模式识别第二次作业提交版

201300086 史浩男 人工智能学院

一、（3.2）K-means

(a)公式抽象

不妨假设μ_i为聚类中心，我们的目标就是把数据点根据到中心距离分类，形式化目标函数如下：

$$D=\sum_{\mathrm{j}=1}^{\mathrm{M}}\sum_{\mathrm{i}=1}^{\mathrm{K}}\gamma_{\mathrm{ij}}\|\boldsymbol{x}_{\mathrm{j}}-\boldsymbol{\mu}_{\mathrm{i}}\|^2 \tag{1}$$

因此只需要最小化这个目标函数，就能求出对应的γ_{ij}和μ_{i}，从而完成聚类

$$\underset{\gamma_{\mathrm{ij}},\boldsymbol{\mu}_{\mathrm{i}}}{\arg\min}\sum_{\mathrm{i}=1}^{\mathrm{K}}\sum_{\mathrm{j}=1}^{\mathrm{M}}\gamma_{\mathrm{ij}}\|\boldsymbol{x}_{\mathrm{j}}-\boldsymbol{\mu}_{\mathrm{i}}\|^2 \tag{2}$$

(b)迭代规则

固定μ_{i}：

此时只需找到每个数据点距离最近的中心是哪个，所有中心都是不变的，因此表达式为：

$$\gamma_{ij}=\begin{cases}1, & \text{if } i=\arg\min_i\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|^2\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

固定γ_{ij}：

此时所有类别包含哪些点已经确定，只需在每个类中找到类中点最近的中心位置，可以解出：

$$\boldsymbol{\mu}_i=\frac{\sum_{j=1}^M\gamma_{ij}\boldsymbol{x}_j}{\sum_{j=1}^M\gamma_{ij}} \tag{4}$$

(c)证明收敛

只需证明，（b）中的两个迭代步骤，都会使目标函数D不增（单调递减有下界的函数必然收敛。而如果两个步骤都不增不减，说明已经收敛。如果至少有一个是递减的，那么满足条件单调递减有下界，一定会最终收敛）

固定μ_{i}：

$$\begin{aligned}D'-D&=\sum_{i=1}^K\sum_{j=1}^M\gamma'_{ij}\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|^2-\sum_{i=1}^K\sum_{j=1}^M\gamma_{ij}\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|^2\\&=\sum_{i=1}^K\sum_{j=1}^M\left(\gamma'_{ij}-\gamma_{ij}\right)\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|^2\\&=\sum_{i=1}^K\left(\|\boldsymbol{x}_j-\boldsymbol{\mu}'_i\|^2-\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|^2\right)\\&\le 0\end{aligned} \tag{5}$$

固定γ_{ij}：

$$\begin{aligned}
D_j' - D_j &= \sum_{i=1}^{K} \gamma_{ij} \|\boldsymbol{x}_j - \boldsymbol{\mu}_i'\|^2 - \sum_{i=1}^{K} \gamma_{ij} \|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 \\
&= \sum_{i=1}^{K} \gamma_{ij} \left(\|\boldsymbol{x}_j - \boldsymbol{\mu}_i'\|^2 - \|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 \right) \\
&= \sum_{i=1}^{K} \gamma_{ij} \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i' + \boldsymbol{x}_j - \boldsymbol{\mu}_i \right)^{\mathrm{T}} \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i' - \boldsymbol{x}_j + \boldsymbol{\mu}_i \right) \\
&= \sum_{i=1}^{K} \gamma_{ij} \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_i' \right)^{\mathrm{T}} \left(2\boldsymbol{x}_j - \boldsymbol{\mu}_i' - \boldsymbol{\mu}_i \right) \\
&= \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_i' \right)^{\mathrm{T}} \left(2 \left(\sum_{i=1}^{K} \gamma_{ij} \boldsymbol{x}_j \right) - \left(\sum_{i=1}^{K} \gamma_{ij} \right) \left(\boldsymbol{\mu}_i' + \boldsymbol{\mu}_i \right) \right) \\
&= \left(\sum_{i=1}^{K} \gamma_{ij} \right) \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_i' \right)^{\mathrm{T}} \left(2 \frac{\sum_{i=1}^{K} \gamma_{ij} \boldsymbol{x}_j}{\sum_{i=1}^{K} \gamma_{ij}} - \boldsymbol{\mu}_i' - \boldsymbol{\mu}_i \right) \\
&= \left(\sum_{i=1}^{K} \gamma_{ij} \right) \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_i' \right)^{\mathrm{T}} \left(2\boldsymbol{\mu}_i' - \boldsymbol{\mu}_i' - \boldsymbol{\mu}_i \right) \\
&\leq 0
\end{aligned} \tag{6}$$

