模式识别与计算机视觉 HW2

姓名：石睿

学号：211300024

所属院系：人工智能学院

1. 习题 1（第三章 3.2）

(a) $\sum_{i=1}^{K}$ 即这 K 组，$\sum_{j=1}^{M} r_{ij}$ 即 K 组中每组中的样本（若 x_j 不在第 i 组，$r_{ij}=0$，不计入考量）

$\|x_j-\mu_i\|^2$ 即样本 x_j 到当前组中心（代表）μ_i 的距离

综上：$\arg\min_{r_{ij}, \mu_i} \sum_{i=1}^{K}\sum_{j=1}^{M} r_{ij}\cdot\|x_j-\mu_i\|^2$ 也即 K 组的所有组内的样本和代表的差异最小化，

也即 k-means 中要求的每组样本彼此相似的优化目标

(b)

i. 由于 $r_j = \begin{pmatrix} r_{1j} \\ \vdots \\ r_{kj} \end{pmatrix}$ 和 $r_{j'} = \begin{pmatrix} r_{1j'} \\ \vdots \\ r_{kj'} \end{pmatrix}$ 相互独立，上述优化式在固定 μ_i 时可以化为 M 个 r_j (i∈[… 单独的优化问题

也即 $\begin{cases} \min_{r_{ij}} \sum_{i=1}^{K} r_{ij}\cdot\|x_j-\mu_i\|^2, \ \forall j\in[M] \\ s.t. \ \sum_{i=1}^{K} r_{ij}=1 \end{cases}$ 由其实际语义可知其最优解为

$$r_{ij}^* = \begin{cases} 1, & i=\arg\min_i \|x_j-\mu_i\|_2^2 \\ 0, & 否则 \end{cases}$$

ii. 由于原优化式中 μ_i 无交叉项，也即原式可单独优化 K 个 μ_i

故优化问题变为 $\min_{\mu_i} \sum_{j=1}^{M} r_{ij}\cdot\|x_j-\mu_i\|^2, \forall i\in[K]$，由其语义可知即"找 μ_i 和实际在 i 组中… 样本距离最小"

$$J \equiv \min_{\mu_i} \sum_{j=1}^{M} r_{ij}\left(x_j^T x_j - 2\mu_i^T\cdot x_j + \mu_i^T\cdot\mu_i\right)$$

令 $\frac{\partial J}{\partial \mu_i} = \sum_{j=1}^{M} r_{ij}\cdot 2\mu_i - \sum_{j=1}^{M} 2x_j = 0$

$$\therefore \mu_i^* = \frac{\sum_{j=1}^{M} x_j}{\sum_{j=1}^{M} r_{ij}}$$

c) Lloyd 算法是在不断优化样本 x_j 分配到 i 组的情况，而聚类的整体分布至多有 K^M 个（每个样本分配到不同组时），故最差情况经 K^M+1 次一定收敛。

详细说明：

(1) 若 r_{ij} 和 μ_i 均不变化，则收敛

(2) 若变化，其条件为 $\|x_j-\mu_i\| \geq \|x_j-\mu_k\|$，则使 $r_{ij}=1 \rightarrow r_{kj}=1$.
且最小

进而使得原优化式更小

故每次更新，优化式不增，且更新次数有上限，故可收敛

2. 习题2（第四章 4.2）

(a) $\min\limits_{\beta} \sum\limits_{i=1}^{n} \left(y_i - x_i^T \cdot \beta \right)^2$

(b) $\min\limits_{\beta} (y - X\cdot\beta)^T \cdot (y - X\cdot\beta) \equiv \min\limits_{\beta} \|y - X\cdot\beta\|^2$

(c) $J = (y^T - \beta^T \cdot X^T)\cdot(y - X\cdot\beta) = y^Ty - 2\cdot\beta^T\cdot X^Ty + \beta^TX^TX\cdot\beta$.

令 $\frac{\partial J}{\partial \beta} = -2X^Ty + 2X^TX\cdot\beta = 0$

$\therefore \beta^* = (X^TX)^{-1}\cdot X^Ty$

(d) $\because X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}_{n\times d}$.

$\therefore X^TX \in \mathbb{R}^{d\times d}$.

$\because rank(X^TX) = rank(X) \leq \min(n, d)$

$\because d > n$

$\therefore rank(X^TX) \leq n < d$ 对 $X^TX \in \mathbb{R}^{d\times d}$ 来说一定不可逆.

(e) $R(\beta) = \|\beta\|^2$

① 惩罚 β 中过大的参数，让 model 在训练集上不过度拟合

② 解决 X^TX 不可逆问题，稳定 β^* 求解过程，具体在下面展开.

