix $\boldsymbol{A}$, then λ is an eigenvalue of $\boldsymbol{A}^T$.

- Intepretation: an eigenvector represents an invariant direction in the vector space.积分：特征向量表示向量空间中不变的方向。
  - Any point lying on the direction defined by $\boldsymbol{v}$ remains on that direction.位于由v定义的方向上的任何点都保持在该方向上。
  - Its magnitude is multiplies by the corresponding eigenvalue λ

其大小乘以相应的特征值λ



PCA的结果

生成的z1,z2,z3...分别是特征提取之后的m个数据点的第i维数值

SVD

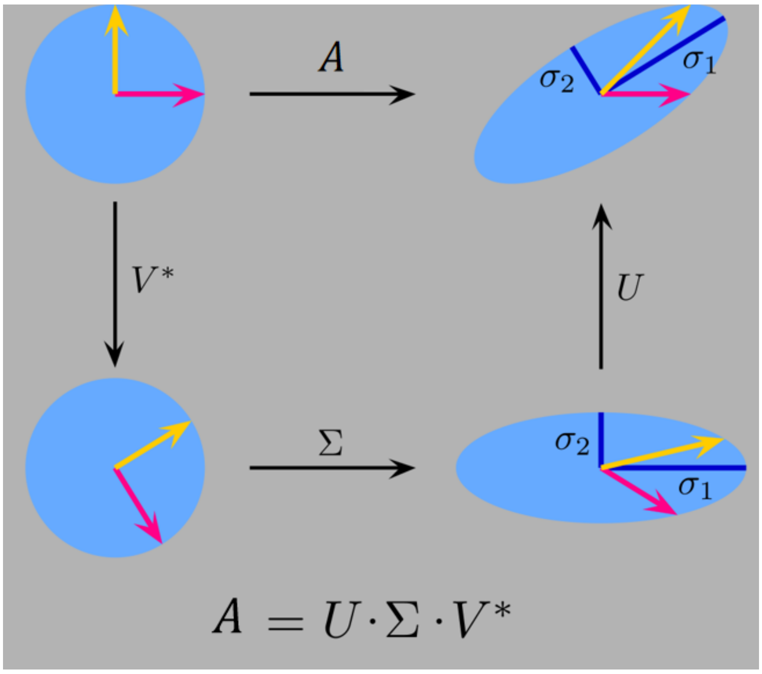
我们通过对中心数据矩阵的奇异值分解（SVD）来计算PC。

线性变换A可以解释为三个几何变换的组合

1.旋转或反射V^T

2.逐坐标缩放的坐标Σ

3.另一个旋转或反射U



- Any $m \times n$ matrix $\boldsymbol{A}$ of rank r can be decomposed into: $\boldsymbol{A} = \boldsymbol{U\Sigma V^T}$
  - For $m > n$ $\quad \boldsymbol{U}_{m\times m} \quad \boldsymbol{\Sigma}_{m\times n} \quad \boldsymbol{V}_{n\times n}$

$$\mathbf{A} = \begin{bmatrix} \vdots & & \vdots & \vdots & & \vdots \\ \mathbf{u}_1 & \cdots & \mathbf{u}_r & \mathbf{u}_{r+1} & \cdots & \mathbf{u}_m \\ \vdots & & \vdots & \vdots & & \vdots \end{bmatrix} \begin{bmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix} \begin{bmatrix} \cdots & \mathbf{v}_1^\top & \cdots \\ & \vdots & \\ \cdots & \mathbf{v}_r^\top & \cdots \\ \cdots & \mathbf{v}_{r+1}^\top & \cdots \\ & \vdots & \\ \cdots & \mathbf{v}_n^\top & \cdots \end{bmatrix}$$

- **Special Properties**:
  - The columns of $\boldsymbol{U}$ (i.e., left singular vectors) are eigenvectors of $\boldsymbol{AA^T}$.
  - The columns of $\boldsymbol{V}$ (i.e., right singular vectors) are eigenvectors of $\boldsymbol{A^T A}$.
  - Eigenvalues $\lambda_1, \ldots, \lambda_r$ of $\boldsymbol{AA^T}$ are the eigenvalues of $\boldsymbol{A^T A}$.
  - Singular value $\sigma_i = \sqrt{\lambda_i}$.

