浙江工商大学
ZHEJIANG GONGSHANG UNIVERSITY

高级人工智能

统计机器学习:无监督学习

目录

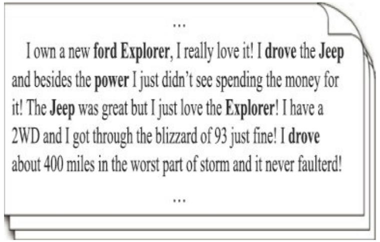
机器学习：从数据中学习映射函数

图像数据

$$f\left\{\begin{matrix}81&116&\cdots&133\\104&130&\cdots&159\\\vdots&\vdots&\ddots&\vdots\\155&189&\cdots&218\\197&221&\cdots&216\end{matrix}\right\}$$

- Person
- Dog
- …

类别分类



…

I own a new **ford Explorer**, I really love it! I **drove** the **Jeep** and besides the **power** I just didn't see spending the money for it! The **Jeep** was great but I just love the **Explorer**! I have a 2WD and I got through the blizzard of 93 just fine! I **drove** about 400 miles in the worst part of storm and it never faulterd!

…

文本数据

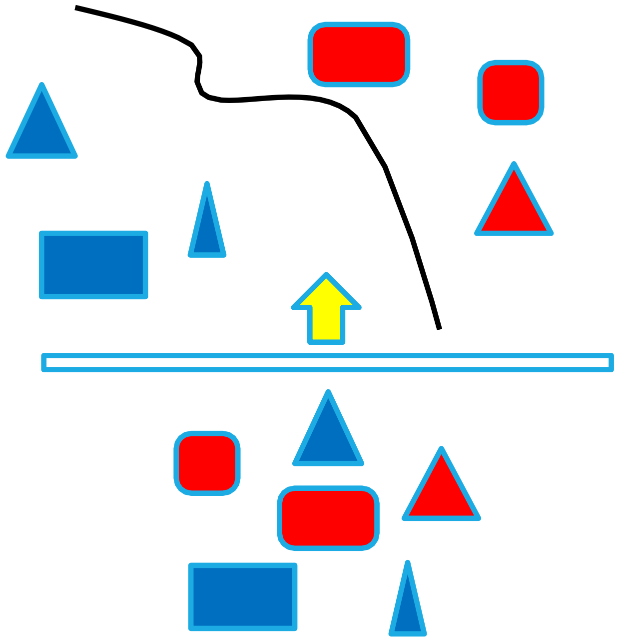
f {car, money, drive,…}

- 喜悦
- 愤怒
- …

情感分类

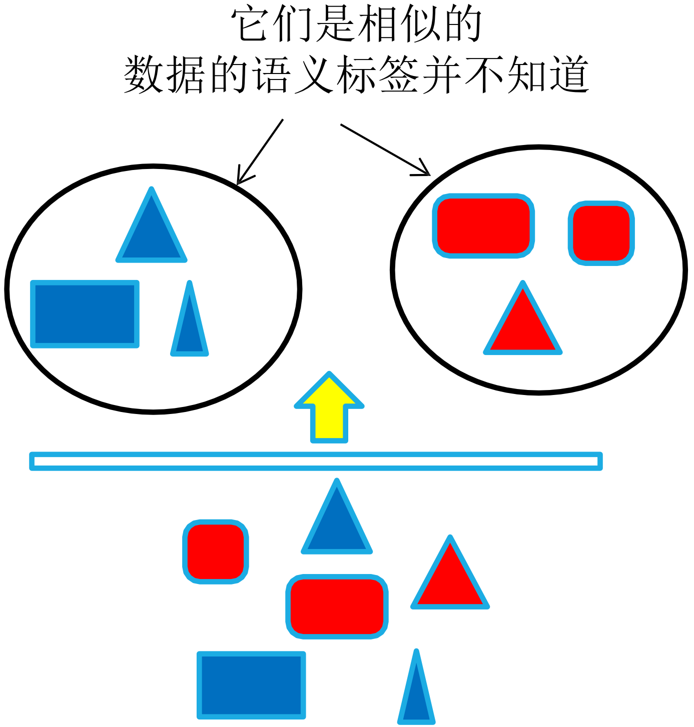
1. 原始数据中提取特征
2. 学习映射函数f
3. 通过映射函数f将原始数据映射到语义空间，即寻找数据和任务目标之间的关系

