CSE6740 Computational Data Analysis

Homework 1

Kai Lu - 9/17/16

1.Probability

(a)

X: resigning employee working in Store C
Y: an employee is a woman.

$$P(X \cap Y) = \frac{100}{225} \cdot 70\% = \frac{70}{225}$$

$$P(Y) = \frac{50}{225} \cdot 50\% + \frac{75}{225} \cdot 60\% + \frac{100}{225} \cdot 70\% = \frac{130}{225}$$

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{\frac{70}{225}}{\frac{130}{225}} = \frac{7}{13}$$

(b)

$$p = 0.5\% * 95\% + 99.5\% * 1\% = 1.47\%$$

(c)

$$p = (1/2)^3 + (1/2)^3 * 6 * \left[(1/2)^3 * 6 + (1/2)^3 * 2 * (1/2) \right]$$
$$= 25/32$$

(d)

$$p = (1/2)^3 * 6 * (1/2)^3 * 2 = 3/16$$

2.Maximum Likehood

(a)

$$L(\theta \mid x_1,\cdots,x_n)=\prod_{i=1}^{n} f(x_i \mid \theta)$$

$$\therefore \ln L(\theta \mid x_1,\cdots,x_n)=\ln(\prod_{i=1}^{n} f(x_i \mid \theta))=\sum_{i=1}^{n}\ln f(x_i \mid \theta)=\sum_{i=1}^{n}(x_i \ln\lambda-\lambda-\ln x_i!)$$

$$=\ln\lambda\sum_{i=1}^{n}x_i-n\lambda-\sum_{i=1}^{n}\ln x_i!$$

$$\therefore \frac{\partial \ln L}{\partial\lambda}=\frac{1}{\lambda}\sum_{i=1}^{n}x_i-n=0$$

$$\therefore \lambda=\frac{1}{n}\sum_{i=1}^{n}x_i$$

(b)

$$L(\theta_1,\cdots,\theta_k \mid n,x_1,\cdots,x_k)=f(x_1,\cdots,x_k;n,\theta_1,\cdots,\theta_k)=\frac{n!}{x_1!x_2!\cdots x_k!}\prod_{j=1}^{k}\theta_j^{x_j}$$

$$l(\theta_1,\cdots,\theta_k)=\ln(\frac{n!}{x_1!x_2!\cdots x_k!}\prod_{j=1}^{k}\theta_j^{x_j})=\ln(\frac{n!}{x_1!x_2!\cdots x_k!})+\sum_{j=1}^{k}x_j\ln\theta_j$$

Since $\sum_{j=1}^{k}\theta_j=1$, introducing Lagrange Multiplier:

Note $Lag(\theta_1,\cdots,\theta_k,\lambda)=\ln(\frac{n!}{x_1!x_2!\cdots x_k!})+\sum_{j=1}^{k}x_j\ln\theta_j+\lambda(1-\sum_{j=1}^{k}\theta_j)$

$$\therefore \frac{\partial Lag(\theta_1,\cdots,\theta_k,\lambda)}{\partial\theta_j}=x_j/\theta_j-\lambda=0(j=1,\cdots,k) \text{ and } \frac{\partial Lag(\theta_1,\cdots,\theta_k,\lambda)}{\partial\lambda}=1-\sum_{j=1}^{k}\theta_j=0$$

$$\because \sum_{j=1}^{k}x_j=n$$

$$\therefore \lambda=n,\ \theta_j=\frac{x_j}{n}$$

(c)

$$L(\mu,\sigma^2 \mid x_1,\cdots,x_n)=\prod_{i=1}^{n} f(x_i \mid \mu,\sigma^2)$$

$$\therefore \ln L(\mu,\sigma^2 \mid x_1,\cdots,x_n)=\ln(\prod_{i=1}^{n} f(x_i \mid \mu,\sigma^2))=\sum_{i=1}^{n}\ln f(x_i \mid \mu,\sigma^2)=\sum_{i=1}^{n}(-\frac{(x_i-\mu)^2}{2\sigma^2}-\ln(\sigma\sqrt{2\pi}))$$

$$=-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}-\frac{n}{2}\ln(2\pi\sigma^2)$$

suppose $\frac{\partial \ln L}{\partial \mu}=\frac{-1}{2\sigma^2}\sum_{i=1}^{n}2(\mu-x_i)=0$ and $\frac{\partial \ln L}{\partial \sigma^2}=\frac{1}{\sigma^4}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2}-\frac{n}{2\sigma^2}$

$$\therefore \mu=\frac{1}{n}\sum_{i=1}^{n}x_i \ , \ \sigma^2=\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n}$$

