《数据挖掘技术》

⭐ CH16 主成分分析

➡️ Create by *Wang JingHui*

➡️ Last Revision Time: *2021.04.08*

主要内容

1. 总体主成分分析
    i. 基本想法
    ii. 定义和导出
    iii. 主要性质
    iv. 主成分的个数
    v. 规范化变量的总体主成分
2. 样本主成分分析
    i. 样本主成分的定义和性质
    ii. 相关矩阵的特征值分解算法
    iii. 数据矩阵的奇异值分解算法

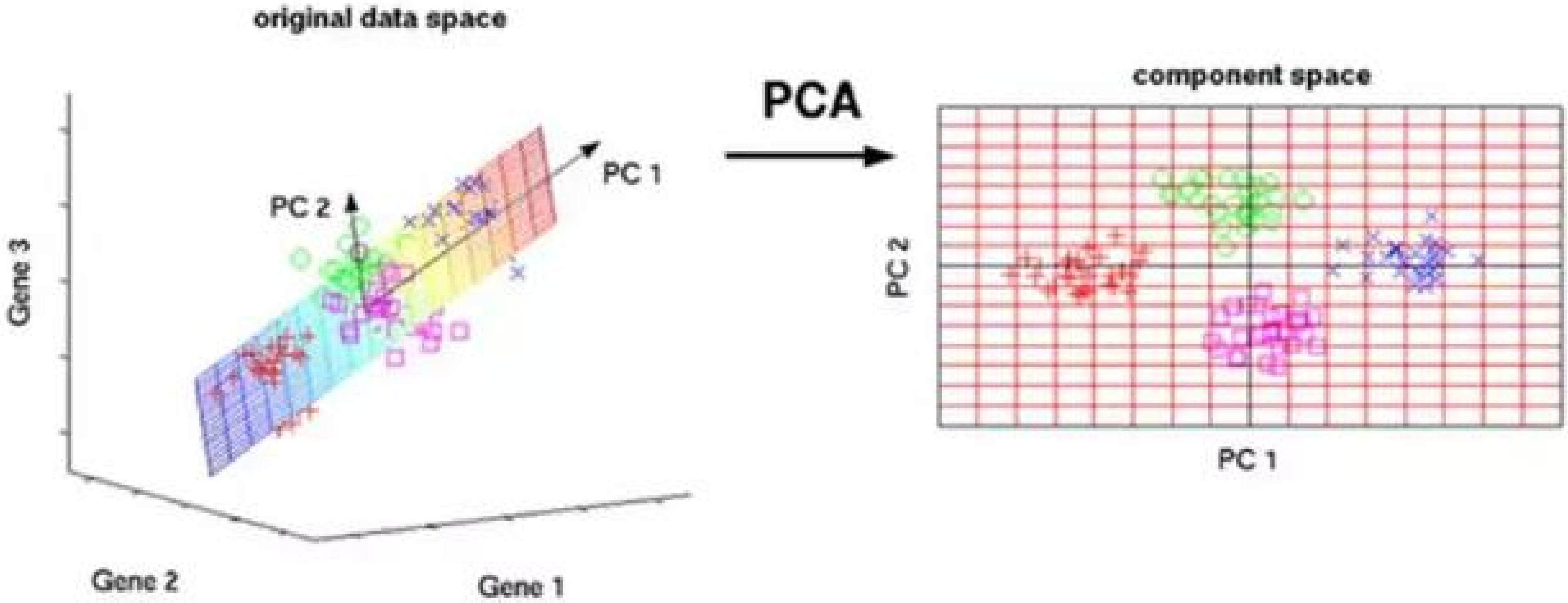
导读

- PCA的基本想法是由少数不相关的变量来代替相关的变量，用来表示数据，并且要求能够保留数据中的大部分信息。注意这个不是特征选择，得到的主成分是线性无关的新变量。
- 所谓线性相关的x_1和x_2就是说知道x_1的值的情况下，x_2的预测不是完全随机的。
- 主成分分析的结果可以作为其他机器学习方法的输入。
- 关于主成分的性质，规范化的变量总体主成分主要是围绕特征值和特征向量展开的。
- 关于总体和样本的说明可以参考一下Strang的书[^1]中第十二章部分说明。
- 关于k的选择，2000年有一个文章自动选择[^2]。

