无监督学习

KMeans

- [ ] 过一遍KMeans的流程

基于质心

sensitive to outliers

- 特别大的值会大幅扰乱数据分布

改进：K-mediods

k-mediods 每次选取的质心，必须是样本点，
而 k-means每次选取的质心可以是样本点之外的点

K-Medoids 的重点，就在于中心点的选取。

> 选取簇中心点的准则函数是 : **当前簇中所有其他点到该中心点的距离之和最小**
> 需要遍历簇中所有点

PCA

皮尔逊相关系数：刻画变量之间线性相关程度

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

性质：

- 绝对值不大于1
- 绝对值=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
- 取值越大，二者在线性相关的意义下相关程度越大

将d维特征数据映射到l维空间，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（保证方向的正交性）

- 最优化的方差等于原始样本数据X的协方差矩阵$\sum$的特征根之和
- 要使方差最大，求协方差矩阵$\sum$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W

算法步骤：

1. 对每个样本数据xi进行中心化处理
2. 计算原始样本数据的协方差矩阵：$\sum = \frac{1}{n-1}X^TX$
3. 对协方差矩阵∑进行特征值分解，对所得特征根按其值大到小排序λ1 >= λ2 > = ... λd
4. 取前l个最大特征根所对应特征向量w1, w2, ..., wl组成映射矩阵W

5. 将每个样本数据xi按照如下方法降维：xi(1*d) W(d*l) -----> 1* l

- [ ] PCA流程

线性区别分析(LDA)

类内散度矩阵：K个类，每个类中的样本-均值之和

类间散度矩阵

非负矩阵分解(NMF)

$D=WH$ 将D分解成非负的W,H

利用任意一个可逆矩阵A以及他的逆矩阵A^{-1}，可得$WH=WAA^{-1}H$。很容易找到A使得$\hat{W}=WA$和$\hat{H}=A^{-1}H$均非负。这样就找到另一种分解，$D=\hat{W}\hat{H}$

两种代价函数：

- 欧氏距离平方函数

$$\|D-WH\|_F^2=\sum_{0\le i<n,0\le j<d}\left(D_{i,j}-(WH)_{i,j}\right)^2$$

- KL散度函数

$$D_{KL}(D||WH)=\sum_{0\le i<n,0\le j<d}D_{i,j}\,log\frac{D_{i,j}}{(WH)_{i,j}}-D_{i,j}+(WH)_{i,j}$$

对应的两种迭代方式

书p168

> 策略：
>
> 对于欧氏距离平方函数，保证分子分母的形状和目标矩阵相同，分子有D，分母无D
>
> 对于KL散度的那种迭代，感觉不太对

多维尺度法(MDS)

保持原始数据之间两两距离不变，**无法对新数据集合进行降维**

方法：

1. 计算距离矩阵A
2. $H=I-\frac{1}{n}l_nl_n^T$
3. $B=-\frac{1}{2}HAH$
4. $B=\Gamma\Lambda\Gamma^T$
5. 其中只取前l个非零特征根，前面的是n*l矩阵，每一列都是特征根对应的特征向量

于是$X = [x_1, \ldots, x_i, \ldots, x_n]$中$d$维原始数据可通过如下变换转变为$l$维数据，从而实现降维：

6. $$\bar{X}^T = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n]^T = \Gamma \Lambda^{1/2}$$

其中$\bar{x}_i (1 \le i \le n)$表示原始数据$x_i$降维后的结果。

局部线性嵌入（LLE）

LLE假设数据在较小的局部是线性的：某一个数据可以由它邻域中的几个样本来线性表示。

线性关系只在样本的附近起作用（所以用了K近邻）

ppt中的x是原始数据，y是其到低维的投影，我们希望其仍能够保持线性关系，也就是先前在x上应用的w权重我们希望仍能应用到y上，即对y的均方差损失函数最小，因此二者损失函数有着一样的形式

区别是高维的式子中，高维数据已知，目标是求最小值对应的权重系数W,而我们在低维是权重系数W已知，求对应的低维数据。注意，这里的W已经是m×m维度，之前的W是m×k维度，我们将那些不在邻域位置的W的位置取值为0，将W扩充到m×m度。

ppt(4)中的对降维结果增加一定约束，是指让y标准化

$M=(I-W)^T(I-W)$，使用拉格朗日乘子，可得$MY^T=\lambda Y^T$

最终通过拉格朗日乘子法，取矩阵M**最小的前k个非零特征根**所对应向量构成的矩阵

https://www.cnblogs.com/pinard/p/6266408.html

特征人脸方法

基于PCA

Y(n * l) = X (n * d) * W (d * l)

X的协方差矩阵，求取前l个特征值，并用特征向量组成W

x*w 获取到降维的人脸 d -> l

w(d*l)   可以变成l个特征人脸

> 降维后的人脸是指Y，人脸数目不变，但是维数变少
>
> 特征人脸实际上是W，L个人脸图片 => L个特征人脸

相关系数的获得：每个原来的人脸图片都与每个特征人脸（W的一维）做矩阵乘法得到一个相关系数

(1 * d) × (d * l) = (1 * l) 总共获得l个相关系数

在后续人脸识别分类中，计算两张人脸是否类似，是看各自的这l个相关系数是否相似

> 6.16：
>
> l个相关系数的取得：每个原来的人脸图片都与W(d*l)相乘，获得(1*l)的向量，每个值就是一个相关系数
>
> l个相关系数指的是l维向量，相比原始的d其数目会大减，达成了降维的效果

奇异值分解（svd）

核心就是求取左右的矩阵U和V，该过程可以通过$A^TAV = \lambda V$的形式来求

$AA^T = \lambda U$ 求U

U和V乘自身转置都是单位阵

奇异值的获取：任一种特征值开根号，特征值组成的矩阵是$\sum^T \sum$，但是只有对角线上有元素，对应的就是奇异值的平方

> 6.16:
>
> 非负奇异值的个数：min(m, n)
>
> 过程是从A出发，但因为A不是方阵，所以要写成A^TA和AA^T的方阵形式，分别求出右奇异矩阵V和左奇异矩阵U，这二者分别代表右奇异向量和左奇异向量的合并。其对应的特征值能拼成一个矩阵，矩阵对角线为所有特征值，该矩阵即是$\Sigma\Sigma^T$或$\Sigma^T\Sigma$，奇异值便是特征值开根号

- [ ] 奇异值分解实现特征人脸

潜在语义分析

1. 计算单词——文档矩阵
2. 奇异值分解/非负矩阵分解
3. 重建矩阵
4. 挖掘语义关系
   计算任意两个文档之间的皮尔逊相关系数，得到文档——文档相关系数矩阵

期望最大化算法

- 最大似然估计
- 最大后验估计：与最大似然估计相比，增加了一项与θ相关的先验概率$P(\theta)$。

无论上面哪种做法，都是充分利用已有数据，在参数模型确定（只是参数值未知）情况下，对所优化目标中的参数求导，令导数为0，求取模型的参数值

隐变量

计算隐变量（EM算法的E步），最大化似然函数和更新模型参数（EM算法的M步）

p104

$$
\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \\
\mathrm{KL}(q \| p) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}
\end{aligned}
$$

- [ ] 自监督学习