Compact SVD压缩版的SVD

- Only the $r = rank(\boldsymbol{A})$ column vectors of $\boldsymbol{U}$ and r row vectors of $\boldsymbol{V^T}$ corresponding to the non-zero singular values $\boldsymbol{\Sigma}_r$ are calculated.

$$\mathbf{A} = \begin{bmatrix} \vdots & & \vdots & \vdots & & \vdots \\ \mathbf{u}_1 & \cdots & \mathbf{u}_r & \mathbf{u}_{r+1} & \cdots & \mathbf{u}_m \\ \vdots & & \vdots & \vdots & & \vdots \end{bmatrix} \begin{bmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix} \begin{bmatrix} \cdots & \mathbf{v}_1^\top & \cdots \\ & \vdots & \\ \cdots & \mathbf{v}_r^\top & \cdots \\ \cdots & \mathbf{v}_{r+1}^\top & \cdots \\ & \vdots & \\ \cdots & \mathbf{v}_n^\top & \cdots \end{bmatrix} \quad m > n$$

- Economy version $\boldsymbol{A} = \underset{m\times r}{\boldsymbol{U}_r}\underset{r\times r}{\boldsymbol{\Sigma}_r}\underset{r\times n}{\boldsymbol{V}_r{}^{\boldsymbol{T}}}$ $\qquad \boldsymbol{\Sigma}_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$

Any information loss?没有信息损失

Truncated SVD截断SVD

- Only k column vectors of $\boldsymbol{U}$ and k row vectors of $\boldsymbol{V^T}$ corresponding to the non-zero singular values $\boldsymbol{\Sigma}_k$ are calculated, $0 < k < r, r = rank(\boldsymbol{A})$.

$$\mathbf{A} = \begin{bmatrix} \vdots & & \vdots & \vdots & & \vdots \\ \mathbf{u}_1 & \cdots & \mathbf{u}_r & \mathbf{u}_{r+1} & \cdots & \mathbf{u}_m \\ \vdots & & \vdots & \vdots & & \vdots \end{bmatrix} \begin{bmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix} \begin{bmatrix} \cdots & \mathbf{v}_1^\top & \cdots \\ & \vdots & \\ \cdots & \mathbf{v}_r^\top & \cdots \\ \cdots & \mathbf{v}_{r+1}^\top & \cdots \\ & \vdots & \\ \cdots & \mathbf{v}_n^\top & \cdots \end{bmatrix} \quad m > n$$

- More Economical $\boldsymbol{A} = \underset{m\times k}{\boldsymbol{U}_k}\underset{k\times k}{\boldsymbol{\Sigma}_k}\underset{k\times n}{\boldsymbol{V}_k{}^{\boldsymbol{T}}}$ $\qquad \boldsymbol{\Sigma}_k = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$

Truncated SVD is no longer an exact decomposition of the original matrix.

SVD有信息损失的情况：舍去了非0的奇异值

SVD Application3-PCA

- In practice, we compute the PCs via singular value decomposition (SVD) on the centered data matrix.
- Form the centered data matrix:

$$\boldsymbol{X} = [\,(\boldsymbol{x}_1 - \overline{\boldsymbol{x}}); \dots; (\boldsymbol{x}_m - \overline{\boldsymbol{x}})] \in \mathbb{R}^{d\times m}$$

- Compute its SVD:

$$\boldsymbol{X} = \boldsymbol{U}_{d\times d}\boldsymbol{D}_{d\times m}(\boldsymbol{V}_{m\times m})^{\boldsymbol{T}}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, $\boldsymbol{D}$ is a diagonal matrix.