因此我们简洁地证明了，两个迭代步骤都使目标函数D不增。

综上，一定会最终收敛

二、（4.2）LR

(a，b)优化问题与重写

$$\arg\min_{\boldsymbol{\beta}} (y_i - x_i^T \boldsymbol{\beta})^2 \tag{7}$$

矩阵表示重写：

$$\arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{8}$$

(c)求解

$$\frac{\partial E}{\partial \boldsymbol{\beta}} = 2\boldsymbol{X}^{\mathrm{T}} (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}) = \mathbf{0} \tag{9}$$

由于 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$可逆，解出：

$$\boldsymbol{\beta}^* = \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{y} \tag{10}$$

(d)维度大于样本导致不可解

由矩阵的性质可知：$\operatorname{rank}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right) = \operatorname{rank}(\boldsymbol{X}) \leq n < d$ ，而$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$是一个 $d \times d$的矩阵, 不满秩的矩阵必然不可逆

(e)正则化项作用

正则化项度量了模型复杂度，是用于对抗过拟合的关键手段。正则化表示了对模型的一种偏好, 可以对模型的复杂度进行约束, 因此可以在性能相同的模型中，选择出模型复杂度最低的一个。

(f)求解岭回归优化问题

$$\begin{gathered}\arg\min_{\boldsymbol{\beta}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})+\lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} \\ \frac{\partial E}{\partial \boldsymbol{\beta}}=2\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{y})+2\lambda\boldsymbol{\beta}=\mathbf{0}\end{gathered} \tag{11}$$

解得最优

$$\boldsymbol{\beta}^*=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{12}$$

(g)正则化在可逆方面的作用

加入岭回归正则项后, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\lambda\boldsymbol{I}$ 几乎总是可逆, 总能求解，解决了无法获得唯一的模型参数的问题。

同时正则化是用于对抗过拟合的关键手段

(h)极端λ的影响

如果$\lambda=0$, 岭回归退化为普通线性回归
如果$\lambda=\infty$, 则优化问题变为 $\arg\min_{\boldsymbol{\beta}}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}$, 解出 $\boldsymbol{\beta}=\mathbf{0}$

(i)λ为什么必须是超参数

因为正则化项$\lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}$恒正，目标函数中只有正则化项中出现了λ，最优化目标函数时一定会将λ优化为0，失去了正则化的意义

三、（4.5）AUC

(a)

下标	标记	得分	P	R	AUC-PR	AP
0			1	0		
1	1	1	1	0.2	0.2	0.2
2	2	0.9	0.5	0.2	0	0
3	1	0.8	0.67	0.4	0.1167	0.1333
4	1	0.7	0.75	0.6	0.1417	0.15
5	2	0.6	0.6	0.6	0	0
6	1	0.5	0.67	0.8	0.1267	0.1333

下标	标记	得分	P	R	AUC-PR	AP
7	2	0.4	0.57	0.8	0	0
8	2	0.3	0.5	0.8	0	0
9	1	0.2	0.56	1	0.1056	0.111
10	2	0.1	0.5	1	0	0

(b)AP&PR

相似是正常的，而且AP比PR总是稍微大一点点

原因是他们的计算方式只有细微区别：

$$
\begin{aligned}
AP - PR &= \sum_{i=1}^{n}(r_i - r_{i-1})p_i - \sum_{i=1}^{n}(r_i - r_{i-1})\frac{p_i + p_{i-1}}{2} \\
&= \sum_{i=1}^{n}\frac{1}{2}(r_i - r_{i-1})\left(p_i - p_{i-1}\right)
\end{aligned}
\tag{13}
$$

(c)

交换了第 9 行和第 10 行的类别标记之后, AUC-PR=0.6794, AP=0.7167.