监督学习 versus 无监督学习



红色：汽车　　蓝色：飞机

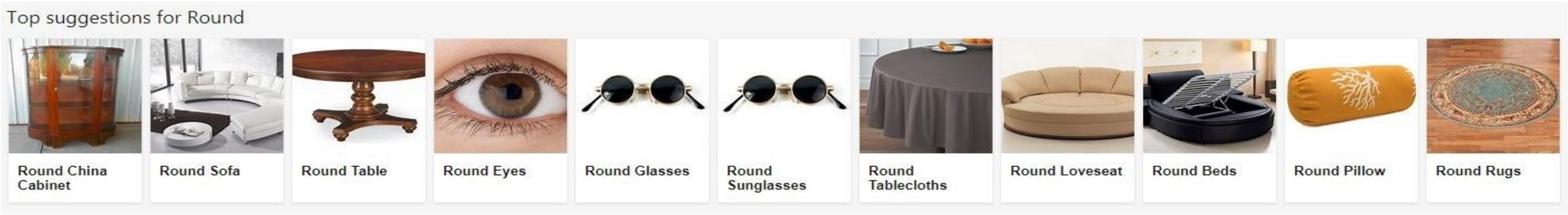
左：监督学习



右：无监督学习

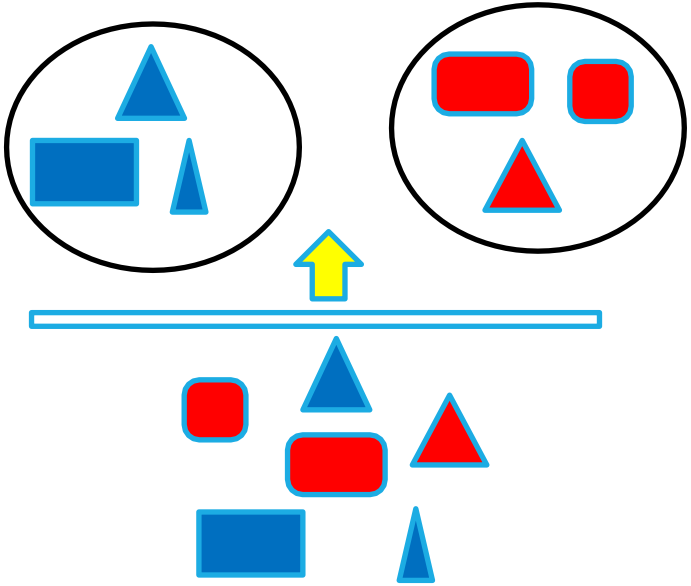
无监督学习的重要因素

数据特征	图像中颜色、纹理或形状等特征	听觉信息中旋律和音高等特征	文本中单词出现频率等特征
相似度函数	定义一个相似度计算函数，基于所提取的特征来计算数据之间的相似性		

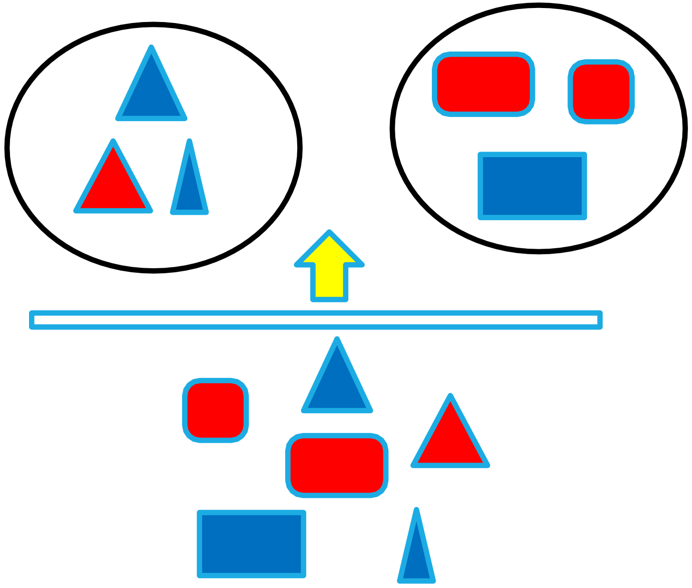


无监督学习：数据特征和相似度函数都很重要

相似度函数：颜色相似



相似度函数：形状相似



无监督学习

K均值聚类(K-means 聚类)

- 物以类聚，人以群分(《战国策·齐策三》)
- 输入：n个数据（无任何标注信息）
- 输出：k个聚类结果
- 目的：将n个数据聚类到k个集合（也称为类簇）

K均值聚类算法描述

- 若干定义：
  - n个m-维数据$\{x_1, x_2, \dots, x_n\}$，$x_i \in R^m (1 \le i \le n)$
  - 两个m维数据之间的欧氏距离为

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{im} - x_{jm})^2}$$

$d(x_i, x_j)$ 值越小，表示x_i和x_j越相似；反之越不相似

  - 聚类集合数目k
- 问题：如何将n个数据依据其相似度大小将它们分别聚类到k个集合，使得每个数据仅属于一个聚类集合。

K均值聚类算法：初始化

■ 第一步：初始化聚类质心

- 初始化k个聚类质心$c = \{c_1, c_2, \ldots, c_k\}$，$c_j \in R^m (1 \leq j \leq k)$
- 每个聚类质心c_j所在集合记为G_j

