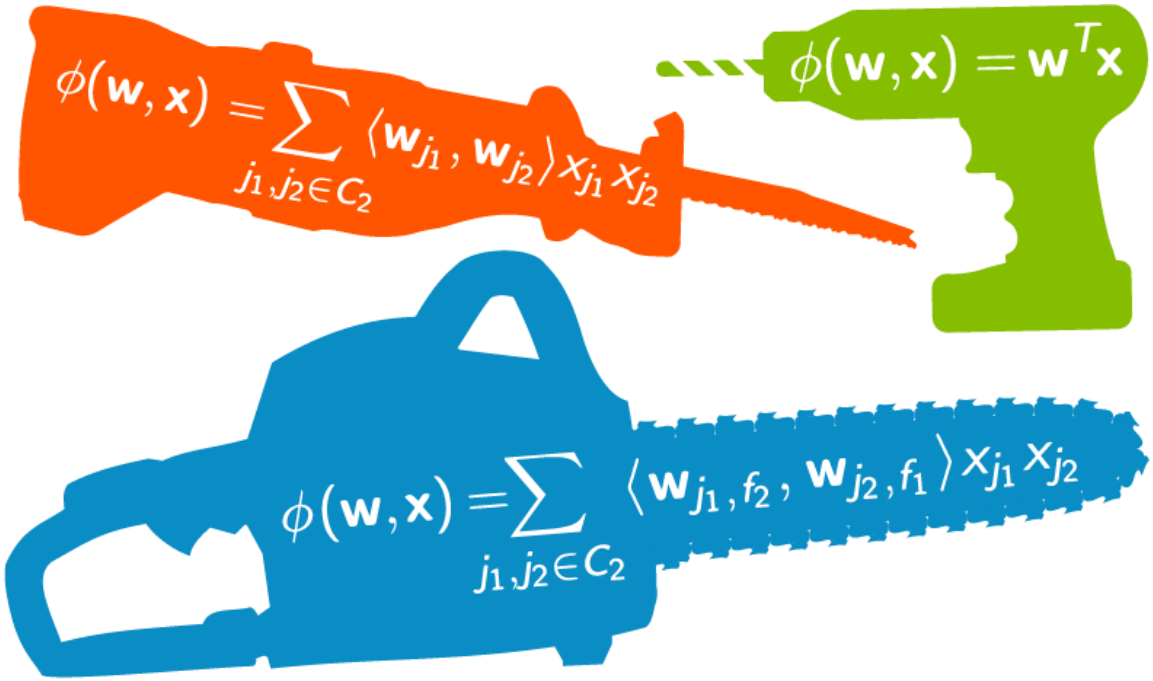
法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

关注 **小象学院**


$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2 \in C_2} \langle \mathbf{w}_{j_1}, \mathbf{w}_{j_2} \rangle x_{j_1} x_{j_2}$
$\phi(\mathbf{w},\mathbf{x}) = \mathbf{w}^T\mathbf{x}$
$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2 \in C_2} \langle \mathbf{w}_{j_1,f_2}, \mathbf{w}_{j_2,f_1} \rangle x_{j_1} x_{j_2}$

