Pattern Recognition*
Homework II

* Teacher: Jianxin Wu. TA: ...

1st 张逸凯 171840708 计算机科学与技术系 本科生

Department of Computer Science and Technology

Nanjing University

zykhelloha@gmail.com

张逸凯 ***171840708*** 计算机科学与技术系 本科生

03_Framework 2. *K-means*

解:

1. 对于$\forall x_i$，需要找到x_i的类别$c_i = arg\,min_j||x_i - \mu_j||$，等价于:

$$J(x_i) = \begin{cases} ||x_i - \mu_j|| & (j = c_i) \\ 0 & (j \neq c_i) \end{cases}$$

   对于c_i, μ_j的计算，令C_i为属于第i类样本的集合，详细地说为:

$$\begin{aligned} C_i &= \{x_n : \|x_i - \mu_k\| \leq all\,\|x_i - \mu_j\|\} \\ \mu_k &= \tfrac{1}{|C_i|} \textstyle\sum_{x_i \in C_i} x_i \end{aligned} \tag{1}$$

   此时可以将$J(x_i)$对i的每一项**同乘**γ_{ij}，当$j = c_i$时$\gamma_{ij} = 1$，反之$\gamma_{ij} = 0$。这样就得到了*K-means*需要优化的目标:

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

2. **update** γ_{ij}:

$$\gamma_{ij} = \begin{cases} 1 & (\underset{j}{\arg\min}\ ||x_i - \mu_j|| = j) \\ 0 & (others) \end{cases}$$

   update μ_i:

$$\mu_i = \frac{1}{\sum_j \gamma_{ij}} \sum_j \gamma_{ij} x_j \tag{3}$$

3. 由上述K-means goal，可以定义**loss**函数如下:

$$f(\gamma,\mu)=\sum_{i=1}^{K}\sum_{j=1}^{M}\gamma_{ij}\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|^2 \tag{4}$$

首先证明以下几个引理从而说明迭代步中**loss函数是减小的**.

> **Lemma 1**: $\mathbb{E}\|X-z\|^2=\mathbb{E}\|X-\mathbb{E}X\|^2+\|z-\mathbb{E}X\|^2$, *其中$z$是其他数据点.*
> 证明:
>
> $$\begin{aligned}&\mathbb{E}\|X-\mathbb{E}X\|^2+\|z-\mathbb{E}X\|^2\\&=\mathbb{E}\left[\|X\|^2+\|\mathbb{E}X\|^2-2X\cdot\mathbb{E}X\right]+\left[\|z\|^2+\|\mathbb{E}X\|^2-2z\cdot\mathbb{E}X\right]\\&=\mathbb{E}\|X\|^2+\|\mathbb{E}X\|^2-2\mathbb{E}X\cdot\mathbb{E}X+\|z\|^2+\|\mathbb{E}X\|^2-2z\cdot\mathbb{E}X\\&=\mathbb{E}\|X-z\|^2\end{aligned}$$
>
> **Lemma 2**: *以上代价在k-means的迭代(**b题中的两个步骤**)中都是单调减的.*
>
> 证明: 令$\gamma^{(t)}$是第t次迭代的结果, 对于算法的两个步骤:
>
> - **update** γ_{ij}之后$f(\gamma^{(t+1)},\mu)\le f(\gamma^{(t)},\mu)$(易见因为$x$被分配给了最近的中心, 显然比之前更远的中心欧氏距离更小).
> - **update** μ_i之后$f(\gamma,\mu^{t+1})\le f(\gamma,\mu^t)$(因为**Lemma 1**).
>
> $\therefore$原命题得证.

以上证明了loss function是减小的, 下面证明收敛性: 我们知道最多只有 $\left\{\begin{matrix}n\\k\end{matrix}\right\}$ (第二类Stirling数)种方法将N个数据划分成k类, 这是**有限的**. 注意到每次迭代, **有且仅有**以下两种影响:

1. 上述f值在两次迭代之间不变, 则接下来也不会变, 收敛.
2. 由上Lemma2的证明我们知道, 如果两次迭代后f值不同, 则迭代后f减小. 又划分数有限, 所以收敛.

综上所述, 收敛性得证.