K均值聚类算法：对数据进行聚类

■ 第二步：将每个待聚类数据放入唯一一个聚类集合中

- 计算待聚类数据x_i和质心c_j之间的欧氏距离

  $d(x_i, c_j)\ (1 \le i \le n, 1 \le j \le k)$

- 将每个x_i放入与之距离最近聚类质心所在聚类集合中，

  即 $\underset{c_j \in C}{\operatorname{argmin}}\, d(x_i, c_j)$

K均值聚类算法：更新聚类质心

■ 第三步：根据聚类结果、更新聚类质心

- 根据每个聚类集合中所包含的数据，更新该聚类集合质心值，即：$c_j = \frac{1}{|G_j|}\sum_{x_i \in G_j} x_i$

K均值聚类算法：继续迭代

- 第四步：算法循环迭代，直到满足条件

- 在新聚类质心基础上，根据欧氏距离大小，将每个待聚类数据放入唯一一个聚类集合中
- 根据新的聚类结果、更新聚类质心

聚类迭代满足如下任意一个条件，则聚类停止：

- 已经达到了迭代次数上限
- 前后两次迭代中，聚类质心基本保持不变

K均值聚类算法的另一个视角：最小化每个类簇的方差

■ 方差：用来计算变量（观察值）与样本平均值之间的差异

$$\arg\min_{G} \sum_{i=1}^{k} \sum_{x \in G_i} ||x - c_i||^2 = \arg\min_{G} \sum_{i=1}^{k} |G_i| \, Var \, G_i$$

第i个类簇的方差: $\mathrm{var}(G_i) = \frac{1}{|G_i|} \sum_{x \in G_i} ||x - c_i||^2$

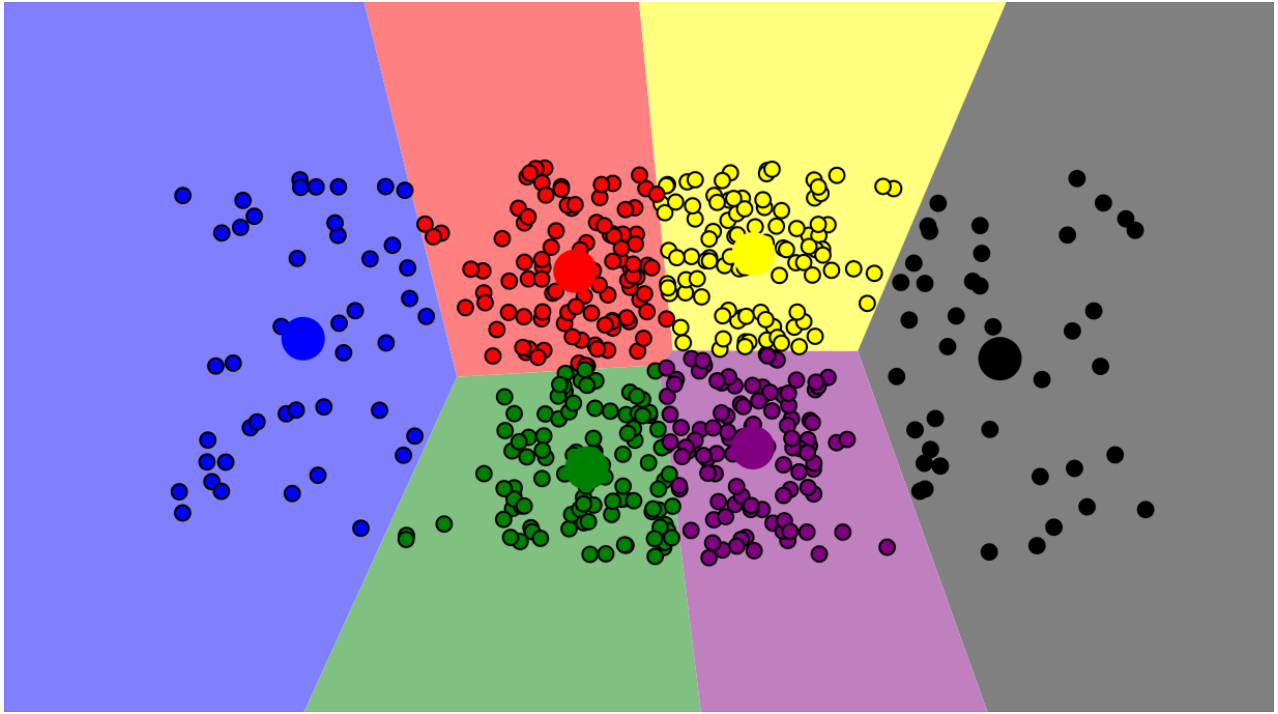
- 欧氏距离与方差量纲相同
- 最小化每个类簇方差将使得最终聚类结果中每个聚类集合中所包含数据呈现出来差异性最小。

