主成分分析PCA

Principle Component Analysis is popular, famous

是一种**特征降维方法**:
把高维数据投影到低维空间。

> 奥卡姆剃刀：如无必要，勿增实体

降维的结果要保持**原有结构**

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

  关于共现：https://blog.csdn.net/tian_panda/article/details/81127034

1. 有关统计的术语

1. **方差**，一维的
   n 个数据：$X=\{x_1,...x_n\} var(X)=1/n\sum_{i=1}^{n}(x_i-u)^2$(u是样本均值)
2. **协方差**，定义在n维数据上的
   衡量俩变量之间的相关度
   以二维为例:
   n 个2维变量数据：$(X,Y)=\{(x_i,y_i),\ldots,(x_n,y_n)\}$
   $cov(X,Y)=1/n\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$
   判断相关性:
   1. 当cov>0，正相关；
   2. 当cov<0，负相关；
   3. 当cov=0，不相关（线性）

2. pearson相关系数

Pearson相关系数可以把两组变量之间的关联度（协方差可以算出）规整到一定的取值范围内。$[-1,1]$

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

例子:

编号	x_i	y_i	$x_i-E(X)$	$y_i-E(Y)$	$[x_i-E(X)][y_i-E(Y)]$	$corr(X,Y)$
1	1	7	-8.33	-16.67	-16.67	1.0
2	3	11	-6.33	-12.67	-12.67	
3	6	17	-3.33	-6.67	-6.67	$y_i=2\times x_i+5$
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	

> 上图的**相关系数**是1，这表示相关性很强。

1. Pearson相关系数的性质

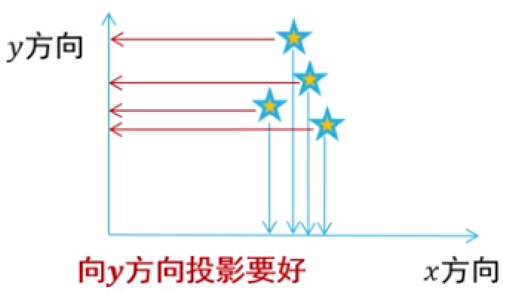
1. $|corr(X,Y)| \leq 1$
2. $corr(X,Y) = 1 \Leftrightarrow \exists a,b \quad Y = aX + b$
3. $corr(X,Y) = corr(Y,X)$
4. $|corr(X,Y)|$越大，说明二者的相关程度越大，=0，那么不存在线性相关的关系.
5. 正线性相关：X增→Y增

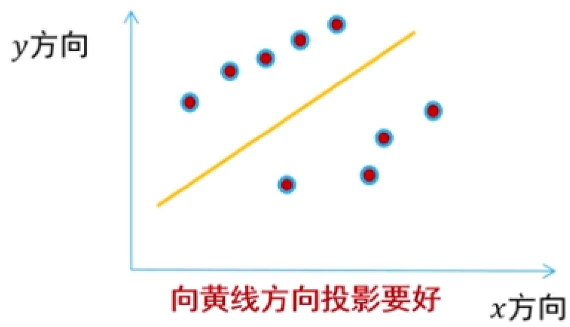
2. 相关性与独立性

- X和Y线性不相关，则$corr(X,Y) = 0$
- X和Y独立，一定$corr(X,Y) = 0$，而且X和Y啥关系都没有
- **不相关比独立要弱**，独立一定不相关，但是不相关不一定独立（可能有其他的复杂关系）

3. 算法动机

保持结构不变（去除冗余性），就是要把方差小的方向忽略掉，**尽量向方差最大的方向投影。**





> 如图所示，投影完之后，每个样本点尽量彼此离散。

- 要将 n 维投影到 l 维，先向方差最大的维度投影
- 然后向方差第二的维度投影......

4. 算法描述

1. 理论介绍

条件

有 n 个 d 维的样本数据，$D = \{x_1, ..., x_n\}$，$x_i \in R^d$ 。 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都是0（已经标准化）。

目的

是求取且使用一个 $d \times l$ 的映射矩阵 $\mathbf{W}$ 。有了这个矩阵，就能把**给定的 d 维的 x 映射到 l 维空间。**

降维后的数据用 $n \times l$ 的矩阵 $\mathbf{Y}$ 表示，$\mathbf{Y} = \mathbf{XW}$。

Y的方差和正交性

1. 我们希望降维以后方差最大，所以就计算方差。

   > 第一行式子在最后有解释。

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

   降维之前的矩阵 $\mathbf{X}$ 的协方差矩阵$\Sigma = \frac{1}{n-1}\mathbf{X^TX}$:

   那么代入上式可得$var(\mathbf{Y}) = trace(\mathbf{W}^T\Sigma\mathbf{W})$

而我们希望找到一个矩阵 $\mathbf{W}$，使得 $trace(\mathbf{W}^T\Sigma\mathbf{W})$ 最大。

2. 同时，还要一个条件，就是 $\mathbf{W}$ 需要满足：对于 $\mathbf{W}$ 中的任意一列 $\mathbf{w}_i$ ，都有$\mathbf{w}_i^T\mathbf{w}_i=1$ ，这是为了让所得到的的映射结果相关性更小（**正交性**），因为两个维度相关性大的话就意味着这二者存在冗余。

拉格朗日函数

又要方差大，又要相关性小。

$$L(W,\lambda)=trace(W^T\Sigma W)-\sum_{i=1}^{l}\lambda_i(w_i^Tw_i-1)$$

λ_i 是拉格朗日乘子

对上述函数中的w_i求偏导，并且令导数为0，得：

$$\Sigma w_i=\lambda_i w_i$$

为什么一个矩阵乘以一个向量等于一个常数乘以一个向量呢？这说明了： w_i 是 Σ 的一个特征向量，而 λ_i 是这个特征向量对应的特征值。

2. 算法实现

1. 算法的输入和输出

- input：X, l
- output：W={w_1,...,w_l}

2. 算法步骤

1. 中心化处理，把平均值搞成0
2. 计算$\Sigma=1/(n-1)X^TX$
3. 对Σ进行特征值分解，将其特征根 λ 按照从大到小排序，有d个
4. 取前 l 个最大的特征根 λ 对应的特征向量 w ，组成映射矩阵 $\mathbf{W}$
5. 把每个样本数据 x 都用 $\mathbf{W}$ 来降维

> **关于公式第一行的解释**
>
> Y的转置乘以Y，结果中把对角线的元素取出来，结果就是矩阵Y的方差。
>
> 原因：已经标准化了，均值就是0。
>
> 假设Y是下式，也就是说，有3个样本数据，每个数据4维
>
> $$\begin{bmatrix}1&2&3&4\\4&5&6&6\\7&8&9&6\end{bmatrix}$$
>
> Y^T：
>
> $$\begin{bmatrix}1&4&7\\2&5&8\\3&6&9\\4&6&6\end{bmatrix}$$
>
> Y^TY出来的结果是一个3x3的矩阵

> $$\begin{bmatrix} o & . & . \\ . & o & . \\ . & . & o \end{bmatrix}$$
>
> 那么左上角就是Y第一列的平方和，也就是此时数据第一维度的方差。