Pattern Recognition HW 2

171240510 Yuheng Ma

Math Major, Kuang Yaming Honors School

Spring 2020

1 03 Framework 2

1. There are K μ's that represent K points which best describe the centers of K groups. A x_j is only assigned to one group, and the "variance" can be shown by $\sum_{j=1}^{M} \gamma_{ij}||x_j - \mu_i||^2$. Total "in-group variance" can be shown by $\sum_{i=1}^{K}\sum_{j=1}^{M} \gamma_{ij}||x_j - \mu_i||^2$. Less variance shows better in-group connection and thus we tends to minimize this.

2. When μ fixed, $\min \sum_{i=1}^{K}\sum_{j=1}^{M} \gamma_{ij}||x_j - \mu_i||^2 \leq \sum_{j=1}^{M} \min \sum_{i=1}^{K} \gamma_{ij}||x_j - \mu_i||^2 = \sum_{j=1}^{M} \min_i ||x_j - \mu_i||^2$, this is a attained when
$$\gamma_{ij} = \begin{cases} 1 & i = \arg\min ||x_j - \mu_i|| \\ 0 & i = \text{ others} \end{cases}$$
. When γ fixed,
$$\begin{aligned} &\frac{\partial \sum_{i=1}^{K}\sum_{j=1}^{M} \gamma_{ij}||x_j - \mu_i||^2}{\partial \mu_i} \\ =&\sum_{j=1}^{M} \gamma_{ij} 2(x_j - \mu_i) \\ =&2\sum_{x_j \in G_i} x_j - \mu_i \end{aligned}$$
when set to 0, μ_i is set to $\bar{x}_i$, the mean of x's in group i.

3. The state space of clustering E is a finite state space and |E|= K^M. Let the state after ith iteration be s_i, then there will be a convergent subsequence $\{s_{n_k}\}$, which is just $s_{n_1} = s_{n_2} = \cdots s_{n_k} = \cdots$. However, each iteration will reduce the loss function $\sum_{i=1}^{K}\sum_{j=1}^{M} \gamma_{ij}||x_j - \mu_i||^2$, for the reason that step i and ii both find the sufficient statistics for γ and μ. Thus, $Loss(s_{n_i}) < Loss(s_{n_i+k}) < Loss(s_{n_{j+1}})$, a contradiction. Thus s does not change after reaching s_{n_1}.

2 04 Error 2

1. As an otimization problem, this is equivalent to finding $\tilde{\beta}$
$$\sum_{i=1}^{n} ||x_i^T \tilde{\beta} - y_i||^2 = \min_{\mathbb{R}^d} \sum_{i=1}^{n} ||x_i^T \beta - y_i||^2$$

2. As an otimization problem, this is equivalent to finding $\tilde{\beta}$
$$||X\tilde{\beta} - y||_2^2 = \min_{\mathbb{R}^d} ||X\beta - y||_2^2$$

3. Take derivative and let it be zero, we have
$$\tilde{\beta} = (X^T X)^{-1} X^T y$$

Table 1: Calculation of AUC-PR and AP

index	label	score	precision	recall	AUC-PR	AP
0			1	0	-	-
1	1	1.0	1	0.2	0.2	0.2
2	2	0.9	0.5	0.2	0	0
3	1	0.8	0.66	0.4	0.116	0.132
4	1	0.7	0.75	0.6	0.141	0.15
5	2	0.6	0.6	0.6	0	0
6	1	0.5	0.66	0.8	0.126	0.132
7	2	0.4	0.5714	0.8	0	0
8	2	0.3	0.5	0.8	0	0
9	1	0.2	0.5556	1	0.10556	0.11112
10	2	0.1	0.5	1	0	0
					0.6905	0.7277

4. No. X^TX and XX^T have same group of eigenvalue except some zeros. If d>n, X^TX have 0 as its eigenvalue and thus not invertible.

5. It gives weight to norm of β, which usually helps to to solve an ill-posed problem or to prevent overfitting. Here, it gives a unique solution despite the relationship of n and d.