K均值聚类算法的不足

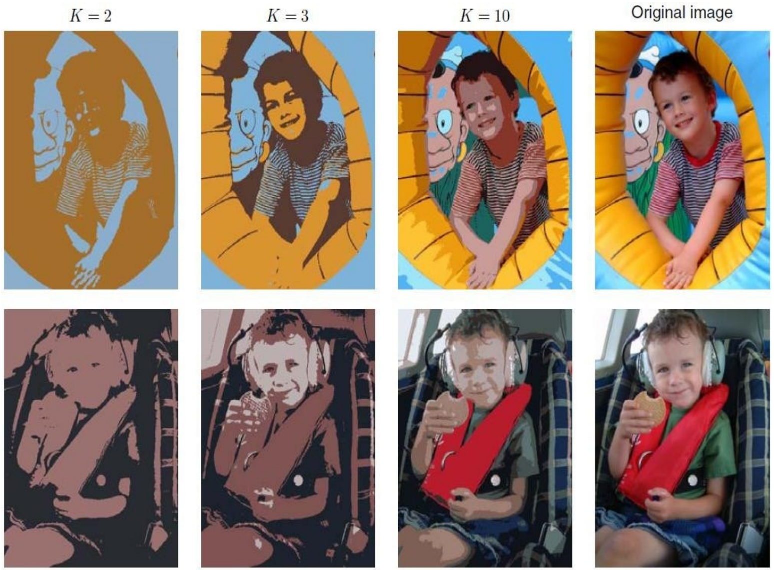
- 需要事先确定聚类数目，很多时候我们并不知道数据应被聚类的数目
- 需要初始化聚类质心，初始化聚类中心对聚类结果有较大的影响
- 算法是迭代执行，时间开销非常大
- 欧氏距离假设数据每个维度之间的重要性是一样的
- 对离群点不够鲁棒

Demo演示

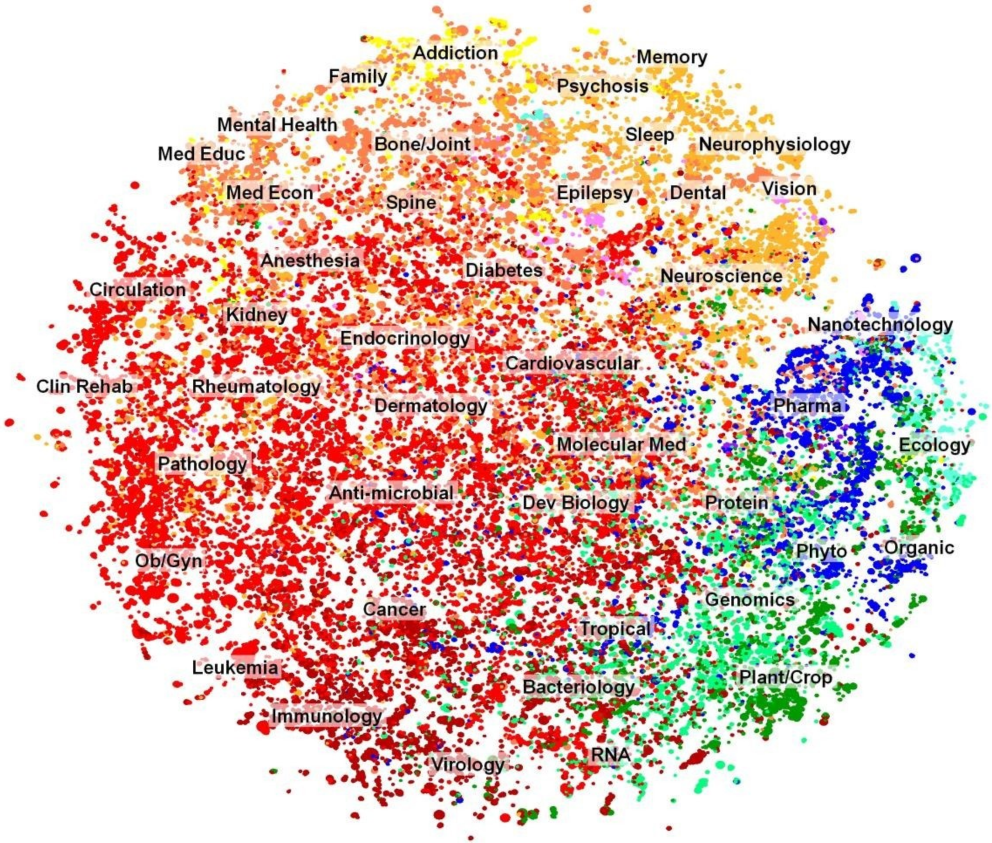
- https://www.naftaliharris.com/blog/visualizing-k-means-clustering/



K均值聚类算法的应用



图像压缩

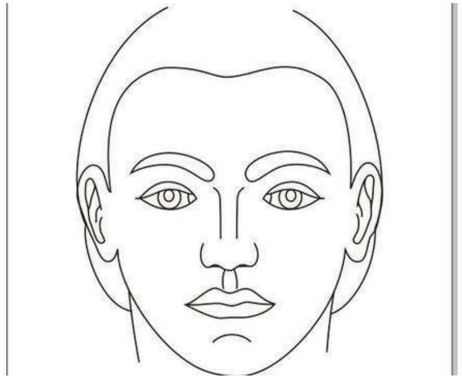


文本分类：将200多万篇论文聚类到29,000个类别，包括化学、工程、生物、传染疾病、生物信息、脑科学、社会科学、计算机科学等及给出了每个类别中的代表单词

目录

主成分分析: Principle Component Analysis （PCA）

- 主成分分析是一种特征降维方法。人类在认知过程中会主动“化繁为简”
- 奥卡姆剃刀定律（Occam’s Razor）：“如无必要，勿增实体”，即“简单有效原理”