主成分分析

PCA(Principal Component Analysis)，即主成分分析方法，是一种使用最广泛的数据降维算法。

- PCA的主要思想是将n维特征映射到k维上，这k维是全新的正交特征也被称为主成分，是在原有n维特征的基础上重新构造出来的k维特征。
- PCA的工作就是从原始的空间中顺序地找一组相互正交的坐标轴，新的坐标轴的选择与数据本身是密切相关的。


original data space
PCA
component space
PC 1
PC 2
Gene 3
Gene 2
Gene 1
PC 2
PC 1

总体主成分性质

PCA选取包含信息量最多的方向对数据进行投影。其投影方向可以从最大化方差或者最小化投影误差两个角度理解。

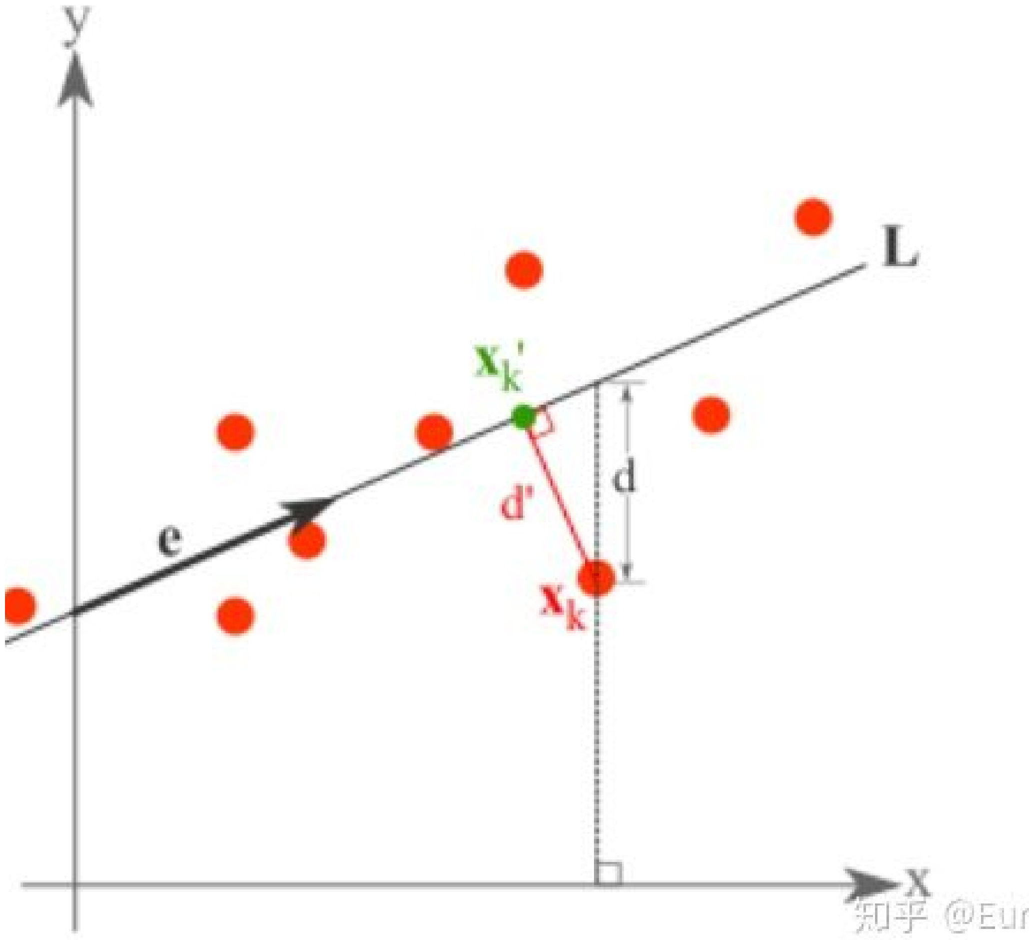
- 第一种解释是样本点到这个直线的距离足够近;
- 第二种解释是样本点在这个直线上的投影能尽可能的分开。