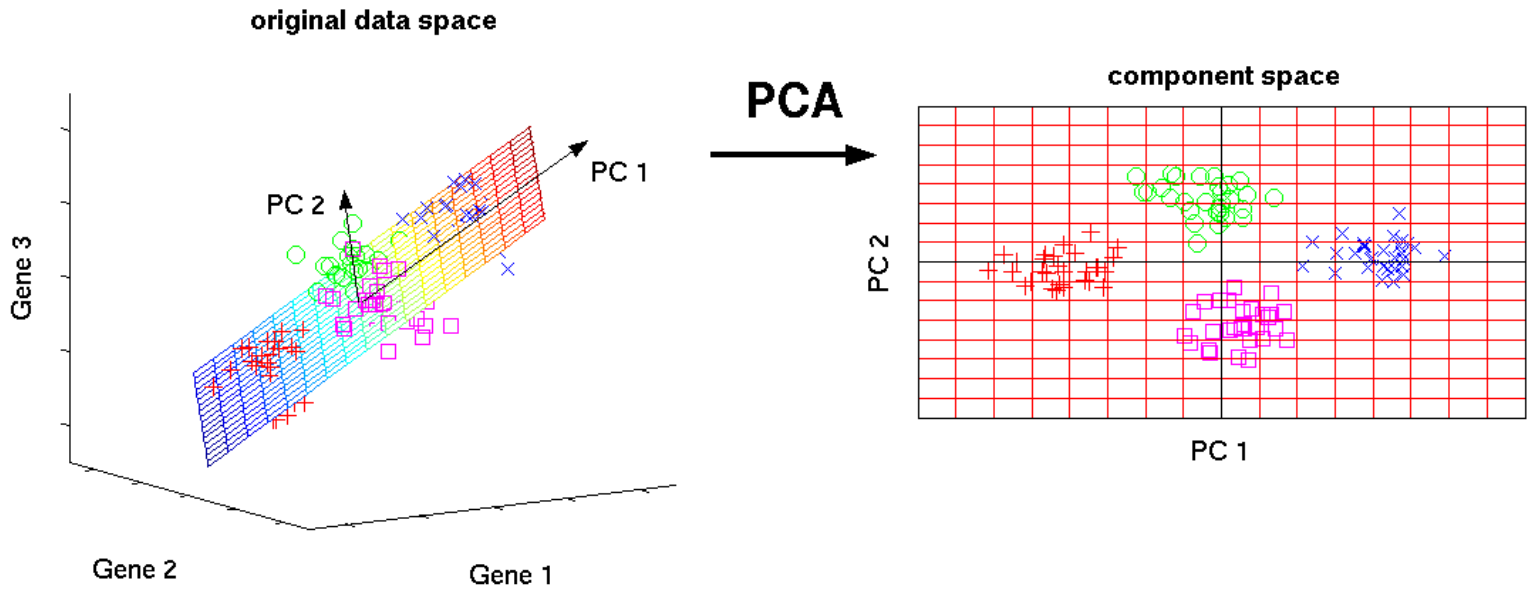
特征工程

--Robin

特征降维

Principal components analysis (PCA)

- 用于减少数据集的维度，同时保持数据集中的对方差贡献最大的特征
- 保留低阶主成分，忽略高阶成分，这样的低阶成分往往能够保留住数据的最重要方面



特征降维

方差与协方差

- 用于衡量一系列点在它们的重心或均值附近的分散程度
- 方差：衡量这些点在一个维度的偏差
- 协方差：衡量一个维度是否会对另一个维度有所影响，从而查看这两个维度之间是否有关系
  - 某个维度和自身之间的协方差就是其方差

$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

协方差矩阵

- 如果数据集是d维的，（x1，x2，…，xd），则可计算出(x1，x2)，（x1，x3），…，(x1，xd)，（x2，x3），…(x2，xd)，…(xd-1，xd)之间的协方差。由于协方差的对称性，再加上各维度自身的协方差，可以构成协方差矩阵

特征降维

$$\begin{bmatrix} \mathrm{cov}(x1,x1) & \mathrm{cov}(x1,x2) & \mathrm{cov}(x1,x3) \\ \mathrm{cov}(x2,x1) & \mathrm{cov}(x2,x2) & \mathrm{cov}(x2,x3) \\ \mathrm{cov}(x3,x1) & \mathrm{cov}(x3,x2) & \mathrm{cov}(x3,x3) \end{bmatrix}$$

协方差矩阵 (续)

- 其中对角线上的是方差
- 协方差为正,代表两个变量变化趋势相同；反之亦然

PCA

- 通过线型变换将原数据映射到新的坐标系统中，使映射后的第一个坐标上的方差最大（即第一个主成分）,第二个坐标上的方差第二大（即第二个主成分）…

特征降维

PCA步骤：

1. 数据集 $\mathbf{X} \in R^{m \times n}$ ，其中每个样本 $\mathbf{x}^{(\mathrm{i})} = [\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, ..., \mathbf{x}_n^{(i)}]$

   计算每个维度的均值

$$\overline{\mathbf{x}} = \frac{1}{m} \sum_{i=1}^{m} [\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, ..., \mathbf{x}_n^{(i)}] \in R^n$$

   每个维度减去这个均值，得到一个矩阵

   相当于将坐标系进行了平移

$$\mathbf{Y} = \begin{bmatrix} \mathbf{x}^{(1)} - \overline{\mathbf{x}} \\ \mathbf{x}^{(2)} - \overline{\mathbf{x}} \\ \cdots \\ \mathbf{x}^{(\mathrm{m})} - \overline{\mathbf{x}} \end{bmatrix}$$

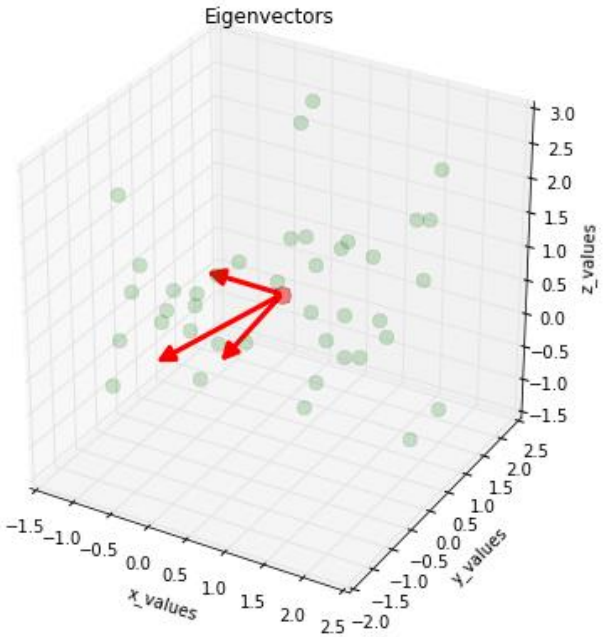
特征降维

PCA步骤：

2. 构建协方差矩阵 $\mathbf{Q} = \mathbf{Y}^T\mathbf{Y} = \begin{bmatrix} \mathbf{x}^{(1)} - \overline{\mathbf{x}} & \mathbf{x}^{(2)} - \overline{\mathbf{x}} & \dots & \mathbf{x}^{(m)} - \overline{\mathbf{x}} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)} - \overline{\mathbf{x}} \\ \mathbf{x}^{(2)} - \overline{\mathbf{x}} \\ \dots \\ \mathbf{x}^{(m)} - \overline{\mathbf{x}} \end{bmatrix}$

