尹浚宇 161130118 计算机科学与技术 本科生

Problem 1

(a) 由题意，$K-means$的目标是在样例集上找到一种分组策略，使得每个组内部样例之间相似度最高. 由题目的定义可知，分组策略由所有的γ_{ij}给出，且每个组存在一个代表元素$\boldsymbol{\mu}$. 于是，组内部样例之间的相似度可以等价于它们在该组代表元素附近的集中程度. 假设使用欧氏距离作为相似度指标，对于任意一个组而言(这里假设为i)，其内部相似度就可表示为

$$\sum_{j=1}^{M}\gamma_{ij}||\boldsymbol{x}_j-\boldsymbol{\mu}_i||^2$$

该值越小则说明该组样例越集中，也就是样例之间的相似度越高. 考虑所有分组，我们得到

$$\sum_{i=1}^{K}\sum_{j=1}^{M}\gamma_{ij}||\boldsymbol{x}_j-\boldsymbol{\mu}_i||^2$$

于是找到最优的分组策略等价于找到恰当的γ_{ij}和$\boldsymbol{\mu}_i$使得上式的值最小，形式化之后得到

$$\arg\min_{\gamma_{ij},\boldsymbol{\mu}_i}\sum_{i=1}^{K}\sum_{j=1}^{M}\gamma_{ij}||\boldsymbol{x}_j-\boldsymbol{\mu}_i||^2$$

(b) 令$L(\Gamma,\mathrm{M})=\sum_{i=1}^{K}\sum_{j=1}^{M}\gamma_{ij}||\boldsymbol{x}_j-\boldsymbol{\mu}_i||^2$，其中$\Gamma=[\boldsymbol{\gamma}_{\cdot 1}\quad\cdots\quad\boldsymbol{\gamma}_{\cdot M}]$，$M=[\boldsymbol{\mu}_1\quad\cdots\quad\boldsymbol{\mu}_K]$，$\boldsymbol{\gamma}_{\cdot j}=[\gamma_{1j}\quad\cdots\quad\gamma_{Kj}]^T$.

当M固定时，要更新Γ使得$L(\Gamma,\mathrm{M})$最小，等价于重新计算每个样例应该属于的组. 由于Γ的每一列对应于一个样例的分组情况，于是逐列对Γ进行更新，对于第j列，规则如下:

$$i^*=\underset{i}{argmin}||\boldsymbol{x}_j-\boldsymbol{\mu}_i||^2,1\le i\le K$$

$$\gamma_{i^*j}=1,\text{该列其他元素设为 }0$$

当Γ固定时，要更新M使得$L(\Gamma,\mathrm{M})$最小，等价于重新计算每个组的代表元素. 这里是一个无约束的优化问题，考虑求导进行求解，有

$$\frac{\partial L}{\partial\boldsymbol{\mu}_i}=\sum_{j=1}^{M}\gamma_{ij}2\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)=0,1\le i\le K$$

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

(c) 首先观察到$L(\Gamma,\mathrm{M})=\sum_{i=1}^{K}\sum_{j=1}^{M}\gamma_{ij}||\boldsymbol{x}_j-\boldsymbol{\mu}_i||^2\ge 0$，设第$t$轮迭代时的输入分别是$\Gamma^{(t)}$和$\mathrm{M}^{(t)}$，由(b)的讨论可知，Γ和M在每一次更新时只会使得$L(\Gamma,\mathrm{M})$下降(或不变)，于是有$L\left(\Gamma^{(t+1)},\mathrm{M}^{(t+1)}\right)\le L\left(\Gamma^{(t)},\mathrm{M}^{(t)}\right)$，即整个过程中$L(\Gamma,\mathrm{M})$单调递减且有下界，于是该算法收敛.

Problem 2

(a) $\underset{\boldsymbol{\beta}}{argmin}\sum_{i=1}^{n}\left(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2$

(b) $\underset{\boldsymbol{\beta}}{argmin}(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})$

(c) 令$L(\boldsymbol{\beta})=(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})$，求导并令导数为 0 有

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = -2X^T(\boldsymbol{y} - X\boldsymbol{\beta}) = 0 \quad \Rightarrow \quad \boldsymbol{\beta} = (X^TX)^{-1}X^T\boldsymbol{y}$$

(d) $\because rank(X^TX) \leq \min\{rank(X), rank(X^T)\} \leq rank(X) \leq \min\{n, d\} = n$, $\therefore X^TX$不可逆.

(e) 可以降低原模型的过拟合程度.