(ps: *k-means*确实是很有趣的东西, 偶然看了shen教授的Yinyang K-Means: A Drop-In Replacement of the Classic K-Means with Consistent Speedup, 优化方法很妙)

04_Error 2. *Linear regression*

解:

1. $$cost=\frac{1}{N}\sum_{i=1}^{n}\left(\hat{y}_i-\left(x_i^T\beta+\epsilon\right)\right)^2 \tag{5}$$

   其中$\hat{y}_i$是Ground Truth. 可以对β,ϵ求偏导来得到最优解.
2. 令$\mathrm{E}\in\mathbb{R}^d$, $\mathrm{E}=(\epsilon,\ \ldots,\ \epsilon)^T$. 则有

   $$\begin{aligned}\boldsymbol{y}&=\boldsymbol{X}^T\boldsymbol{\beta}+\mathrm{E}\\cost&=\frac{1}{N}(\hat{\boldsymbol{y}}-\boldsymbol{X}^T\boldsymbol{\beta}-\mathrm{E})^T(\hat{\boldsymbol{y}}-\boldsymbol{X}^T\boldsymbol{\beta}-\mathrm{E})\end{aligned} \tag{6}$$
3. 不妨令$E=N\,loss$, 因为同时对$\boldsymbol{\beta},\epsilon$进行优化, 可以做一个吸收: $\boldsymbol{\beta}:=(\beta;\ \epsilon)$, 并把数据$X$表示成一个$n\times(d+1)$维度的矩阵. 即:

   $$\mathbf{X}=\begin{pmatrix}x_{11}&x_{12}&\ldots&x_{1d}&1\\x_{21}&x_{22}&\ldots&x_{2d}&1\\\vdots&\vdots&\ddots&\vdots&\vdots\\x_{n1}&x_{n2}&\ldots&x_{nd}&1\end{pmatrix}=\begin{pmatrix}\boldsymbol{x}_1^{\mathrm{T}}&1\\\boldsymbol{x}_2^{\mathrm{T}}&1\\\vdots&\vdots\\\boldsymbol{x}_n^{\mathrm{T}}&1\end{pmatrix} \tag{7}$$

$$
\begin{aligned}
E &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\
\frac{\partial E_\beta}{\partial \beta} &= \frac{\partial \boldsymbol{y}^T\boldsymbol{y}}{\partial \beta} - \frac{\partial \boldsymbol{y}^T\mathbf{X}\boldsymbol{\beta}}{\partial \beta} - \frac{\partial \boldsymbol{\beta}^T\mathbf{X}^T\boldsymbol{y}}{\partial \beta} + \frac{\partial \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}}{\partial \beta} + 0 \\
&= 0 - \mathbf{X}^T\boldsymbol{y} - \mathbf{X}^T\boldsymbol{y} + \left(\mathbf{X}^T\mathbf{X} + \mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta} \\
&= 2\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y})
\end{aligned} \tag{8}
$$

上式求偏导后为0，等价于$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$，其最后一列即为吸收的$\epsilon$的值.

4. $\because \boldsymbol{X} \in \mathbb{R}^{n\times d}$. $\therefore \mathrm{rank}(\mathrm{X}) = \mathrm{rank}\left(\mathrm{X}^T\mathrm{X}\right) \le n < d$

   $\boldsymbol{X}^T\boldsymbol{X}$不满秩，所以不可逆.

5. ridge regression所加的L2可以防止过拟合，因为大多数的拟合是因为参数过大，使新的数据与原数据有一点差异就带来输出的巨大变化，加入正则化项后限制了参数β的大小，从另一个角度也减小了模型的复杂度，有效防止过拟合.

6. ridge regression的loss function可表达为：

$$
cost = \frac{1}{N}(\hat{\boldsymbol{y}} - \boldsymbol{X}^T\boldsymbol{\beta} - \mathrm{E})^T(\hat{\boldsymbol{y}} - \boldsymbol{X}^T\boldsymbol{\beta} - \mathrm{E}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} \tag{9}
$$

   同上题使用lagrange乘子法求得闭式解：

$$
\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda I)^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{10}
$$

7. $$(Xx)^T(Xx) = \|Xx\|^2 \ge 0 \tag{11}$$

   所以$\boldsymbol{X}^T\boldsymbol{X}$是半正定的，$\boldsymbol{X}^T\boldsymbol{X} + \lambda I$是正定的，可逆. 这样$\beta$的闭式解存在.

8. - $\lambda = 0$，则ridge regression退化为一般的linear regression.
   - $\lambda \to +\infty$，则loss function中正则化项的影响很大，优化时$\beta \to 0$.
9. λ是不能作为普通的参数被模型学得的.
   - 从梯度下降优化角度：$\lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$是线性的，所以沿着负梯度方向，$\lambda$会一直减小，这在$\lambda \ge 0$下是无意义的.
   - 从λ的作用角度：λ是可以调整正则化项对模型复杂程度(或者说优化速率)的影响，使SGD作用在其他参数上时在真正的目标上获得了良好的性能，而不是通过直接优化λ来最小化loss function.