主成分分析: 降维后的结果要保持原始数据固有结构

- 原始数据中的结构
  - 图像数据中结构：视觉对象区域构成的空间分布
  - 文本数据中结构：单词之间的(共现)相似或不相似

200万像素点

约减

60个像素点

主成分分析: 若干概念-方差与协方差

数据样本的方差
variance

假设有n个数据，记为$X = \{x_i\}$ $(i = 1, \dots, n)$

- 方差等于各个数据与样本均值之差的平方和之平均数
- 方差描述了样本数据的波动程度

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - u)^2$$

其中u是样本均值， $u = \frac{1}{n}\sum_{i=1}^{n} x_i$

主成分分析: 若干概念-方差与协方差

数据样本的协方差
covariance

假设有n个两维变量数据，记为$(X,Y)=\{(x_i,y_i)\}$ $(i=1,\ldots,n)$

- 衡量两个变量之间的相关度

$$\mathrm{cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-E(X))(y_i-E(Y))$$

其中$E(X)$和$E(Y)$分别是X和Y的样本均值, 分别定义如下

$$\mathrm{E(X)}=\frac{1}{n}\sum_{i=1}^{n}x_i, \qquad \mathrm{E(Y)}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

主成分分析: 协方差例子

编号	x_i	y_i	$x_i - E(X)$	$y_i - E(Y)$	$[x_i - E(X)][y_i - E(Y)]$
1	1	7	-8.33	-16.67	138.89
2	3	11	-6.33	-12.67	80.22
3	6	17	-3.33	-6.67	22.22
4	10	25	0.67	1.33	0.89
5	15	35	5.67	11.33	64.22
6	21	47	11.67	23.33	272.22
	$E(X)$=9.33	$E(Y)$=23.67	$Var(X)$=48.22	$Var(Y)$=192.89	$E([x_i - E(X)][y_i - E(Y)])$ =96.44

$$X = \{x_i\}, Y = \{y_i\}$$

主成分分析: 协方差例子

- 对于一组两维变量（如广告投入-商品销售、天气状况-旅游出行等），可通过计算它们之间的协方差值来判断这组数据给出的两维变量是否存在关联关系：
- 当协方差$cov(X,Y) > 0$时，称X与Y正相关
- 当协方差$cov(X,Y) < 0$时，称X与Y负相关
- 当协方差$cov(X,Y) = 0$时，称X与Y不相关（线性意义下）

主成分分析: 从协方差到相关系数

我们可通过**皮尔逊相关系数**（Pearson Correlation coefficient ）将两组变量之间的关联度规整到一定的取值范围内。皮尔逊相关系数定义如下：

$$\mathrm{corr}(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

编号	x_i	y_i	$x_i - E(X)$	$y_i - E(Y)$	$[x_i - E(X)][y_i - E(Y)]$	$\mathrm{corr}(X,Y)$
1	1	7	-8.33	-16.67	-16.67	1.0 $y_i = 2 \times x_i + 5$
2	3	11	-6.33	-12.67	-12.67	
3	6	17	-3.33	-6.67	-6.67	
4	10	25	0.67	1.33	1.33	
5	15	35	5.67	11.33	11.33	
6	21	47	11.67	23.33	23.33	
	$E(X)$=9.33	$E(Y)$ =23.67	$Var(X)$=48.22	$Var(Y)$=192.89	$E([x_i - E(X)][y_i - E(Y)])$ =96.44	

主成分分析: 从协方差到相关系数

皮尔逊相关系数所具有的性质如下：

- $|corr(X,Y)| \le 1$
- $corr(X,Y) = 1$的充要条件是存在常数a和b,使得$Y = aX + b$
- 皮尔逊相关系数是对称的，即$corr(X,Y) = corr(Y,X)$

- 由此衍生出如下性质：皮尔逊相关系数刻画了变量X和Y之间线性相关程度，如果$|corr(X,Y)|$的取值越大，则两者在线性相关的意义下相关程度越大。$|corr(X,Y)| = 0$表示两者不存在线性相关关系（可能存在其他非线性相关的关系）。