基于上面的两种标准，我们可以得到PCA的两种等价推导。

PCA的推导：基于最小投影距离

- 样本点到超平面的距离足够近。
- 假设二维样本点（红色点），要从二维降到一维，本质是求一个向量（或一条直线），线性回归时最小二乘法度量的是样本点到直线的坐标轴距离，即图中的 d，PCA里选取指标是图中的 d'。
- 数据维数等于 m，样本大小是 n 的数据 $x_1, ..., x_n$。将样本点 x_k 在直线上的投影记为 x'_k，那么就是要最小化：

$$\sum_{k=1}^{m} ||x'_k - x_k||^2$$



最近重构性

- 假定数据样本进行了中心化，即$\sum_i x_i = 0$;
- 假定投影变换后得到的新坐标系为$w_1, w_2, ..., w_d$，其中w_i是标准正交基向量，$||w_i||_2 = 1, w_i^T w_j = 0 \ (i \neq j)$
- 若丢弃新坐标中的部分坐标，将维数降低到$d' < d$，则样本点x_i在低维坐标系中的投影是$z_i = (z_{i1}; z_{i2}; ...; z_{id'})$,其中$z_{ij} = w_j^T x_i$是$x_j$在低维坐标系下第$j$维的坐标；
- 基于z_i来重构x_i;

考虑整个训练集，原样本点x_i与基于投影重构的样本点$\hat{x}_i$之间的距离

$$\begin{aligned}\sum_{i=1}^{m}\left\|\sum_{j=1}^{d'} z_{ij}\boldsymbol{w}_j-\boldsymbol{x}_i\right\|_2^2 &= \sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i-2\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i+\ \mathrm{const}\\ &\propto -\operatorname{tr}(\mathbf{W}^{\mathrm{T}}(\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}})\mathbf{W})\end{aligned}$$

[推导]：已知$\mathbf{W}^{\mathrm{T}}\mathbf{W}=\mathbf{I},\boldsymbol{z}_i=\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i$，则

$$
\begin{aligned}
\sum_{i=1}^{m}\left\|\sum_{j=1}^{d'} z_{ij}\boldsymbol{w}_j-\boldsymbol{x}_i\right\|_2^2 &= \sum_{i=1}^{m}\|\mathbf{W}\boldsymbol{z}_i-\boldsymbol{x}_i\|_2^2=\sum_{i=1}^{m}(\mathbf{W}\boldsymbol{z}_i-\boldsymbol{x}_i)^{\mathrm{T}}(\mathbf{W}\boldsymbol{z}_i-\boldsymbol{x}_i)\\
&=\sum_{i=1}^{m}\left(\boldsymbol{z}_i^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\boldsymbol{z}_i-\boldsymbol{z}_i^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{x}_i^{\mathrm{T}}\mathbf{W}\boldsymbol{z}_i+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i\right)=\sum_{i=1}^{m}\left(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i-2\boldsymbol{z}_i^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i\right)\\
&=\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i-2\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i+\sum_{i=1}^{m}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i=\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i-2\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i+\text{ const}\\
&=\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i-2\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i+\text{ const }=-\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i+\text{ const}\\
&=-\sum_{i=1}^{m}\operatorname{tr}\left(\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\right)+\text{ const }=-\operatorname{tr}\left(\sum_{i=1}^{m}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\right)+\text{ const}\\
&=-\operatorname{tr}\left(\sum_{i=1}^{m}\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\mathbf{W}\right)+\text{ const }=-\operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\left(\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right)\mathbf{W}\right)+\text{ const}\\
&\propto-\operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\left(\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right)\mathbf{W}\right)
\end{aligned}
$$