- Note that the scatter/covariance matrix can be written as

$$\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}} = \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^{\boldsymbol{T}} \qquad \boldsymbol{X} = \boldsymbol{U}_{d\times d}\boldsymbol{D}_{d\times m}(\boldsymbol{V}_{m\times m})^{\boldsymbol{T}}$$

- The eigenvectors of $\boldsymbol{S}$ are the columns of $\boldsymbol{U}$ and the eigenvalues are the diagonal elements of $\boldsymbol{D}^2$.
- Take only a few significant eigenvalue-eigenvector pairs $p \ll d$. The new reconstructed sample from low-dim space is:

$$\widehat{\boldsymbol{x}}_i = \overline{\boldsymbol{x}} + \boldsymbol{U}_{d\times p}\left(\boldsymbol{U}_{d\times p}\right)^{\boldsymbol{T}}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})$$

Advantages of Using SVD for PCA

- No need to compute the covariance matrix $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}$不需要计算方差矩阵
- Numerically more accurate, since the formation of $\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}$ can cause loss of precision. 从数字上来说更准确，因为XX^T的形成会导致精度的损失。
  - For example, the Läuchli matrix:

$$\begin{pmatrix} 1 & 1 & 1 \\ \epsilon & 0 & 0 \\ 0 & \epsilon & 0 \\ 0 & 0 & \epsilon \end{pmatrix}^{T}$$

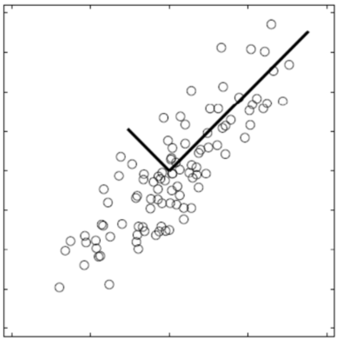
  where ϵ is a tiny number.

主成分的可视化

Visualize PCs

The columns of $\boldsymbol{U}$ are eigenvectors of $\boldsymbol{S} = \boldsymbol{A}\boldsymbol{A}^{\boldsymbol{T}}$.

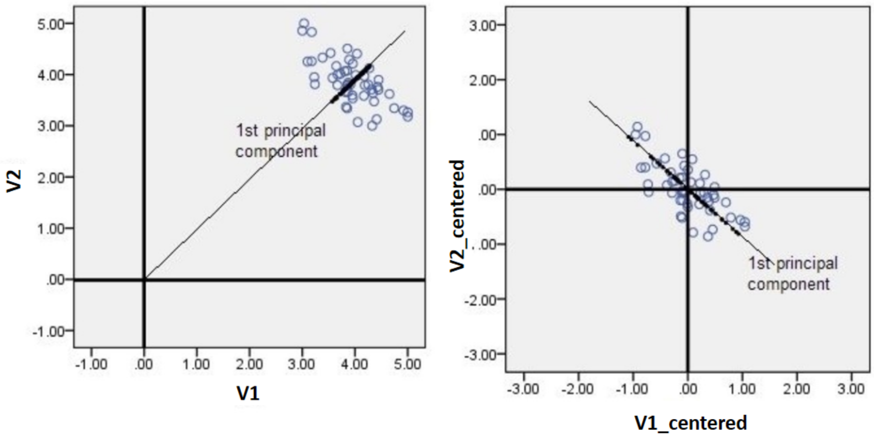
$$\boldsymbol{y} = \boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{x}$$



Data points are represented in a rotated orthogonal coordinate system: the origin is the mean of the data points and the axes are provided by the eigenvectors. 数据点在旋转的正交坐标系中表示：原点是数据点的平均值，轴由特征向量提供。

中心化的作用

The Necessity of Centralization中心化的必要性



中心化是对每一个样本向量做的，而预处理是对每一特征维度做的，这也可以反应中心化并不算是预处理。算法核心是在低纬度空间上依据方差最大使得数据之间差异体现，即降维不会丧失数据之间的差异性，如果没有中心化，则是没有体现方差的含义，不能体现数据的离散程度，所以不能够正确分类