- 正线性相关意味着变量X增加的情况下，变量Y也随之增加；负线性相关意味着变量X减少的情况下，变量Y随之增加。

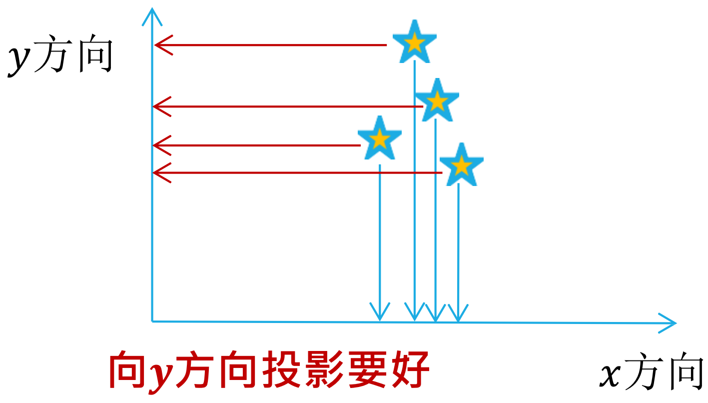
主成分分析: 从协方差到相关系数

- 相关性(correlation)与独立性(independence)
  - 如果$\boldsymbol{X}$和$\boldsymbol{Y}$的线性不相关，则$|corr(X,Y)| = 0$
  - 如果$\boldsymbol{X}$和$\boldsymbol{Y}$的彼此独立，则一定$|corr(X,Y)| = 0$，且$\boldsymbol{X}$和$\boldsymbol{Y}$不存在任何线性或非线性关系
  - “不相关”是一个比“独立”要弱的概念，即独立一定不相关，但是不相关不一定相互独立（可能存在其他复杂的关联关系）。独立指两个变量彼此之间不相互影响。

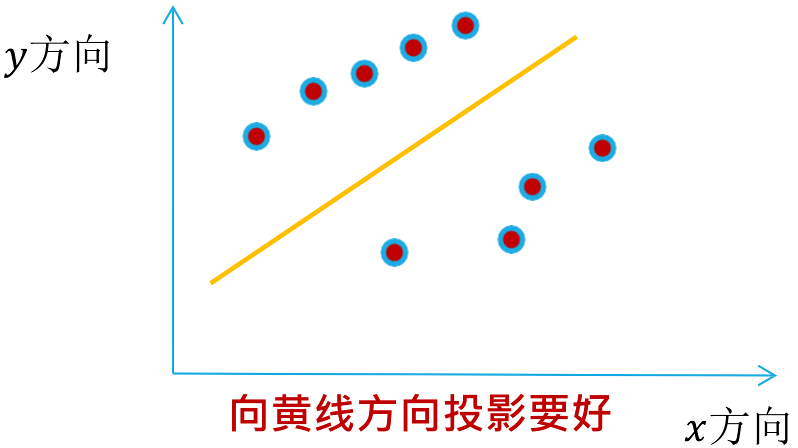
主成分分析: 算法动机

保证样本
投影后方差最大

- 在数理统计中，方差被经常用来度量数据和其数学期望（即均值）之间偏离程度，这个偏离程度反映了数据分布结构。
- 在许多实际问题中，研究数据和其均值之间的偏离程度有着很重要的意义。
- 在降维之中，需要尽可能将数据向方差最大方向进行投影，使得数据所蕴含信息没有丢失，彰显个性。如左下图所示，向y方向投影（使得二维数据映射为一维）就比向x方向投影结果在降维这个意义上而言要好；右下图则是黄线方向投影要好。



向y方向投影要好



向黄线方向投影要好

主成分分析: 算法动机

- 主成分分析思想是将n维特征数据映射到l维空间（$n \gg l$），去除原始数据之间的冗余性（通过去除相关性手段达到这一目的）。
- 将原始数据向这些数据方差最大的方向进行投影。一旦发现了方差最大的投影方向，则继续寻找保持方差第二的方向且进行投影。
- 将每个数据从n维高维空间映射到l维低维空间，每个数据所得到最好的k维特征就是使得每一维上样本方差都尽可能大。

主成分分析: 算法描述

- 假设有n个d维样本数据所构成的集合$D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$，其中$\boldsymbol{x}_i (1 \le i \le n) \in R^d$。
- 集合D可以表示成一个$n \times d$ 的矩阵$\mathbf{X}$。
- **假定每一维度的特征均值均为零（已经标准化）。**
- 主成分分析的目的是求取一个且使用一个$d \times l$的映射矩阵$\mathbf{W}$。
- 给定一个样本$\boldsymbol{x}_i$，可将$\boldsymbol{x}_i$ 从d维空间如下映射到l 维空间：$(\boldsymbol{x}_i)_{1 \times d} (\mathbf{W})_{d \times l}$
- 将所有降维后数据用$\mathbf{Y}$ 表示，有 $\mathbf{Y} = \mathbf{X}\,\mathbf{W}$

？如何求取映射矩阵$\mathbf{W}$

降维结果 原始数据 映射矩阵

- $\mathbf{Y} = n \times l$
- $\mathbf{X} = n \times d$
- $\mathbf{W} = d \times l$

主成分分析: 算法描述

$$\mathbf{Y} = n \times l \quad \mathbf{X} = n \times d \quad \mathbf{W} = d \times l$$

降维后n个l维样本数据$\mathbf{Y}$的方差为:

$$\mathrm{var}(\mathbf{Y}) = \frac{1}{n-1}\mathrm{trace}\left(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\right)$$

$$= \frac{1}{n-1}\mathrm{trace}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{W}\right)$$

$$= trace(\mathbf{W}^{\mathrm{T}}\frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{W})$$

降维前n个d维样本数据$\mathbf{X}$的协方差矩阵记为:

$$\boldsymbol{\Sigma} = \frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{X}$$