6. As an otimization problem, this is equivalent to finding $\tilde{\beta}$

$$||X\tilde{\beta} - y||_2^2 + \lambda\tilde{\beta}^T\tilde{\beta} = \min_{\mathbb{R}^d} ||X\beta - y||_2^2 + \lambda\beta^T\beta$$

. We have the derivative

$$2X^TX\beta - 2X^Ty + 2\lambda\beta \Rightarrow \tilde{\beta} = (X^TX + \lambda I)^{-1}X^Ty$$

7. In 6, we notice that $X^TX + \lambda I$ is positive since X^TX is semi-positive.

8. λ =0 then the result is the same with ordinary linear regression. $\lambda = \infty$ then we have result $\beta = 0$.

9. No, since $\beta^T\beta$ is always positive, cost function will always be less if $\lambda = 0$.

3 04 Error 5

1. As in Table 1.

2. As in Table 1. They are acceptably close. Generally AP is greater than AUC-PR since precision is generally decreasing.

3. AUC-PR: 0.6794, AP: 0.7166.

4.

```
import numpy as np
def calculate(label, score,target):
    if target=="AUC-PR":
        n=len(label)
        AUCPR=np.zeros(n)
        precision=np.zeros(n+1)
        recall=np.zeros(n+1)
        precision[0]=1
        recall[0]=0
        for i in range(1,n+1):
            precision[i]=np.sum(label[0:i])/i
            recall[i]=np.sum(label[0:i])/np.sum(label)
            AUCPR[i-1]=(recall[i]-recall[i-1])*(precision[i]+precision[i-1])/2
        return np.sum(AUCPR)
    elif target=="AP":
```

```python
        n=len(label)
        AP=np.zeros(n)
        precision=np.zeros(n+1)
        recall=np.zeros(n+1)
        precision[0]=1
        recall[0]=0
        for i in range(1,n+1):
            precision[i]=np.sum(label[0:i])/i
            recall[i]=np.sum(label[0:i])/np.sum(label)
            AP[i-1]=(recall[i]-recall[i-1])*precision[i]
        return np.sum(AP)
    else:
        return 0
```

4 04 Error 6

1.
$$\begin{aligned}
&E\left[(y-f(x;D)^2\right]\\
=&E[](F(x)-f(x;D)+\varepsilon)^2]\\
=&E\left[(F(x)-E_Df(x;D))^2\right]+E\left[(E_Df(x;D)-f(x;D))^2\right]+\sigma^2
\end{aligned}$$

2.
$$\begin{aligned}
&E[f]\\
=&E[\frac{1}{k}\sum_{i=1}^{k}y_{nn(i)}]\\
=&E[\frac{1}{k}\sum_{i=1}^{k}F(x_{nn(i)})+\varepsilon]\\
=&\frac{1}{k}\sum_{i=1}^{k}F(x_{nn(i)})
\end{aligned}$$

3.
$$\begin{aligned}
&E[(F(x)-E_Df(x;D))^2]+E\left[(E_Df(x;D)-f(x;D))^2\right]+\sigma^2\\
=&E\left[(F(x)-\frac{1}{k}\sum_{i=1}^{k}F(x_{nn(i)}))^2\right]+E\left[\left(\frac{1}{k}\sum_{i=1}^{k}F(x_{nn(i)})-\frac{1}{k}\sum_{i=1}^{k}y_{nn(i)}\right)^2\right]+\sigma^2\\
=&E\left[(F(x)-\frac{1}{k}\sum_{i=1}^{k}F(x_{nn(i)}))^2\right]+E[(\sum_{i=1}^{k}\varepsilon_i)^2]+\sigma^2\\
=&E\left[(F(x)-\frac{1}{k}\sum_{i=1}^{k}F(x_{nn(i)}))^2\right]+k\sigma^2+\sigma^2
\end{aligned}$$

4. $k\sigma^2$. It grows linearly.

5. $E\left[(F(x)-\frac{1}{k}\sum_{i=1}^{k}F(x_{nn(i)}))^2\right]$. When k=n, this is $Var[F(x)]$. Also, as k grows, the squared bias term grows from 0 to $Var[F(x)]$.

5 05 PCA 5

1. Let $\{e_1, e_2, \cdots, e_n\}$ be n unit eigenvector for G.

$$\begin{aligned}
||Gx|| &= ||G \cdot (x_1 e_1 + \cdots + x_n e_n)|| \\
&= ||\lambda_1 e_1 x_1 + \cdots + \lambda_n e_n x_n|| \\
&= |x_1 \lambda_1|^2 + \cdots + |x_n \lambda_n| \\
&= |x_1|^2 + \cdots + |x_n|^2 \\
&= ||x||
\end{aligned}$$

   G^T is also orthogonal so the result holds.