(d)

代码：

```
from collections import Counter

v = [1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1]
label = [1, 2, 1, 1, 2, 1, 2, 2, 1, 2]#[1, 2, 1, 1, 2, 1, 2, 2, 2, 1]

P = [1.0]
R = [0.0]
TPR = [0.0]
FPR = [0.0]
for i in range(1, len(v) + 1):
    pos_count = Counter(label[:i])
    neg_count = Counter(label[i:])
    TP = pos_count.get(1, 0)
    FP = pos_count.get(2, 0)
    FN = neg_count.get(1, 0)
    TN = neg_count.get(2, 0)
    P.append(TP / (TP + FP))
    R.append(TP / (TP + FN))
AUC_PR = [0.5 * (R[i] - R[i - 1]) * (P[i] + P[i - 1]) for i in range(1, len(R))]
AP = [(R[i] - R[i - 1]) * P[i] for i in range(1, len(R))]

print('P:', [*P])
```

```
print('R:', [*R])
print('AUC_PR:', [*AUC_PR])
print('AP:', [*AP])
```

四、（4.6）KNN

(a)偏置-方差分解

首先给出误差表达式

$$\mathbb{E}_D[(y-f(\boldsymbol{x};D))] = \mathbb{E}_D\left[(F(\boldsymbol{x})-f(\boldsymbol{x};D)+\epsilon)^2\right] = \mathbb{E}_D\left[(F(\boldsymbol{x})-f(\boldsymbol{x};D))^2\right]+\sigma^2 \tag{14}$$

展开可得

$$\mathbb{E}_D\left[(F(\boldsymbol{x})-f(\boldsymbol{x};D))^2\right] = (\mathbb{E}_D[F(\boldsymbol{x})-f(\boldsymbol{x};D)])^2+\operatorname{Var}(F(\boldsymbol{x})-f(\boldsymbol{x};D)) \tag{15}$$

由于 $F(\boldsymbol{x})$ 是确定的, 与训练集 D 无关, 即 $\mathbb{E}_D[F(\boldsymbol{x})]=F(\boldsymbol{x})$, 则上式进一步简化为:

$$\begin{aligned}
&(\mathbb{E}_D[F(\boldsymbol{x})-f(\boldsymbol{x};D)])^2+\operatorname{Var}(F(\boldsymbol{x})-f(\boldsymbol{x};D)) \\
&=(F(\boldsymbol{x})-\mathbb{E}_D[f(\boldsymbol{x};D)])^2+\operatorname{Var}(f(\boldsymbol{x};D)) \\
&=(F(\boldsymbol{x})-\mathbb{E}_D[f(\boldsymbol{x};D)])^2+\mathbb{E}_D\left[(f(\boldsymbol{x};D)-\mathbb{E}_D[f(\boldsymbol{x};D)])^2\right]
\end{aligned} \tag{16}$$

综上,得到偏置-方差分解

$$\begin{aligned}
&\mathbb{E}_D[(y-f(\boldsymbol{x};D))] \\
=&(F(\boldsymbol{x})-\mathbb{E}_D[f(\boldsymbol{x};D)])^2+\mathbb{E}_D\left[(f(\boldsymbol{x};D)-\mathbb{E}_D[f(\boldsymbol{x};D)])^2\right]+\sigma^2
\end{aligned} \tag{17}$$

(b)带入，缩写

$$\mathbb{E}[f]=\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}y_{nn(i)}\right]=\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[F\left(\boldsymbol{x}_{nn(i)}\right)+\epsilon\right]=\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[F\left(\boldsymbol{x}_{nn(i)}\right)\right] \tag{18}$$