主成分的保留

How Many PCs to Keep?设置一个保留主成分阈值

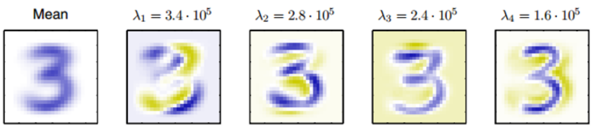
To choose p based on percentage of energy to retain, we can use the following criterion (smallest p):

$$\frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geq Threshold \quad (e.g.,\ 0.95)$$

PCA的应用

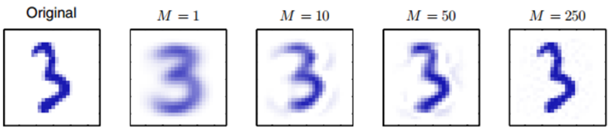
数据压缩

Data Compression数据压缩



We represent the eigenvectors as images of the same size as the data points.

The mean vector $\overline{\mathbf{x}}$ along with the first four PCA eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_4$ for the off-line digits data set, together with the corresponding eigenvalues.



An original example from the off-line digits data set together with its PCA reconstructions obtained by retaining M principal components for various values of M. As M increases the reconstruction becomes more accurate and would become perfect when $M = D = 28 \times 28 = 784$.

数据预处理

Data Preprocessing数据预处理

- The goal is **not** dimensionality reduction but rather the transformation of a data set in order to **standardizing** the data.目标不是降维，而是为了标准化数据而对数据集进行转换。
- Important in allowing subsequent pattern recognition algorithms to be applied successfully to the data set. 重要的是允许后续模式识别算法成功应用于数据集。
- Typically, it is done when the original variables are measured in different order of magnitudes or have significantly different variability.

通常，当原始变量以不同的数量级进行测量或具有显著不同的可变性时，就会进行预处理。

Data Preprocessing

- The goal is **not** dimensionality reduction but rather the transformation of a data set in order to **standardizing** the data.归一化

Traditionally, we can made a linear re-scaling of the individual variables such that each variable had zero mean and unit variance.通常处理： 均值为0，方差为1的正态分布

$$\frac{x_{ni} - \bar{x}_i}{\sigma_i}$$

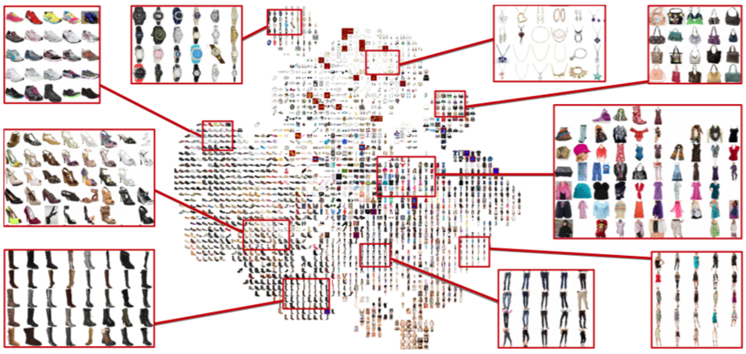
However, using PCA we can make a more substantial normalization of the data to give it zero mean and unit covariance, so that variables become decorrelated.
然而，使用主成分分析，我们可以对数据进行更实质性的归一化，
使其为零均值和单位协方差，从而使变量变得去相关。

数据可视化

PCA-Applications

Data Visualization数据可视化

- Each data point is projected onto a two-dimensional principal subspace.



分类问题

将训练集和测试集数据投射到主成分空间

对于每个测试样本，使用最近邻进行分类

问题：准确性对主成分数量很敏感

PCA对于分类来说，不一定是好的提取技术

- 主成分分析基于样本的协方差，协方差表达了整个数据集的分散性，与每个类的成员无关
- 由PCA选择的投影轴可能不能提供良好的分类能力

监督LDA

Linear Discriminant Analysis线性判别分析（LDA）

线性判别分析，一种找到分离两类或多类对象的特征的线性组合的方法。

标准：最大化类间散度，最小化类内散度

二分类