3.Principal Component Analysis

(a)

Since

$$J=\frac{1}{N}\sum_{n=1}^{N}\left\|x^n-\tilde{x}^n\right\|^2=\frac{1}{N}\sum_{n=1}^{N}(x^n-\tilde{x}^n)^T(x^n-\tilde{x}^n)=\frac{1}{N}\sum_{n=1}^{N}(x^{nT}-\tilde{x}^{nT})(x^n-\tilde{x}^n)$$

$$=\frac{1}{N}\sum_{n=1}^{N}(x^{nT}x^n+\tilde{x}^{nT}\tilde{x}^n-x^{nT}\tilde{x}^n-\tilde{x}^n x^{nT})$$

To minimize J,

$$\frac{\partial J}{\partial z_j^n}=2z_j^n\left\|u_j\right\|^2-2\alpha_j^n\left\|u_j\right\|^2=0$$

$$\therefore z_j^n=\alpha_j^n=x^{nT}u_i \text{ for } j=1,....,M$$

(b)

$$J=\frac{1}{N}\sum_{n=1}^{N}(x^{nT}x^n+\tilde{x}^{nT}\tilde{x}^n-x^{nT}\tilde{x}^n-\tilde{x}^n x^{nT})$$

To minimize J,

$$\frac{\partial J}{\partial b_j}=\frac{1}{N}\sum_{n=1}^{N}(2b_j\left\|u_j\right\|^2-2\alpha_j^n\left\|u_j\right\|^2)=0$$

$$\therefore Nb_j=\sum_{n=1}^{N}\alpha_j^n \text{ for } j=M+1,....,D$$

$$\therefore b_j=\frac{1}{N}\sum_{n=1}^{N}\alpha_j^n=\bar{x}^T u_j \text{ for } j=M+1,....,D$$

(c)

$$\tilde{x}^n = \sum_{i=1}^{M} (x^{nT} u_i) u_i + \sum_{i=M+1}^{D} (\overline{x}^T u_i) u_i$$

$$x^n - \tilde{x}^n = \sum_{i=M+1}^{D} [(x^{nT} - \overline{x}^T) u_i] u_i$$

(d)

$$J = \frac{1}{N} \sum_{n=1}^{N} \left\| x^n - \tilde{x}^n \right\|^2 = \frac{1}{N} \sum_{n=1}^{N} \left\| \sum_{i=M+1}^{D} [(x^{nT} - \overline{x}^T) u_i] u_i \right\|^2$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{i=M+1}^{D} \left\| [(x^{nT} - \overline{x}^T) u_i] u_i \right\|^2$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{i=M+1}^{D} [(x^{nT} - \overline{x}^T) u_i]^2 \left\| u_i \right\|^2$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{i=M+1}^{D} [(x^{nT} - \overline{x}^T) u_i]^2 (\left\| u_i \right\|^2 = 1)$$

$$= \sum_{i=M+1}^{D} u_i^T S u_i \quad (\left\| u_i \right\|^2 = 1, S = \frac{1}{N} \sum_{n=1}^{N} (x^n - \overline{x})(x^n - \overline{x})^T)$$

To minimize J, introduce *Lagrange multiplier* λ_i

$$\tilde{J} = \sum_{i=M+1}^{D} u_i^T S u_i + \lambda_i (1 - u_i^T u_i) \quad (\left\| u_i \right\|^2 = 1, S = \frac{1}{N} \sum_{n=1}^{N} (x^n - \overline{x})(x^n - \overline{x})^T)$$

$$\frac{\partial \tilde{J}}{\partial u_i} = u_i^T S - \lambda_i u_i^T = 0$$

$$\therefore S u_i = \lambda_i u_i$$

$\therefore u_i$ should be an eigenvector of S, and λ_i is the corresponding eigenvalue.$(\left\| u_i \right\|^2 = 1,)$