2.

$$\begin{aligned}
||G^T X G||_F &= \sqrt{tr(G^T X G (G^T X G)^T)} \\
&= \sqrt{tr(G^T X G G^T X^T G)} \\
&= \sqrt{tr(G^T X X^T G)} \\
&= \sqrt{tr(G^{-1} X X^T G)} \\
&= \sqrt{tr(X X^T)} \\
&= ||X||_F
\end{aligned}$$

3. This generally accumulates X to diagonal entries, which is just approximate diagonalization. Eigenvalues appear naturally.

4. If $X_{ii} = a$, $X_{jj} = b$, $X_{ij} = X_{ji} = c$, consider

$$P \equiv P(i,j,\theta) = \begin{bmatrix}
1 & & & & & & & & \\
& \ddots & & & & & & & \\
& & 1 & & & & & & \\
& & & \cos\theta & \cdots & \sin\theta & & & \\
& & & \vdots & \ddots & \vdots & & & \\
& & & -\sin\theta & \cdots & \cos\theta & & & \\
& & & & & & 1 & & \\
& & & & & & & \ddots & \\
& & & & & & & & 1
\end{bmatrix}$$

   where $\theta = \frac{1}{2}\arctan\frac{2c}{b-a}$. Denote P's column vector as P_i,

$$\begin{aligned}
(P^T X P)_{ij} &= P_i X P_j \\
&= \cos\theta sin\theta X_{ii} - \cos\theta sin\theta X_{jj} + \cos^2\theta X_{ij} - sin^2\theta X_{ji} \\
&= \cos\theta sin\theta a - \cos\theta sin\theta b + (\cos^2\theta - sin^2\theta)c \\
&= \frac{a-b}{2} sin2\theta + cos2\theta c \\
&= \frac{a-b}{2}\frac{2c}{b-a} cos2\theta + cos2\theta c \\
&= 0
\end{aligned}$$

   The result holds for $(P^T X P)_{ji}$.

5. Since $P^T X P$ does not change F norm, it suffice to prove one iteration will not decrease $\sum_{i=1}^n X_{ii}^2$. Only X_{ii} and

X_{jj} will change after operation by $P(i,j,\theta)$. Thus the increment will be

$$
\begin{aligned}
&(\cos^2\theta x_{ii} + sin^2\theta x_{jj} + \cos\theta\sin\theta(x_{ij}+x_{ji}))^2 + (\cos^2\theta x_{jj} + sin^2\theta x_{ii} - \cos\theta\sin\theta(x_{ij}+x_{ji}))^2 - x_{ii}^2 - x_{jj}^2 \\
=&\left(\cos^4\theta + \sin^4\theta\right)x_{ii}^2 + \left(\cos^4\theta + \sin^4\theta\right)x_{jj}^2 + \left(4\cos^3\theta\sin\theta - 4\cos\theta\sin^3\theta\right)(x_{ii}-x_{jj})x_{ij} \\
&+ 8\cdot cos^2\theta sin^2\theta x_{ij}^2 + 4\sin^2\theta\cos^2\theta x_{ii}x_{jj} \\
=&8cos^2\theta sin^2\theta x_{ij}^2 + 2sin2\theta cos2\theta(x_{ii}-x_{jj})x_{ij} - (x_{ii}-x_{jj})^2 2cos^2\theta sin^2\theta \\
=&8cos^2\theta sin^2\theta x_{ij}^2 + 4c^2cos^2 2\theta - 2c^2cos^2 2\theta \\
=&2c^2 \geq 0
\end{aligned}
$$

Thus proved.

6. From increment computed in 5, off(X) decrease strictly before off-diagonal entries become all 0, and off(X) converges to 0. After some iterations, off$(X) < \varepsilon$. Then $(P^TXP)_{ij} = \cos^2\theta x_{ii} + sin^2\theta x_{jj} + \cos\theta\sin\theta(x_{ij}+x_{ji}) = (1+O(\varepsilon))x_{ii} + O(\varepsilon)x_{jj} + 2c\cdot O(\varepsilon)$. $|(P^TXP)_{ij} - X_{ij}| = O(\varepsilon)$, thus we can make sure that X converges.