③ 从偏差-方差分解的视角看，其引入了偏差 $F(x) - \mathbb{E}_D[f(x;D)]$

但由于其不过拟合，不对噪声/outliers 学地过多，使其可降低方差 $\mathbb{E}_D[f(x;D) - \mathbb{E}_D[f(x;D)]]^2$

(f) $\min\limits_{\beta} \|y - X\beta\|^2 + \beta^T\beta\cdot\lambda$

令 $J = (y - X\beta)^T\cdot(y - X\beta) + \beta^T\beta\cdot\lambda$

令 $\frac{\partial J}{\partial \beta} = -2X^Ty + 2X^TX\cdot\beta + 2\beta\cdot\lambda = 0$

$\therefore \beta^* = (X^TX + \lambda\cdot I)^{-1}\cdot X^Ty$. 其中 $I = E$. 为单位阵

(g) 由于 $\lambda\cdot I > 0$，故 $X^TX + \lambda\cdot I$ 通常正定. 进而可逆，可以方便 β^* 的求解

(h) $\lambda = 0$ 也即退化成普通线性回归. $\beta^* = (X^TX)^{-1}\cdot X^T\cdot y$ （X^TX 可逆时）

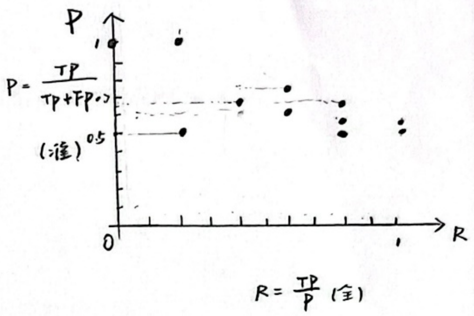
$\lambda = \infty$ 时，$\beta^* = (X^TX + \lambda\cdot I)^{-1}\cdot X^Ty = \frac{1}{\lambda}\cdot I\cdot X^Ty \to 0$，由于惩罚太大，$\beta^*$ 趋近 0 向量.

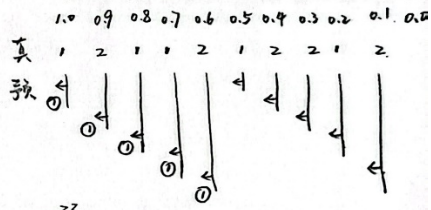
(i) 可用 inner-learning 来做，但直接联合优化 λ 和 β 时，固定 β 优化 λ 的过程中会使优化式非凸，进而优化不稳定且不可解释.

3. 习题 3（第四章 4.5）

a)，b)

下标	查准率	查全率	AUC-PC	AP
0	1.0000	0.0000		
1	1.0000	$\frac{1}{5}$=0.2000	0.20000	0.2000
2	$\frac{1}{2}$=0.5000	$\frac{1}{5}$=0.2000	0	0
3	$\frac{2}{3}$=0.6667	0.4000	0.1167	0.1333
4	0.7500	0.6000	0.1417	0.1500
5	0.6000	0.6000	0	0
6	0.6667	0.8000	0.1267	0.1333
7	0.5714	0.8000	0	0
8	0.5000	0.8000	0	0
9	0.5556	1.0000	0.1056	0.1111
10	0.5000	1.0000	0	0
			0.6906	0.7278





c) AUC-PC 为 0.6794

AP 为 0.7167

d) 代码如下.

真 \ 预	+	−	
+	TP	FN	P
−	FP	TN	


```python
def cal_AUCPR_AP(labels, scores):
    auc_pr = ap = 0
    tp_and_fn = labels.count(1)
    precision = [1, 0]
    recall = [0, 0]
    pairs = sorted(zip(scores, labels), reverse=True)
    for i in range(len(pairs)):
        precision[1] = (precision[0]*i + int(pairs[i][1] == 1))/(i+1)
        recall[1] = recall[0] + int(pairs[i][1] == 1)/tp_and_fn
        auc_pr += (recall[1]-recall[0]) * (precision[1]+precision[0]) / 2
        ap += (recall[1]-recall[0]) * precision[1]
        precision[0], recall[0] = precision[1], recall[1]
    return auc_pr, ap
labels = [1,2,1,1,2,1,2,2,1,2]
scores = [1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1]
print(cal_AUCPR_AP(labels, scores))
```

4. 习题 4（第五章 5.2）

$\boxed{corr1}$

a) scale = 0.001（默认），第一特征向量 和 平均向量 相关性为-0.318169（方向相反 相关性弱）

scale 越大（如1时），相关性为 0.999993.（方向相同 & 极相似）

随着 scale 减小，相关性减小

综上，未减去平均向量时，scale越大（也即构建数据时，平均向量占比更大时），"第一个特征向量"和"平均向量"相关性越大（也即 corr1 增加）

(b) scale 变化时，

$\boxed{corr2}$

随着 scale 的增大，"正确的特征向量"和"未减平均值的特征向量" 不断减小（相关性）

（也即 corr2 减小）

在 scale=1时，new_e1（正确）= $\begin{pmatrix} -0.029 \\ -0.092 \\ 0.3763 \\ \vdots \\ 0.2649 \end{pmatrix}$，二者相关性为-0.07724