(c)x,y带入f

$$\begin{aligned}
&(F(\boldsymbol{x})-\mathbb{E}_D[f(\boldsymbol{x};D)])^2+\mathbb{E}_D\left[(f(\boldsymbol{x};D)-\mathbb{E}_D[f(\boldsymbol{x};D)])^2\right]+\sigma^2 \\
&=\left(F(\boldsymbol{x})-\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_D\left[F\left(\boldsymbol{x}_{nn(i)}\right)\right]\right)^2+\mathbb{E}_D\left[\frac{1}{k}\left(\sum_{i=1}^{k}y_{nn(i)}-\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_D\left[F\left(\boldsymbol{x}_{nn(i)}\right)\right]\right)^2\right]+\sigma^2 \\
&=\left(F(\boldsymbol{x})-\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_D\left[F\left(\boldsymbol{x}_{nn(i)}\right)\right]\right)^2+\frac{1}{k^2}\mathbb{E}_D\left[\left(\sum_{i=1}^{k}\left(y_{nn(i)}-\mathbb{E}_D\left[F\left(\boldsymbol{x}_{nn(i)}\right)\right]\right)\right)^2\right]+\sigma^2
\end{aligned} \tag{19}$$

(d)方差项与k

$$\frac{1}{k^2}\mathbb{E}_D\left[\left(\sum_{i=1}^{k}\left(y_{nn(i)}-\mathbb{E}_D\left[F\left(\boldsymbol{x}_{nn(i)}\right)\right]\right)\right)^2\right] \tag{20}$$

k增大时，方差项系数变小，找到的最近邻更多，方差整体减小

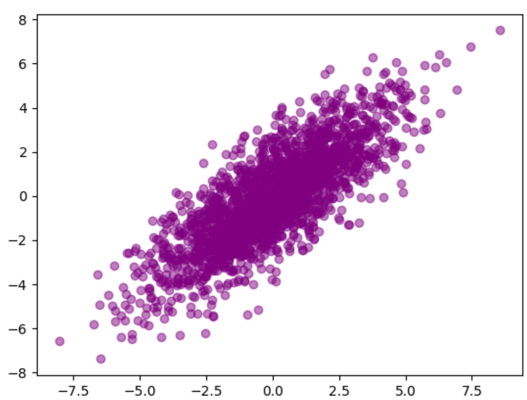
(e)偏差平方项与k

$$\left(F(\boldsymbol{x})-\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_D\left[F\left(\boldsymbol{x}_{nn(i)}\right)\right]\right)^2 \tag{21}$$

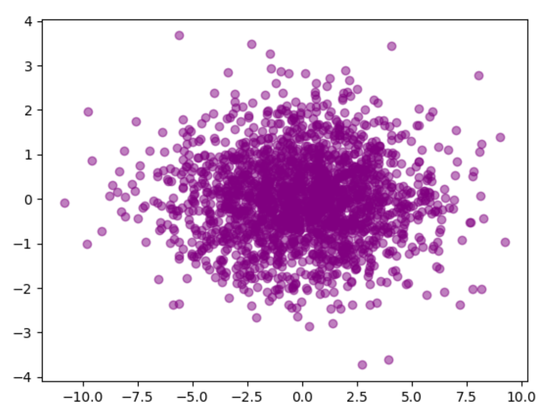
k增大时，这两项的差会越来越大，导致偏差增大。尤其是当k=n时，偏差达到最大，方差达到最小（0）

五、（5.3）编程：PCA&白化

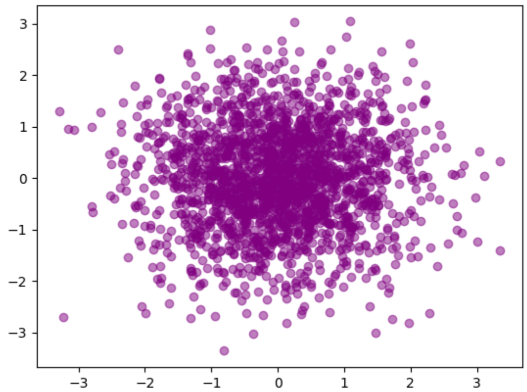
(a)



(b)



(c)