[推导]：由式（10.15）可知，主成分分析的优化目标为

$$\begin{aligned}
\min_{\mathbf{W}} \quad & -\operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}\right) \\
s.t. \quad & \mathbf{W}^{\mathrm{T}}\mathbf{W}=\mathbf{I}
\end{aligned}$$

其中，$\mathbf{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m)\in\mathbb{R}^{d\times m},\mathbf{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_{d'})\in\mathbb{R}^{d\times d'}$，$\mathbf{I}\in\mathbb{R}^{d'\times d'}$为单位矩阵。对于带矩阵约束的优化问题，可得此优化目标的拉格朗日函数为

$$\begin{aligned}
L(\mathbf{W},\Theta) &= -\operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}\right)+\langle\Theta,\mathbf{W}^{\mathrm{T}}\mathbf{W}-\mathbf{I}\rangle \\
&= -\operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}\right)+\operatorname{tr}\left(\Theta^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\mathbf{W}-\mathbf{I})\right)
\end{aligned}$$

$$
\begin{aligned}
\min_{\mathbf{W}} - \operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}\right) &= \max_{\mathbf{W}}\ \operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}\right) \\
&= \max_{\mathbf{W}} \sum_{i=1}^{d'} \boldsymbol{w}_i^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\boldsymbol{w}_i \\
&= \max_{\mathbf{W}} \sum_{i=1}^{d'} \boldsymbol{w}_i^{\mathrm{T}} \cdot \lambda_i \boldsymbol{w}_i \\
&= \max_{\mathbf{W}} \sum_{i=1}^{d'} \lambda_i \boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{w}_i \\
&= \max_{\mathbf{W}} \sum_{i=1}^{d'} \lambda_i
\end{aligned}
$$

使用拉格朗日得到：

$$\mathbf{X}\mathbf{X}^{\mathrm{T}}\boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i$$

于是，只需要对协方差矩阵XX^T进行特征值分解；

- 将求出的特征值排序：$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d$
- 取出前d'个特征值对应的特征向量构成$W^* = (w_1, w_2, ..., w_{d'}$就是主成分分析的解。

样本主成分分析

观测数据上进行主成分分析就是样本主成分分析。

给定样本矩阵X，可以**估计**样本均值以及样本协方差。

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

相关矩阵的特征值分解算法

1. 观测数据规范化处理，得到规范化数据矩阵X
2. 计算相关矩阵R

$$R=[r_{ij}]_{m\times m}=\frac{1}{n-1}XX^{\mathrm{T}}$$

$$r_{ij}=\frac{1}{n-1}\sum_{l=1}^{n}x_{il}x_{lj},i,j=1,2,\cdots,m$$

3. 求R的特征值和特征向量

$$|R-\lambda I|=0$$
$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$$

求累计方差贡献率达到预定值的主成分个数k

$$\sum_{i=1}^{k} \eta_i = \frac{\sum\limits_{i=1}^{k} \lambda_i}{\sum\limits_{i=1}^{m} \lambda_i}$$

求前k个特征值对应的单位特征向量

$$a_i = (a_{1i}, a_{2i}, \cdots, a_{mi})^{\mathrm{T}}$$

4. 求k个样本主成分

$$y_i = a_i^{\mathrm{T}} \boldsymbol{x}$$

其实算法到这就完事了，剩下两部分是输出。**前面是fit部分，后面是transform部分。**具体可以看下P_{319}中的关于相关矩阵特征值分解算法部分内容，构造正交矩阵之后就得到了主成分。