主成份分析的求解目标函数为

$$\max_{\mathbf{W}} trace(\mathbf{W}^{T}\boldsymbol{\Sigma}\mathbf{W})$$

满足约束条件

$$\mathbf{w}_i{}^{T}\mathbf{w}_i = 1 \quad i \in \{1,2,\ldots,l\}$$

主成分分析: 算法描述

$$\frac{\partial \mathbf{x}^T \mathbf{B} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{B} + \mathbf{B}^T)\mathbf{x}$$

所有带约束的最优化问题，可通过拉格朗日乘子法将其转化为无约束最优化问题

$$\mathbf{Y} = n \times l \quad \mathbf{X} = n \times d \quad \mathbf{W} = d \times l$$

主成份分析求解目标函数为

$$\max_{\mathbf{W}} trace(\mathbf{W}^T \mathbf{\Sigma} \mathbf{W})$$

$$= \max_{\mathbf{W}} \sum_{i=1}^{l} \mathbf{w}_i^T \mathbf{\Sigma} \mathbf{w}_i$$

拉格朗日函数

$$L(\mathbf{W}, \boldsymbol{\lambda}) = \max_{\mathbf{W}} \sum_{i=1}^{l} \mathbf{w}_i^T \mathbf{\Sigma} \mathbf{w}_i - \lambda_i \sum_{i=1}^{l} (\mathbf{w}_i^T \mathbf{w}_i - 1)$$

其中$\lambda_i (1 \le i \le l)$为拉格朗日乘子，$\mathbf{w}_i$为矩阵$\mathbf{W}$第$i$列。

满足约束条件

$$\mathbf{w}_i^T \mathbf{w}_i = 1 \quad i \in \{1, 2, \dots, l\}$$

投影向量是单位向量（关心方向）

对上述拉格朗日函数中变量$\mathbf{w}_i$求偏导并令导数为零，得到

$$\mathbf{\Sigma}\, \mathbf{w}_i = \lambda_i\, \mathbf{w}_i$$

上式表明：每一个$\mathbf{w}_i$都是n个d维样本数据$\mathbf{X}$的协方差矩阵$\mathbf{\Sigma}$的一个特征向量，λ_i是这个特征向量所对应的特征值。

主成分分析: 算法描述

$\mathbf{Y} = n \times l \quad \mathbf{X} = n \times d \quad \mathbf{W} = d \times l$

$$\mathbf{\Sigma}\,\mathbf{w}_i = \lambda_i\,\mathbf{w}_i，\ 且 trace(\mathbf{W}^T\mathbf{\Sigma}\mathbf{W}) = \sum_{i=1}^{l} \mathbf{w}_i^T\mathbf{\Sigma}\mathbf{w}_i = \sum_{i=1}^{l} \lambda_i$$

- 可见，在主成份分析中，最优化的方差等于原始样本数据$\mathbf{X}$的协方差矩阵$\mathbf{\Sigma}$的特征根之和。
- 要使方差最大，我们可以求得协方差矩阵$\mathbf{\Sigma}$的特征向量和特征根，然后取前l个最大特征根所对应的特征向量组成映射矩阵$\mathbf{W}$即可。
- 注意，每个特征向量$\mathbf{w}_i$与原始数据$\boldsymbol{x}_i$的维数是一样的，均为d。

主成分分析: 算法描述

$\mathbf{Y} = n \times l \quad \mathbf{X} = n \times d \quad \mathbf{W} = d \times l$

- 输入： n个d维样本数据所构成的矩阵$\mathbf{X}$，降维后的维数l
- 输出：映射矩阵$\mathbf{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_l\}$
- 算法步骤：

1：对于每个样本数据$\boldsymbol{x}_i$进行中心化处理： $\boldsymbol{x}_i = \boldsymbol{x}_i - \mu$, $\mu = \frac{1}{n}\sum_{j=1}^{n} x_j$

2: 计算原始样本数据的协方差矩阵：$\boldsymbol{\Sigma} = \frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{X}$

3: 对协方差矩阵$\boldsymbol{\Sigma}$进行特征值分解，对所得特征根按其值大到小排序$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_l$

4: 取前l个最大特征根所对应特征向量$\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_l$组成映射矩阵$\mathbf{W}$

5：将每个样本数据$\boldsymbol{x}_i$按照如下方法降维：$(\boldsymbol{x}_i)_{1\times d}(\mathbf{W})_{d\times l} = 1 \times l$

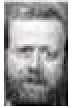
目录

特征人脸方法: 动机

- 特征人脸方法是一种应用主成份分析来实现人脸图像降维的方法，其本质是用一种称为“特征人脸(eigenface)”的特征向量按照线性组合形式来表达每一张原始人脸图像，进而实现人脸识别。
- 由此可见，这一方法的关键之处在于如何得到特征人脸。

用（特征）人脸表示人脸，而非用像素点表示人脸

特征人脸方法: 算法描述



$$\Rightarrow \begin{bmatrix}45 & \cdots & 68\\ \vdots & \ddots & \vdots \\ 36 & \cdots & 86\end{bmatrix}_{32\times32=1024} \Rightarrow \begin{bmatrix}45\\ \cdots\\ \cdots\\ 68\\ \cdots\\ \cdots\\ \cdots\\ 86\end{bmatrix}_{1024\times1}$$