(f) $\underset{\boldsymbol{\beta}}{argmin}[(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda\boldsymbol{\beta^T\beta}]$. 令$L'(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda\boldsymbol{\beta^T\beta}$, 求导求解:

$$\frac{\partial L'}{\partial \boldsymbol{\beta}} = -2X^T(\boldsymbol{y} - X\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} = 0 \quad \Rightarrow \quad \boldsymbol{\beta} = (X^TX + \lambda I)^{-1}X^T\boldsymbol{y}$$

(g) 注意到$\det(X^TX + \lambda I)$是关于λ的d次多项式, 由代数学基本定理, $\det(X^TX + \lambda I) = 0$在复数域上恰有$d$个根(重根以重数计). 考虑到本课程仅涉及实数, 所以使得$\det(X^TX + \lambda I) = 0$的$\lambda$数量至多为$d$个. 由于我们是在正实数范围内选择$\lambda$, 所以从测度的角度来说, $\Pr(\det(X^TX + \lambda I) \neq 0) = 1$, 这等价于$\Pr(X^TX + \lambda I$可逆$) = 1$.

(h) 当$\lambda = 0$时, $\boldsymbol{\beta} = (X^TX)^{-1}X^T\boldsymbol{y}$; 当$\lambda = \infty$时, $\boldsymbol{\beta} = 0$.

(i) 等价于最小化$L'(\boldsymbol{\beta}, \lambda) = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda\boldsymbol{\beta^T\beta}$. 求导有

$$\frac{\partial L'}{\partial \lambda} = \boldsymbol{\beta^T\beta} = 0 \quad \Rightarrow \quad \boldsymbol{\beta} = 0$$

可见若将λ当作参数考虑, 求出的回归模型无意义, 所以不能将λ视作参数.

Problem 3

(a)(b)

index	label	score	precision	recall	AUC-PR	AP
0			1.0000	0.0000	-	-
1	1	1.0	1.0000	0.2000	0.2000	0.2000
2	2	0.9	0.5000	0.2000	0.0000	0.0000
3	1	0.8	0.6667	0.4000	0.1167	0.1333
4	1	0.7	0.7500	0.6000	0.1417	0.1500
5	2	0.6	0.6000	0.6000	0.0000	0.0000
6	1	0.5	0.6667	0.8000	0.1267	0.1333
7	2	0.4	0.5714	0.8000	0.0000	0.0000
8	2	0.3	0.5000	0.8000	0.0000	0.0000
9	1	0.2	0.5556	1.0000	0.1056	0.1111
10	2	0.1	0.5000	1.0000	0.0000	0.0000
					0.6907	0.7277

AUC-PR 和 AP 确实很相似.

(c)

index	label	score	precision	recall	AUC-PR	AP
9	2	0.2	0.4444	0.8000	0.0000	0.0000
10	1	0.1	0.5000	1.0000	0.0944	0.1000
					0.6795	0.7166

最后三行改变如上, 其余不变.

(d) 运行结果如下(代码见附件), 可见程序结果与手算结果仅求和的最后一位不同. 这是因为手算时每一个值都进行了舍入, 而程序只在求和最后才进行舍入.

```
>> auc_pr
AUC-PR
    0.2000         0    0.1167    0.1417         0    0.1267         0         0    0.1056         0

    0.6906

AP
    0.2000         0    0.1333    0.1500         0    0.1333         0         0    0.1111         0

    0.7278
```

Problem 4

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

(b) $\mathbb{E}[f]=\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}y_{nn(i)}\right]=\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[y_{nn(i)}\right]=\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[F\left(\boldsymbol{x}_{nn(i)}\right)+\varepsilon_i\right]$. 由于$\varepsilon$独立于其余所有随机变量，且$\mathbb{E}[\varepsilon]=0$，所以有

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

(c) $\mathbb{E}[(y-f)^2]=\left(F-\frac{1}{k}\sum_{i=1}^{k}F\left(\boldsymbol{x}_{nn(i)}\right)\right)^2+\mathbb{E}\left[\left(\frac{1}{k}\sum_{i=1}^{k}y_{nn(i)}-\frac{1}{k}\sum_{i=1}^{k}F\left(\boldsymbol{x}_{nn(i)}\right)\right)^2\right]+\sigma^2$

(d) $\mathbb{E}\left[\left(\frac{1}{k}\sum_{i=1}^{k}y_{nn(i)}-\frac{1}{k}\sum_{i=1}^{k}F\left(\boldsymbol{x}_{nn(i)}\right)\right)^2\right]=\mathbb{E}\left[\left(\frac{1}{k}\sum_{i=1}^{k}\varepsilon_i\right)^2\right]=\frac{1}{k^2}\mathbb{E}\left[\left(\sum_{i=1}^{k}\varepsilon_i\right)^2\right]$. 由于$\varepsilon_i$之间互相独立，且$\mathbb{E}[\varepsilon]=0$，所以有$\left(\sum_{i=1}^{k}\varepsilon_i\right)^2$展开式中交叉项的期望均为 0，方差项即为

$$\frac{1}{k^2}\mathbb{E}\left[\left(\sum_{i=1}^{k}\varepsilon_i\right)^2\right]=\frac{1}{k^2}\sum_{i=1}^{k}\mathbb{E}[\varepsilon_i^2]=\frac{1}{k^2}\sum_{i=1}^{k}Var(\varepsilon)=\frac{\sigma^2}{k}$$

根据上式，当k变大时，方差项会变小.