5. 计算k个主成分y_i与原变量x_i的相关系数$\rho(x_i, y_i)$以及k个主成分对原变量x_i的贡献率ν_i

$$\nu_i = \rho^2(x_i, (y_1, y_2, \cdots, y_k)) = \sum_{j=1}^{k} \rho^2(x_i, y_j) = \sum_{j=1}^{k} \lambda_j a_{ij}^2$$
$$i = 1, 2, \cdots, m$$

6. 计算n个样本的k个主成分值

第j个样本，$\boldsymbol{x}_j = (x_{1j}, x_{2j}, \cdots, x_{mj})^{\mathrm{T}}$的第$i$个主成分值是

$$y_{ij} = (a_{1i}, a_{2i}, \cdots, a_{mi})(x_{1j}, x_{2j}, \cdots, x_{mj})^{\mathrm{T}} = \sum_{l=1}^{m} a_{li} x_{lj}$$
$$i = 1, 2, \cdots, m, j = 1, 2, \cdots, n$$

主成分分析算法

输入： n维样本集$X = (x_1, x_2, ..., x_m)$，要降维到的维数为n'

输出： 降维后的样本集体Y

算法过程：

1. 对所有的样本进行中心化$x_i = x_i - \frac{1}{m}\sum_{j=1}^{m} x_j$
2. 计算样本的协方差矩阵$C = \frac{1}{m}XX^T$
3. 求出协方差矩阵的特征值及对应的特征向量
4. 将特征向量按对应特征值大小从上到下按行排列成矩阵，取前k行组成矩阵P
5. $Y = PX$为降维到K维后的数据

例：

原始数据集矩阵X：

$$\mathbf{A}=\begin{bmatrix}1 & 1 & 2 & 4 & 2\\ 1 & 3 & 3 & 4 & 4\end{bmatrix}$$

求均值后：

$$\mathbf{A}=\begin{bmatrix}-1 & -1 & 0 & 2 & 0\\ -2 & 0 & 0 & 1 & 1\end{bmatrix}$$

再求协方差矩阵

$$\mathbf{C}=\begin{bmatrix}-1 & -1 & 0 & 2 & 0\\ -2 & 0 & 0 & 1 & 1\end{bmatrix}\begin{bmatrix}-1 & -2\\ -1 & 0\\ 0 & 0\\ 2 & 1\\ 0 & 1\end{bmatrix}=\begin{bmatrix}\frac{6}{5} & \frac{4}{5}\\ \frac{4}{5} & \frac{6}{5}\end{bmatrix}$$

特征值：

$$\lambda_1 = 2, \lambda_2 = 25$$

对应的特征向量：

$$c_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}, c_1 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

标准化:

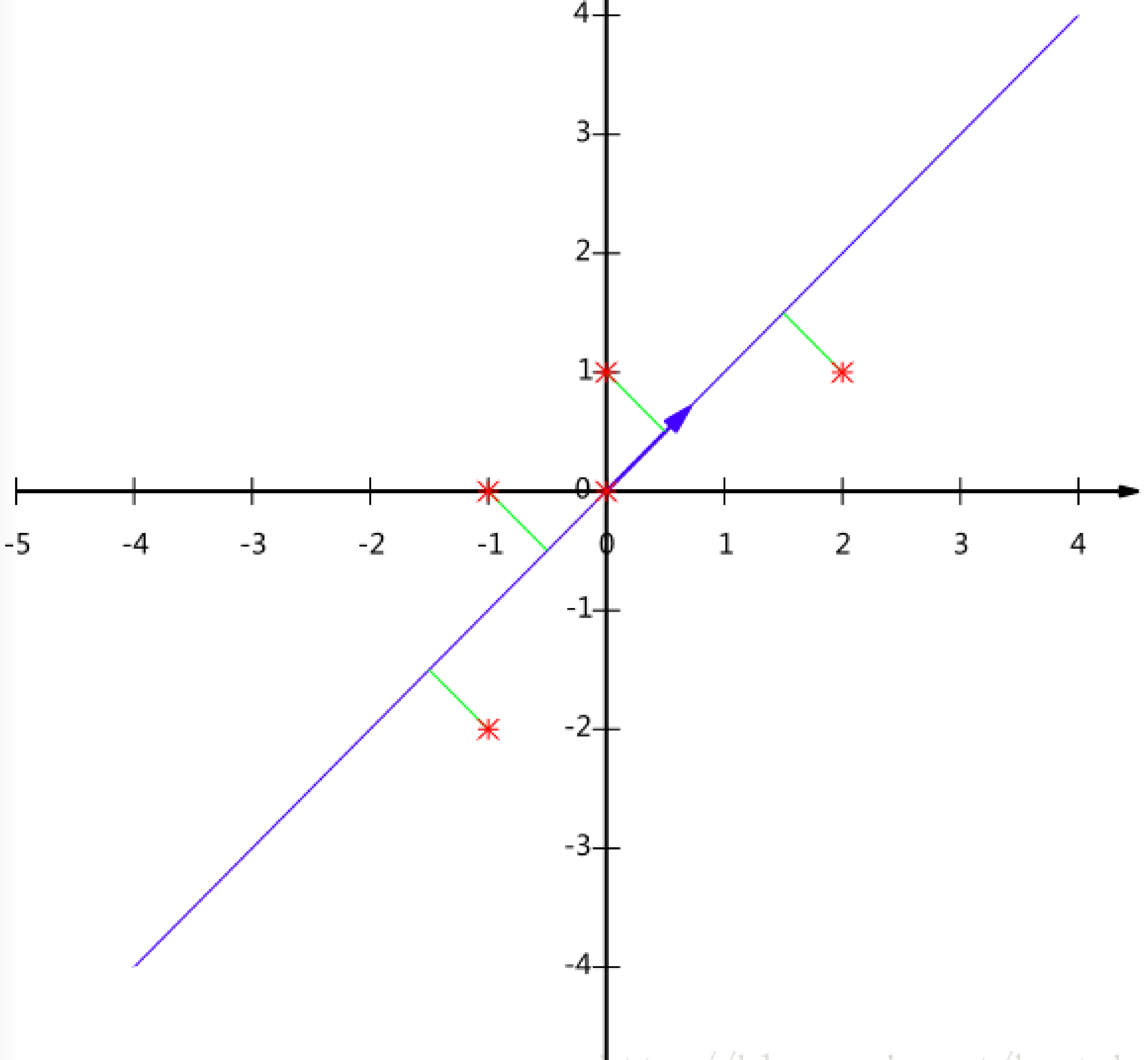
$$c_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