- 将每幅人脸图像转换成列向量
- 如将一幅32×32的人脸图像转成1024×1的列向量

特征人脸: 算法描述

$\mathbf{Y} = n \times l \quad \mathbf{X} = n \times d \quad \mathbf{W} = d \times l$

- 输入： n个1024维人脸样本数据所构成的矩阵$\mathbf{X}$，降维后的维数l
- 输出：映射矩阵$\mathbf{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_l\}$ (其中每个$\boldsymbol{w}_j(1 \le \mathrm{j} \le l)$是一个特征人脸)
- 算法步骤：

1：对于每个人脸样本数据$\boldsymbol{x}_i$进行中心化处理： $\boldsymbol{x}_i = \boldsymbol{x}_i - \mu$, $\mu = \frac{1}{n}\sum_{j=1}^{n} x_j$

2: 计算原始人脸样本数据的协方差矩阵：$\boldsymbol{\Sigma} = \frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{X}$

3: 对协方差矩阵$\boldsymbol{\Sigma}$进行特征值分解，对所得特征根从到小排序$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_d$

4: 取前l个最大特征根所对应特征向量$\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_l$组成映射矩阵$\mathbf{W}$

5：将每个人脸图像$\boldsymbol{x}_i$按照如下方法降维：$(\boldsymbol{x}_i)_{1\times d}(\mathbf{W})_{d\times l} = 1 \times l$

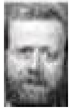
特征人脸: 算法描述

- 每个人脸特征向量$\mathbf{w}_i$与原始人脸数据$\boldsymbol{x}_i$的维数是一样的，均为1024。
- 可将每个特征向量还原为32 × 32的人脸图像，称之为特征人脸，因此可得到l个特征人脸。

400个人脸（左）和与之对应的36个特征人脸

基于特征人脸的降维

- 将每幅人脸分别与每个特征人脸做矩阵乘法，得到一个相关系数
- 每幅人脸得到l个相关系数⇒每幅人脸从1024维约减到l维



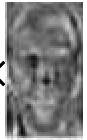
$$\begin{bmatrix}\cdots\\\cdots\\\cdots\\\cdots\\\cdots\\\cdots\\\cdots\\\cdots\end{bmatrix}^{\mathrm{T}}_{1024\times1}$$

$$\begin{bmatrix}\cdots\\\cdots\\\cdots\\\cdots\\\cdots\\\cdots\\\cdots\\\cdots\end{bmatrix}_{1024\times1}$$

每幅人脸图像与特征人脸做矩阵乘法得到一个相关系数

基于特征人脸的降维

- 由于每幅人脸是所有特征人脸的线性组合，因此就实现人脸从“像素点表达”到“特征人脸表达”的转变。每幅人脸从1024维约减到l维。

 $=\ \alpha_{i1}\times$  $+\alpha_{i2}\times$  $+\cdots\ +\alpha_{il}\times$ 

$\boldsymbol{x}_i$

使用l个特征人脸的线性组合来表达原始人脸数据$\boldsymbol{x}_i$

 ⇨ $(\alpha_{i1},\ldots\alpha_{il})$

$\boldsymbol{x}_i$的像素点空间表达32×32

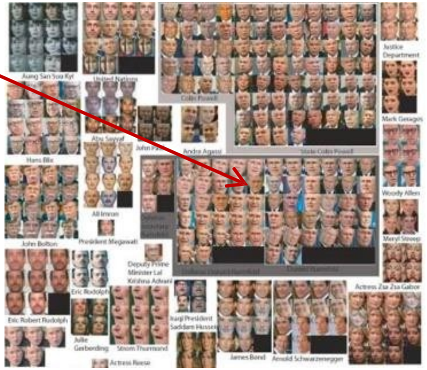
$\boldsymbol{x}_i$的人脸子空间的l个系数表达

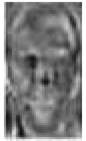
在后续人脸识别分类中，就使用这l个系数来表示原始人脸图像。即计算两张人脸是否相似，不是去计算两个32×32矩阵是否相似，而是计算两个人脸所对应的l个系数是否相似

人脸表达的方法对比：聚类、主成份分析、非负矩阵分解

$\boldsymbol{x}_i$

聚类表示：用待表示人脸最相似的聚类质心来表示



 $=\ \alpha_{i1}\times$  $+\alpha_{i2}\times$  $+\cdots\ +\alpha_{il}\times$ 

$\boldsymbol{x}_i$

特征人脸表示：使用l个特征人脸的线性组合来表达原始人脸数据$\boldsymbol{x}_i$

$\boldsymbol{x}_i$

非负矩阵人脸分解方法表示：通过若干个特征人脸的线性组合来表达原始人脸数据 $\boldsymbol{x}_i$，体现了“部分组成整体”

Daniel D. Lee & H. Sebastian Seung, Learning the parts of objects by non-negative matrix factorization, 1999, Nature

人脸表达后的分析与处理