04_Error 5. PR curve

解：

1. (12)

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
					0.6906	0.7278

2. AUC-PR和AP是相似的，因为他们只是在计算PR曲线下面积时 运用了不同的插值计算近似方法.
3. 调换数据位置后的表为：

index	label	score	precision	recall	AUC-PR	AP
0			1.0000	0.0000	–	–
	...		...*Same*	*as*	*the*	*above*...
9	2	0.2	0.4444	0.8000	0.0000	0.0000
10	1	0.1	0.5000	1.0000	0.0944	0.1000
					0.6794	0.7167

(13)

4. 代码如下:

```python
import csv
import numpy as np

def readData():
    x = [
        [1, 1, 1.0],
        [2, 2, 0.9],
        [3, 1, 0.8],
        [4, 1, 0.7],
        [5, 2, 0.6],
        [6, 1, 0.5],
        [7, 2, 0.4],
        [8, 2, 0.3],
        [9, 1, 0.2],
        [10, 2, 0.1]
    ]
    return np.array(x).astype('float')

# 计算当前阈值下的TP, FP, FN, TN
def calculateTPFPFNTN(data, threshold):
    TP, FP, FN, TN = 0, 0, 0, 0
    for i in data:
        if i[2] >= threshold:
            if i[1] == 1:
                TP += 1
            else:
                FP += 1
        else:
            if i[1] == 1:
                FN += 1
            else:
                TN += 1
    return TP, FP, FN, TN

def calculateAUCPRandAP(data):
    befRec, befPre = 0, 1 # 上一步的rec和pre.
    sumAUCPR, sumAP = 0, 0
    for i in data:
        TP, FP, FN, TN = calculateTPFPFNTN(data, i[2])
        curPre, curRec = TP / (TP + FP),  TP / (TP + FN)
        AUCPR = (curRec - befRec) * (curPre + befPre) / 2 # 当前插值梯形的面积.
        AP = curPre * (curRec - befRec) # 当前矩形的面积.
        befRec, befPre = curRec, curPre
        sumAUCPR += AUCPR
        sumAP += AP
        print("%.4f\t%.4f\t%.4f\t%.4f" % (curPre, curRec, AUCPR, AP))
    print("#.##\t#.##\t%.4f\t%.4f" % (sumAUCPR, sumAP))

def main():
    data = readData()
    calculateAUCPRandAP(data)

if __name__ == '__main__':
    main()
```

04_Error 6. bias-variance decomposition

解：(注：本题公式由`mathpix`软件识别)

1.

$$
\begin{aligned}
& \mathbb{E} \Leftrightarrow \mathbb{E}_D \\
& \mathbb{E}\left[(y-f(x ; D))^2\right] \\
= & \mathbb{E}\left[(F(x)-f(x ; D)+\epsilon)^2\right] \\
= & \mathbb{E}\left[(F(x)-f(x ; D))^2\right]+\mathbb{E}\left[\epsilon^2\right]+\mathbb{E}[2(F(x)-f(x ; 0))] \mathbb{E}[\epsilon] \\
= & (\mathbb{E}[F(x)-f(x ; D)])^2+\operatorname{Var}[F(x)-f(x ; D)]+(\mathbb{E}[\epsilon))^2+\operatorname{Var}(\epsilon) \\
= & (\mathbb{E}[F(x)-f(x ; D)])^2+\operatorname{Var}(-f(x ; 0))+\sigma^2 \\
= & (F(x)-\mathbb{E}[(x ; D)])^2+\mathbb{E}\left[(f(x ; D)-\mathbb{E}[f(x ; D)])^2\right]+\sigma^2
\end{aligned}
$$

其中第一项是偏差，第二项是方差，第三项是噪声.

2.

$$
\begin{aligned}
& \because E[F]=F \\
& \mathbb{E}[f] \\
= & \mathbb{E}\left[\frac{1}{k} \sum_{i=1}^k y_{n n(i)}\right] \\
= & \frac{1}{k} \mathbb{E}\left[\sum_{i=1}^k F(x_{n n(i)})+\epsilon\right] \\
= & \frac{1}{k} \mathbb{E}\left[\sum_{i=1}^k F\right]+0 \\
= & \frac{\sum_{i=1} F\left(x_{n n(i)}\right)}{k}
\end{aligned}
$$

3.