(d)PCA本质是旋转

因为PCA 本质是将数据视作了一个多维空间中的类球形, 并把这个球的各个轴按照各方差最大的方向，旋转对齐到坐标轴上。用数学方式解释，就是把数据乘上一个旋转矩阵，

PCA 旋转这一操作有效的原因： PCA数据降维的本质，就是在对齐到坐标轴上后，把短轴对应纬度去掉, 保留几个长轴对应的维度，进而得到新的降维后数据。由于已经进行旋转对齐，所以去除短轴这一过程很简单，只需比较轴长短即可。

六、（6.3）条件数

(a)矩阵 2-范数$=\sigma_{max}$

矩阵 2-范数等于其最大奇异值，可知 $\|\boldsymbol{X}\|_2=\sigma_1$, 且由矩阵的逆的性质可知 $\left\|\boldsymbol{X}^{-1}\right\|_2=\frac{1}{\sigma_n}$

$$\kappa_2(\boldsymbol{X})=\|\boldsymbol{X}\|_2\left\|\boldsymbol{X}^{-1}\right\|_2=\frac{\sigma_1}{\sigma_n} \tag{22}$$

(b)病态线性系统

我们想要解释的是，在 $\kappa_2(\boldsymbol{A})$很大的情况下, 稍微改变 $\boldsymbol{A}$ 或 $\boldsymbol{b}$ 就会使 $\boldsymbol{x}$有很大的改变.

已知$\Delta\boldsymbol{x}=A^{-1}\Delta\boldsymbol{b}$, $\|\boldsymbol{b}\|\leq\|\boldsymbol{A}\|\|\boldsymbol{x}\|, \|\Delta\boldsymbol{x}\|\leq\left\|\boldsymbol{A}^{-1}\right\|\|\Delta\boldsymbol{b}\|$

相乘再除以 $\|\boldsymbol{b}\|\|\boldsymbol{x}\|$ 可得

$$\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leq\|\boldsymbol{A}\|\left\|\boldsymbol{A}^{-1}\right\|\frac{\|\Delta\boldsymbol{b}\|}{\boldsymbol{b}}=\kappa_2(\boldsymbol{A})\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|} \tag{23}$$

再进行扰动$\Delta\boldsymbol{A}$ 可得

$$\begin{gathered}(\boldsymbol{A}+\Delta\boldsymbol{A})(\boldsymbol{x}+\Delta\boldsymbol{x})=\boldsymbol{b}=\boldsymbol{A}\boldsymbol{x}\\ \boldsymbol{A}\Delta\boldsymbol{x}=-\Delta\boldsymbol{A}(\boldsymbol{x}+\Delta\boldsymbol{x})\\ \Delta\boldsymbol{x}=-\boldsymbol{A}^{-1}\Delta\boldsymbol{A}(\boldsymbol{x}+\Delta\boldsymbol{x})\end{gathered} \tag{24}$$

因此我们使用范数不等式并两边除以$\|\boldsymbol{x}+\Delta\boldsymbol{x}\|$ 有

$$\frac{\|\Delta \boldsymbol{x}\|}{\|\boldsymbol{x}+\Delta \boldsymbol{x}\|} \leq \|\boldsymbol{A}\|\left\|\boldsymbol{A}^{-1}\right\| \frac{\|\Delta \boldsymbol{A}\|}{\|\boldsymbol{A}\|}=\kappa_2(\boldsymbol{A}) \frac{\|\Delta \boldsymbol{A}\|}{\|\boldsymbol{A}\|} \tag{25}$$

由此表达式发钱，即使当有较小的扰动$\Delta \boldsymbol{A}$ 或者 $\Delta \boldsymbol{b}$的时候, 也会带来较大的 $\Delta \boldsymbol{x}$, 小的输入变换就会导致较大的输出变化

这一定程度上说明了病态系统的原因

(c)良态正交矩阵

正交矩阵的逆等于其转置，有相同特征值

$$\kappa_2(\boldsymbol{X})=\|\boldsymbol{W}\|_2\left\|\boldsymbol{W}^{-1}\right\|_2=\|\boldsymbol{W}\|_2\left\|\boldsymbol{W}^{\mathrm{T}}\right\|_2=\left(\|\boldsymbol{W}\|_2\right)^2=1 \tag{26}$$