4.Clustering

(a)

$$J=\sum_{n=1}^{N}\sum_{k=1}^{K}r^{nk}\left\|x^n-\mu^k\right\|^2=\sum_{n=1}^{N}\sum_{k=1}^{K}r^{nk}(x^n-\mu^k)^T(x^n-\mu^k)=\sum_{n=1}^{N}\sum_{k=1}^{K}r^{nk}(x^{nT}x^n-2x^{nT}\mu^k+\mu^{kT}\mu^k)$$

$$\frac{\partial J}{\partial \mu^k}=\sum_{n=1}^{N}(-2r^{nk}x^{nT}+2r^{nk}\mu^{kT})=0$$

$$\therefore \mu^{kT}=\frac{\sum_{n=1}^{N}r^{nk}x^{nT}}{\sum_{n=1}^{N}r^{nk}}$$

$$\therefore \mu^{k}=\frac{\sum_{n=1}^{N}r^{nk}x^{n}}{\sum_{n=1}^{N}r^{nk}}$$

(b)

Claim: Each iteration of the K-means algorithm decreases the objective:

$$J=\sum_{n=1}^{N}\sum_{k=1}^{K}r^{nk}\left\|x^n-\mu^k\right\|^2$$

Proof:

(1) Cluster assignment:

$$r^{nk}=\begin{cases}1 & \text{if } k=\arg\min_{j=1,\cdots,K}\left\|x^i-\mu^j\right\|^2\\ 0 & \text{otherwise}\end{cases}$$ which decreases at least one data point i's $\sum_{k=1}^{K}r^{ik}\left\|x^i-\mu^k\right\|^2$

which in turn would decrease J

(2) Center adjustment:

$$\mu^k=\frac{\sum_{n=1}^{N}r^{nk}x^n}{\sum_{n=1}^{N}r^{nk}}=\text{argmin}_{\mu^k}\sum_{n=1}^{N}r^{nk}\left\|x^n-\mu^k\right\|^2$$ which in turn would decrease J

$\therefore$ Claim is proven.

$\therefore$ K-means algorithm converges to a local optimum in finite steps

(c)

The third(Average linkage) would most likely result in clusters most similar to K-means.

Because both single and complete linkage **only focus on one pair of datapoint** from the two clusters and would inevitably ignore the other datapoint in the clusters. Meanwhile, average linkage **takes all the datapoint into consideration**, just like K-means does.

(d)

The single linkage would work in such case.
Because **single linkage focuses more on connectivity** instead of similarity between two clusters than other linkages, which is exactly what we want in this case.

5.Programming: Image compression

(a)

Detailed implementation of K-medoids:
1) choose representatives for each cluster
At the beginning, the medoids are randomly determined.
Then, in every iteration, for each cluster, I randomly pick a non-medoid point in the cluster, and check if the point becomes medoid of the cluster, whether the distortion would decrease or not. If it decreases, I would update the new medoid.
2) distance measures
I have tried 1-norm, 2-norm and 3-norm as distance measurement.
And I chose 1-norm.
3) when to stop
The iterations would stop under either condition: (1) the # of iterations reaches 500; (2) no medoids change in the iteration

(b)

I would use this picture with a size of 320x240.

(c)

The pictures on the right show the results of K-medoid with K=2,3,4 and 5(from top to bottom).
With K increases, the picture gets more details and the colors are more similar to the original ones.
The time they take are 76s, 112s, 147s and 172s.