3. 矩阵分解 (如SVD)，得到特征值(eigenvalues)及特征向量 (eigenvectors)

4. 将特征值从大到小排序，对应的特征向量就是第一个主成分，第二个主成分…



如何选择主成分个数？

- 交叉验证
- 根据主成分的累计贡献率(t)

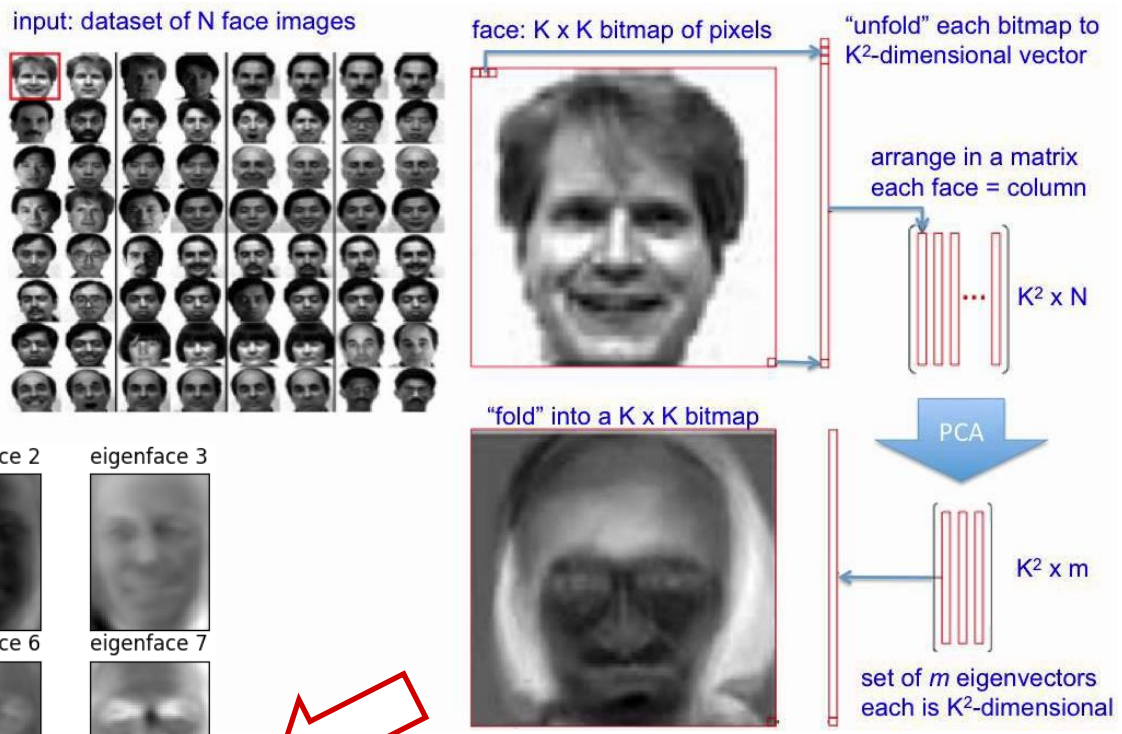
$$\frac{\sum_{i=1}^{d'} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geqslant t$$

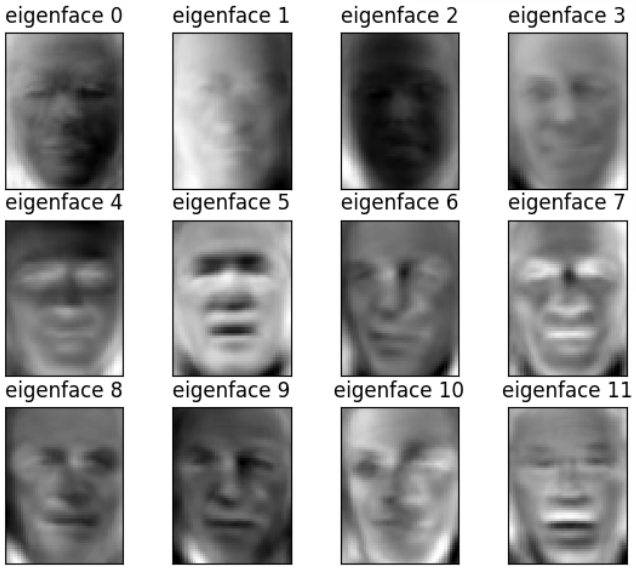
应用: 特征提取、数据降维

特征降维

应用:

- 数据降维
- 特征提取





联系我们

小象学院：互联网新技术在线教育领航者

– 微信公众号：**小象学院**