$$
\begin{aligned}
& (\mathbb{E}[F-f])^2=(F-\mathbb{E}[f])^2 \\
& =\left(F(x)-\frac{\sum_{i=1}^k F\left(x_{n n(i)}\right)}{k}\right)^2
\end{aligned}
$$

$$
\begin{aligned}
& \because \epsilon_i \text { is i.i.d. } \\
& \operatorname{Var}\left(\frac{1}{k} \sum_{i=1}^k y_{n n(i)}\right) \\
& =\frac{1}{k^2} \operatorname{Var}\left(\sum_{i=1}^k F-\epsilon_i\right) \\
& =\frac{1}{k^2} \operatorname{Var}\left(\sum_{i=1}^k \epsilon_i\right) \\
& =\frac{\sigma^2}{k}
\end{aligned}
$$

$$
\begin{aligned}
& (a)^{\prime} s \ result: \\
& \mathbb{E}\left[(y-f)^2\right]=\mathbb{E}\left[(F-f)^2\right]+\sigma^2 \\
& \quad=(E[F-f])^2+\operatorname{Var}[F-f]+\sigma^2 \\
& =(E[F-f])^2+\operatorname{Var}(f)+\sigma^2 \\
& =\left(F(x)-\frac{\sum_k F\left(x_{n n(i)}\right)}{k}\right)^2+\frac{\sigma^2}{k}+\sigma^2
\end{aligned}
$$

4. 由(c)题知，variance term 为：$\frac{\sigma^2}{k}$，它与k成反比．随着k的增大而减小.

5. 由(c)题知，squared bias term 为 $\left(F(x)-\frac{\sum_k F(x_{nn(i)})}{k}\right)^2$，考虑提示中$k=n$时，squared bias term将变大，这是可解释的，因为k较大时考虑的最近邻的点更多了，也就是他们的平均更有可能远离这个点；反之如果k较小，所考虑的是更小最近邻集合，那么他们的平均值更有可能接近这个点，所以bias会更小.

   即squared bias term随着k的增大而增大.

05_PCA 5. Jacobi eigenvalue algorithm

解：(注：此题是在上完Indian Institute of Technology Roorkee的Professor Dr. Sanjeev Kumar的Numerical Methods课程Lecture 13后作答的)

1. $\because G^TG=I, \therefore ||x||=\sqrt{x^Tx}=\sqrt{(Gx)^TGx}=||G^Tx||$

   同理：$||x||=\sqrt{(G^Tx)^TG^Tx}=||G^Tx||$

2.

$$\begin{aligned} tr(AB) &= tr(BA) \\ ||G^TXG||_F &= \sqrt{tr(G^TXG\ G^TX^TG)} \\ &= \sqrt{tr(G^TXX^TG)} \\ &= \sqrt{tr(X^TG\ G^TX)} \\ &= \sqrt{tr(X^TX)} = \sqrt{tr(XX^T)} \\ &= ||X||_F \end{aligned} \tag{14}$$

3. 首先证明一些引理：

   > **Lemma**：if $G^TG=I$，$||GX||_F^2=||X||_F^2$.
   >
   > 证明：
   >
   > $$||GX||_F^2=tr(GXX^TG^T)=tr(X^TG^TGX)=tr(X^TX)=||X||_F^2 \tag{15}$$
   >
   > **Lemma**：若 A,B相似，A和B有相同特征值.
   >
   > 证明：
   >
   > $$\begin{aligned} &A=GBG^{-1} \\ &A\xi_i=\lambda_i\xi_i \Rightarrow BG^{-1}\xi_i=\lambda_iG^{-1}\xi_i \end{aligned} \tag{16}$$

   $\because J^TJ=I$，$\therefore$ J^TXJ, X两矩阵相似，相似矩阵有**相同特征值**.

   对于任意的J，我们有$off(J^TXJ)<off(X)$，也即在J连续作用之后(迭代步)：$off((J^T)^N\ X\ J^N)$，即X除了对角线的元素平方和**在不断减小**，最后只剩下对角线的元素，又因为相似矩阵特征值相同，所以最后对角线留下的是特征值，在最后一小题将证明迭代步是**收敛**的.

4. (由于前几题使用G，本小题中G,J等价.) 因为J是正交的，并且等价于旋转操作，不失一般性考虑Givens rotation matrix J:

$$J(i,j,\theta)=\begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \\ \vdots & \ddots & \vdots & & \vdots & & \vdots & \\ 0 & \cdots & cos(\theta) & \cdots & -sin(\theta) & \cdots & 0 & (idx:i) \\ \vdots & & \vdots & \ddots & \vdots & & \vdots & \\ 0 & \cdots & sin(\theta) & \cdots & cos(\theta) & \cdots & 0 & (idx:j) \\ \vdots & & \vdots & & \vdots & \ddots & \vdots & \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 & \end{bmatrix}$$

其中$J_{ii}=cos(\theta), J_{ij}=sin(\theta)$，依次类推$J_{ji}=-sin(\theta), J_{jj}=cos(\theta)$.