(e) 偏差平方项为$\left(F-\frac{1}{k}\sum_{i=1}^{k}F\left(\boldsymbol{x}_{nn(i)}\right)\right)^2$. 该项实际上受很多因素影响，比如$F(\boldsymbol{x})=c$时，$k$的变化不会影响该项. 又或者$F(\boldsymbol{x})\neq c$时，该项取决于新加入的近邻的标签值与$F(\boldsymbol{x})$的差值大小. 不过不考虑极端情况，并从宏观角度去看的话，k增大会导致该项也变大，当$k=n$时，该回归模型的输出与输入$\boldsymbol{x}$无关，这时偏差项很大.

Problem 5

(a) 因为G是实正交阵，所以有$GG^T=G^TG=I$. 注意到

$$||Gx||=\sqrt{(Gx)^TGx}=\sqrt{x^TG^TGx}=\sqrt{x^Tx}=||x||$$

同理有$||Gx||=||x||$.

(b) 不难注意到

$$\|G^TXG\|_F=\sqrt{tr(G^TXGG^TX^TG)}=\sqrt{tr(G^TXX^TG)}=\sqrt{tr(GG^TXX^T)}=\sqrt{tr(XX^T)}=\|X\|_F$$

(c) 由于X是实对称矩阵，于是一定存在实正交阵G，使得$\Lambda=G^TXG$，其中Λ为对角阵，且满足$\Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_n)$，$\lambda_1$~$\lambda_n$为$X$的$n$个特征值，$G$的列向量组为$X$的特征向量组.

注意到$off(G^TXG)=off(\Lambda)=0$，于是找正交阵$J$，使得$off(J^TXJ)<off(\mathrm{X})$的过程不断重复等价于对$X$进行特征分解. 即假设重复$t$次后有$off(J_t^T\cdots J_1^TXJ_1\cdots J_t)=0$，则我们可以得到$X$的特征值即为$J_t^T\cdots J_1^TXJ_1\cdots J_t$对角线上元素，$X$的特征向量组即为$J_1\cdots J_t$的列向量组.

(d) 考虑令$J=J(i,j,\theta)$(教材第5章习题4中的矩阵)，注意到J^TXJ也为实对称矩阵，于是计算$(J^TXJ)_{ij}=0$即可.

首先计算J^TX的第i行为$[X_{i1}cos\theta - X_{j1}sin\theta, \cdots, X_{in}cos\theta - X_{jn}sin\theta]$.
然后计算$(J^TXJ)_{ij} = \big(X_{ii}cos\theta - X_{ji}sin\theta\big)sin\theta + \big(X_{ij}cos\theta - X_{jj}sin\theta\big)cos\theta$，注意到$X$为实对称矩阵，化简有$(J^TXJ)_{ij} = \frac{(X_{ii}-X_{jj})}{2}sin2\theta + X_{ij}cos2\theta$，令其为0得到$tan2\theta = \frac{2X_{ij}}{X_{jj}-X_{ii}}$，于是我们可以令$J = J(i,j,\frac{\arctan\left(\frac{2X_{ij}}{X_{jj}-X_{ii}}\right)}{2})$.

(e) 不难注意到变换J^TXJ只影响X的p,q行和p,q列，且由于J^TXJ也为实对称矩阵，于是考虑p,q行即可. 观察到off函数不统计矩阵的对角线元素，且由(d)可知$(J^TXJ)_{pq} = (J^TXJ)_{qp} = 0$.
记$J^TXJ = X'$，有如下变换公式成立:

$$\begin{cases} X'_{pi} = X_{pi}cos\theta - X_{qi}sin\theta \\ X'_{qi} = X_{pi}sin\theta + X_{qi}cos\theta \end{cases}, 1 \le i \le n, i \ne p, q$$

考虑p,q行对应位置平方和，有

$$\big(X'_{pi}\big)^2 + \big(X'_{qi}\big)^2 = \big(X_{pi}\big)^2 + \big(X_{qi}\big)^2$$

结合上式和上述讨论，我们可以得到

$$off^2(J^TXJ) = off^2(X) - 2X_{pq}^2$$

于是命题得证.

(f) 根据(e)我们知道每次迭代时$off(X)$的值不会上升，且由(c)我们知道$off(X)$有下界0，由单调有界原则可知，该算法一定收敛.