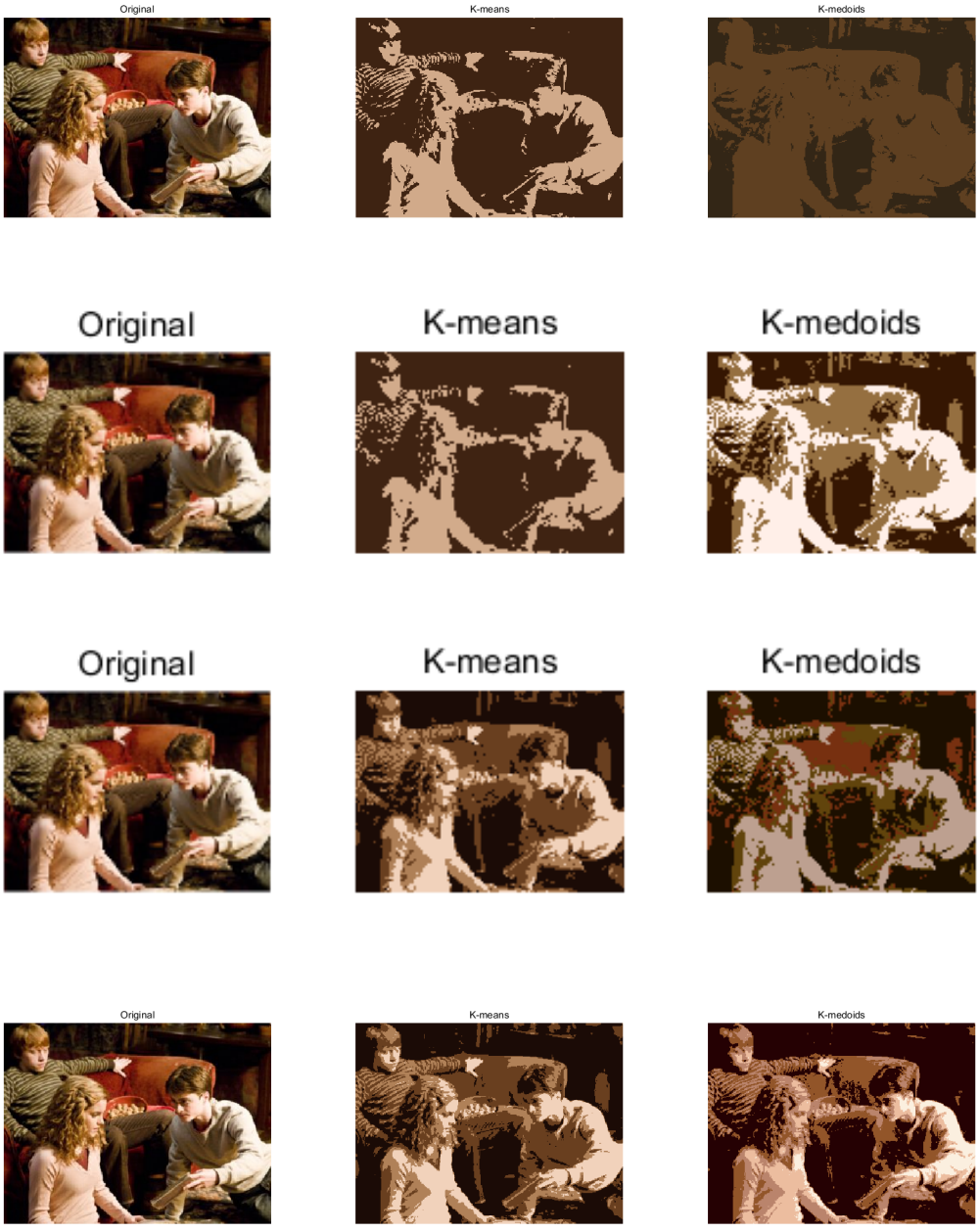
(d)

The initial centroids are randomly chosen so that every run has a different initialization.
It would affect the final result. And different initialization would end up with similar but different results.

(e)

The following pictures are the comparison among original, K-means and K-medoids:

For K-means, (1) the different initializations also affect the results. (2) increasing K would improve the quality of pictures. (3) The time they take are 2.5s, 1.9s, 6.9s and 7.0s.



In generally, the two methods have very close performance in quality and robustness. However, in terms of runtime, K-means has a great advantage over K-medoids.