以上的Jacobi旋转矩阵J作用在X上时，等价于做了一次相似(旋转)变换：J改变了第i,j行和第i,j列.

令$X'=J^TXJ$，注意到变换作用后：

$$\begin{aligned}X'_{ii}&=cos(\theta)^2X_{ii}-2sin(\theta)cos(\theta)X_{ij}+sin(\theta)^2X_{jj}\\X'_{jj}&=sin(\theta)^2X_{ii}+2sin(\theta)cos(\theta)X_{ij}+cos(\theta)^2X_{jj}\\X'_{ij}&=X'_{ji}=\left(cos(\theta)^2-sin(\theta)^2\right)X_{ij}+sin(\theta)cos(\theta)\left(X_{ii}-X_{jj}\right)\end{aligned}\tag{17}$$

要使$X'_{ij}=X'_{ji}=0$，等价于

$$\theta=\frac{1}{2}arctan\left(\frac{2X_{ij}}{X_{jj}-X_{ii}}\right)\tag{18}$$

即θ取上述值即可满足题意.

5. 不妨令$X'=J^TXJ$，我们有$||X'||_F=||X||_F$，注意到当$m,n\neq i,j$时，$X'_{mn}=X_{mn}$，即不在第i,j行或列上的元素不变. 从而我们可以证明$||X'||_F-\sum_{i=1}^n(X'_i)^2<||X||_F-\sum_{i=1}^n(X_i)^2$即可. 也就是对角线元素平方和(不妨令为$X_{diag}$)变大.

   这里p,q就是上一题中的i,j，X'对应元素变换：

$$\begin{aligned}X'_{pp}&=cos(\theta)^2X_{pp}-2sin(\theta)cos(\theta)X_{pq}+sin(\theta)^2X_{qq}\\X'_{qq}&=sin(\theta)^2X_{pp}+2sin(\theta)cos(\theta)X_{pq}+cos(\theta)^2X_{qq}\\X'_{pq}&=X'_{qp}=\left(cos(\theta)^2-sin(\theta)^2\right)X_{pq}+sin(\theta)cos(\theta)\left(X_{pp}-X_{qq}\right)\end{aligned}\tag{19}$$

   注意$|X_{pq}|=\max_{i\neq j}|X_{ij}|$在证明过程的最后一个$>$中被使用到.

$$\begin{aligned}(X')^2_{diag}&=\textstyle\sum_{i=1}^n(X'_i)^2\\&=\textstyle\sum_{i\neq p,q}(X'_{ii})^2+\left((X'_{pp})^2+(X'_{qq})^2\right)\\&=\textstyle\sum_{i\neq p,q}X_{ii}^2+\left(X_{pp}^2+X_{qq}^2+2X_{pq}^2\right)\\&=\textstyle\sum_{i=1}^nX_{ii}^2+2X_{pq}^2\\&>(X)^2_{diag}\end{aligned}\tag{20}$$

上面的证明过程化简是极麻烦的(二次项的平方)，所以我直接从off来证明：

$$\begin{aligned}&k\neq p,q\\&X'_{pk}=cosX_{pk}-sinX_{qk}\\&X'_{qk}=sinX_{pk}+cosX_{qk}\\&X'\text{的上半角部分}(\text{因为是对称阵})\\=&\sum_k X'^2_{pk}+X'^2_{qk}+X'^2_{pq}\\=&\sum_k\left(sin(\theta)X_{pk}+cos(\theta)X_{qk}\right)^2+\left(cos(\theta)X_{pk}-sin(\theta)X_{qk}\right)^2\\=&X_{pk}^2+X_{qk}^2\\<&X_{pk}+X_{qk}^2+X_{pq}^2\end{aligned}$$

6. 延用上题中$|X_{pq}|=\max_{i\neq j}|X_{ij}|$，不妨令其等于$m$，又$X\in\mathbb{R}^{n\times n}$，所以除了对角线以外的元素有$n(n-1)$个.

$$\begin{aligned}&\text{off}(X)^2\leq n(n-1)X_{pq}^2\\&\Rightarrow\text{off}(X')^2\leq\left(1-\frac{2}{n(n-1)}\right)\text{off}(X)^2\end{aligned}\tag{21}$$

上式已经可以说明收敛性了，不妨令$X'^N=(J)^{N\,T}XJ^N$，对$\forall\,\epsilon>0,\,\exists\,N>N_0$，使$\left(1-\frac{2}{n(n-1)}\right)^N<\epsilon$.

收敛性得证.