选择较大特征值对应的特征向量：

$$c_1 = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]$$

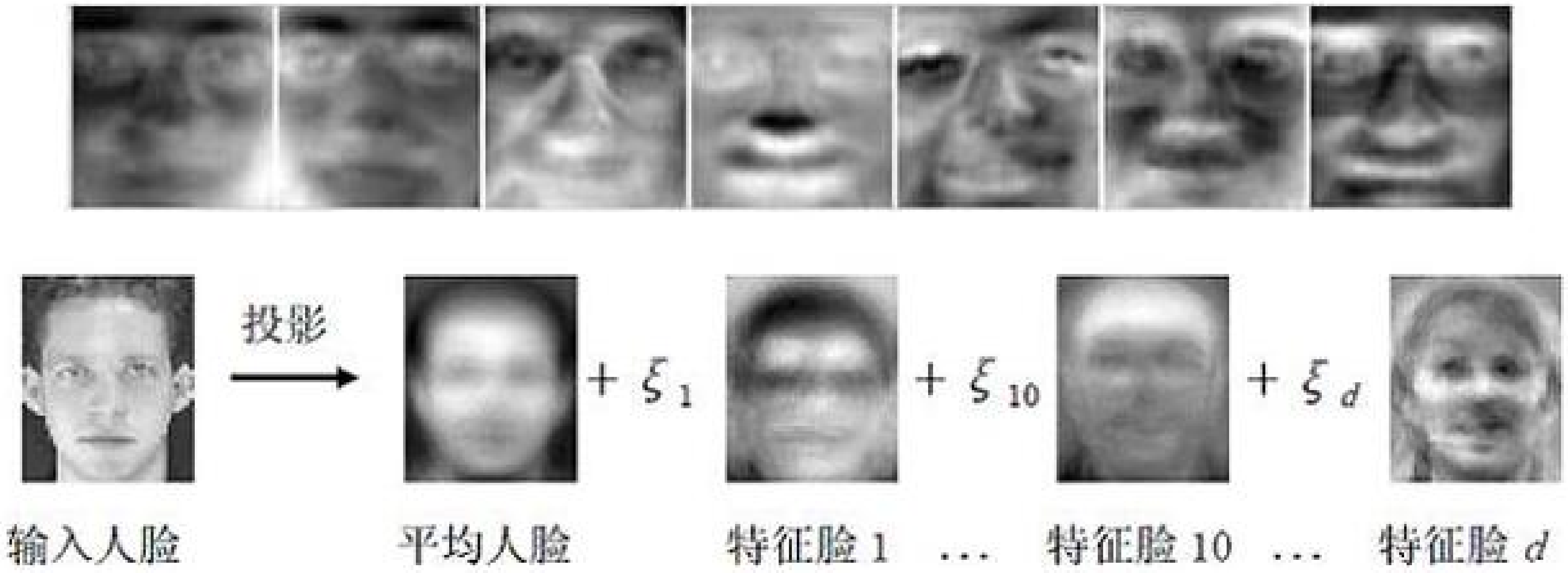
执行PCA变换：Y=PX，得到的Y就是PCA降维后的值数据集矩阵：

$$Y = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} -\frac{3}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \frac{3}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$


4
3
2
1
0
-1
-2
-3
-4
-5
-4
-3
-2
-1
0
1
2
3
4

应用

1. 降维
2. 人脸识别



1）将训练集的每一个人脸图像都拉长一列，将他们组合在一起形成一个大矩阵A。假设每个人脸图像是$M \times M$大小，那么拉成一列后每个人脸样本的维度就是$d = M \times M$大小了。假设有N个人脸图像，那么样本矩阵A的维度就是$d \times N$了。

2）将所有的N个人脸在对应维度上加起来，然后求平均，得到了一个“平均脸”。

3）将N个图像都减去那个平均脸图像，得到差值图像的数据矩阵。

4）计算协方差矩阵。再对其进行特征值分解。就可以得到特征向量（特征脸）了。

5）将训练集图像和测试集的图像都投影到这些特征向量上了，再对测试集的每个图像找到训练集中的最近邻或者k近邻，进行分类即可。

SVD 和 PCA 比较:

- 两者都是矩阵分解的技术，一个直接分解SVD，一个是对协方差矩阵操作后分解PCA
- 奇异值和特征向量存在关系，
- SVD可以获取另一个方向上的主成分，而PCA只能获得单个方向上的主成分，PCA只与SVD的右奇异向量的压缩效果相同
- 通过SVD可以得到PCA相同的结果，但是SVD通常比直接使用PCA更稳定。因为在PCA求协方差时很可能会丢失一些精度。

参考

[1] https://www.jianshu.com/p/471d9bfbd72f

[2] http://www.ruanyifeng.com/blog/2013/03/tf-idf.html

[3] https://www.cnblogs.com/iloveai/p/word2vec.html

[4] 协方差计算 https://blog.csdn.net/ybdesire/article/details/6270328

[5] Introduction to Linear Algebra

[6] Automatic choice of dimensionality for PCA

Enjoy your machine learning!

https://github.com/wjssx/Statistical-Learning-Slides-Code

E-mail: csr_dsp@sina.com