在 scale=0.001时，new_e1（正确）= $\begin{pmatrix} 0.2872 \\ 0.2110 \\ \vdots \\ 0.1836 \end{pmatrix}$，二者相关性为 0.99999

5 习题五（第五章习题 5.4）

(a)

$$y = G(i,j,\theta)^T x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ c\cdot x_i - s\cdot x_j \\ \vdots \\ s\cdot x_i + c\cdot x_j \\ \vdots \\ x_m \end{pmatrix}$$

，更改了 x 中 x_i 和 x_j 的值，更新为
$$\begin{cases} y_i \leftarrow \cos\theta\cdot x_i - \sin\theta\cdot x_j \\ y_j \leftarrow \sin\theta\cdot x_i + \cos\theta\cdot x_j \end{cases}$$

(b) $y_j = 0$，也即 $\sin\theta\cdot x_i + \cos\theta\cdot x_j = 0$

$$\therefore \begin{cases} \sin\theta\cdot x_i + \cos\theta\cdot x_j = 0 \\ \sin^2\theta + \cos^2\theta = 1 \end{cases} \Rightarrow \begin{cases} \sin\theta = s = \pm \dfrac{x_j}{\sqrt{x_i^2 + x_j^2}} \\ \cos\theta = c = \mp \dfrac{x_i}{\sqrt{x_i^2 + x_j^2}} \end{cases}$$

（也可用 arcsin / arccos 表示 θ^*）.

(c)

（啥意思？，同(b)中推导？）

用同(b)中的数值稳定的归一化方法估计 c 和 s

$\forall a,b \in \mathbb{R}$，令
$$\begin{cases} c = \dfrac{a}{\sqrt{a^2+b^2}},\ s = \dfrac{b}{\sqrt{a^2+b^2}}, & a,b \text{ 不同时为 } 0 \\ c \text{ 和 } s \text{ 单独定义},\ c=1, s=0 \,/\, c=0, s=0, & a,b \text{ 同时为 } 0 \end{cases}$$

故估计 c 和 s 的任务，转化成估计 a 和 b 的任务，其中只有开根号和除法.

(d)

d.1 $A' = G(i,j,\theta)^T \cdot A$，其中 $G(i,j,\theta)^T \in \mathbb{R}^{m\times m}$，$A \in \mathbb{R}^{m\times n}$
$$\begin{cases} A'_{ii} = c\cdot a_{ii} - s\cdot a_{ji} \\ A'_{ij} = c\cdot a_{ij} - s\cdot a_{jj} \\ A'_{ji} = s\cdot a_{ii} + c\cdot a_{ji} \\ A'_{jj} = s\cdot a_{ij} + c\cdot a_{jj} \end{cases}$$
，其它元素和 A 保持一致

d.2 同理，这 4 个元素令其 $=0$ 时的 θ 即为目标。如令 $A'_{ij} = 0 \Rightarrow \begin{cases} s = \dfrac{a_{ij}}{\sqrt{a_{jj}^2 + a_{ij}^2}} \\ c = \dfrac{a_{jj}}{\sqrt{a_{jj}^2 + a_{ij}^2}} \end{cases}$

d.3 复杂度为 $O(m^2 \cdot n)$

(e)

$$R = \begin{pmatrix} r_{11} & & & \\ & r_{21} & & \\ & & \ddots & \\ & & & r_{m\times n} \end{pmatrix}_{m\times n}$$

思路：对 A 消成上三角后等式变换.

$$A = \begin{pmatrix} a_{11} & & & \vdots & \\ & a_{22} & & & \\ & & a_{kk} & & \\ & & \vdots & \ddots & \\ & & a_{mk} & & a_{mn} \end{pmatrix}_{m\times n}$$

以下以第 k 列为例进行说明.

目标：消去 a_{ik}，令其 $=0$，$i = k+1, \cdots, m$

令 $i = j-1$，$j = k+1, \cdots, m$，令 $\begin{cases} \cos\theta_j = \dfrac{a_{ik}}{\sqrt{a_{ik}^2 + a_{jk}^2}} \\ \sin\theta_j = \dfrac{a_{jk}}{\sqrt{a_{ik}^2 + a_{jk}^2}} \end{cases}$

使 $A' = G(i,j,\theta_j)^T \cdot A$ 中，$A'_{jk} = 0$

$\therefore$ 对 k 列下三角消元 $A_k = G^T(k,k+1,\theta_{k+1}) \cdots G^T(m-1,m,\theta_m) \cdot A_{k-1}$
$$\triangleq G_k^{1\,T} \cdots G_k^{m-k\,T} \cdot A_{k-1}$$

$$\therefore R = \underbrace{\left(G_1^{1\,T} \cdots G_1^{m-1\,T}\right) \cdot \left(G_2^{1\,T} \cdots G_2^{m-2\,T}\right) \cdots\cdots G_{n-1}^{1\,T}}\ \underset{\sim}{A}.$$

也即从第一列消到第 $n-1$ 列

由于 Givens 矩阵正交，即 $G_m^n \cdot G_m^{n\,T} = I$.

$$\therefore Q = G_{n-1}^{1} \cdots\cdots \left(G_2^{m-2} \cdots G_2^{1}\right) \cdot \left(G_1^{m-1} \cdots G_1^{1}\right)$$

6 习题六 (6.3)

(a) $\|X\|_2$ 即矩阵的 2-范数，值为 σ_1

$\|X^{-1}\|_2$ 即逆的 2-范数，值为 $\frac{1}{\sigma_n}$

$\therefore K_2(X) = \|X\|_2 \cdot \|X^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}$

(b) $K_2(X) = \frac{\sigma_1}{\sigma_n}$ 很大，即 σ_1 很大，σ_n 很小

类比 PCA 中的严格证明. λ_1 对应的 g_1 会使 $\{g_1^T\cdot(x-\bar{x})\}_n$ 的方差最大

λ_n 对应的 g_n 使 $\{g_n^T\cdot(x-\bar{x})\}_n$ 的方差最小

故此时. σ_1 很大, σ_n 很小., 使得 σ_1 对应的奇异向量方向方差大（信息被"放大"）

使得 σ_n 对应的奇异向量方差小，（信息被"压缩"）

$\therefore$ $x^* = A^{-1}\cdot b$ 时, A^{-1} 或 b 在 σ_n 的奇异向量上的小变动，由于原方差小，小扰动会极大影响判定

会很大程度改变 x^* 的值

（同理 σ_1 奇异向量方向）

(c) 证明:

$\because$ 正交矩阵 A, 有 $A^TA = AA^T = I$

而 $I = A^TA$ 所有的特征值均为 1，即 $\sigma_1=\sigma_2=\cdots=\sigma_n=1$

$\therefore$ $\sigma_i = \sqrt{\lambda_i} = 1$, $\forall i \in [n]$

$\therefore$ 故在 2-范数条件下, $K_2(X) = \frac{\sigma_1}{\sigma_n} = 1$

以下拓展到任意范数中

由 $\|\cdot\|$ 的定义，有 $\|Ax\| \le \|A\|\cdot\|x\|$, $\|A\| = \sup\limits_{x\neq0} \frac{\|Ax\|}{\|x\|}$ (由向量定义的范数)

$$K(A) = \sup_{x\neq0} \frac{\|Ax\|}{\|x\|} \cdot \sup_{y\neq0} \frac{\|A^{-1}y\|}{\|y\|} \ge \frac{\|Ax\|}{\|x\|}\cdot\frac{\|A^{-1}y\|}{\|y\|} \ge \frac{\|y\|}{\|x\|}\cdot\frac{\|x\|}{\|y\|} = 1$$

对 $\forall x, y$, $x\neq0, y\neq0$. 令 $x = Q^{-1}y$

由于 $\|Ax\| \le \|A\|\cdot\|x\|$

从 2 范数理解 $\|Ax\|_2^2 = x^TA^TAx = x^Tx = \|x\|_2^2$，推广到多范数时.

多数范数对多数向量，有 $\|Ax\| \approx \|x\|$

综上 $K(A) \ge 1$，但 $K(A)$ 不会很大，即正交矩阵 A 是良态的.

习题 7 ：简答题如下，代码请见压缩包中的.py 文件

1.

代码中 block 1：完成 train 和 test 的划分，并以我的学号（211300024）作为 random seed 从每个类中拿到一个样本

代码中 block 2：完成模型的加载（从 huggingface 上下载模型到本地啦，连接 vpn 好像也没法直接从 huggingface 的 url 下载呢（端口超时）），并对数据集 S 进行前向传播，计算 ViT-Tiny 模型的 CLS Token（默认维度为 192）

2.

代码中 block 3：完成 PCA，和保存 90%方差的降维

结果：

- 原始（未降维）的特征矩阵：torch.Size([200, 192])
- 原始维度: 192
- PCA 后维度: 71
- 保留的维度比: 0.3697916666666667

补充说明： structure.txt，项目结构（此时只上传了.py 文件）。.py 中均以相对路径完成。

CUB_200_2011

attributes

dataset

test

train

train_200

images

parts

model

model.safetensors

pytorch_model.bin

config.json

extract-cls.py（仅上传了本文件）