CSE/ISYE 6740 Homework 1 (Le Song)

Siying Cen

deadline: September 26th

1 Probability

(a) Answer:
Number of woman in store A, B, C:
N(A) = 50 x 50% = 25, N(B) = 75 x 60% = 45, N(C) = 100 x 70% = 70

$$P(woman\ in\ store\ C) = \frac{70}{25 + 45 + 70} = 0.5$$

(b) Answer:
P(true) = 0.5%, P(positive—false) = 1%, P(positive—true) = 95%

$$P(true|positive) = \frac{P(positive|true)P(true)}{P(positive)} = \frac{P(positive|true)P(true)}{P(positive|true)P(true) + P(positive|false)P(false)}$$

$$= \frac{95\% \times 0.5\%}{95\% \times 0.5\% + 1\% \times (1 - 0.5\%) \times 1\%} = 0.323$$

(c) Answer:
A for Atlanta Braves, SF for San Francisco Giants and L for Los Angeles Dodgers.
P(Braves win) = P(A=90) + P(A=89, SF=88 or 87) + P(A=89, SF=89 or LA=89, A win playoff) + P(A=88, SF=88 or LA=88, A win playoff)

P(A=90) = $(\frac{1}{2})^3 = \frac{1}{8}$,   P(A=89, SF=88 or 87) = $(3 \cdot (\frac{1}{2})^3 \cdot 3 \cdot (\frac{1}{2})^3) \cdot 2 = \frac{9}{32}$,

P(A=89, SF=89 or LA=89, A win playoff) = $(3{\cdot}(\frac{1}{2})^3 \cdot (\frac{1}{2})^3 \cdot \frac{1}{2}) \cdot 2 = \frac{3}{64}$,

$P(A = 88, SF = 88orLA = 88, Awinplayoff) = (3 \cdot (\frac{1}{2})^3 \cdot 3 \cdot (\frac{1}{2})^3 \cdot \frac{1}{2}) \cdot 2 = \frac{9}{64}$

$$P(Braves\ win) = \frac{1}{8} + \frac{9}{32} + \frac{3}{64} + \frac{9}{64} = \frac{19}{32}$$

(d) Answer:
P(playoff game) = P(A=89, SF=89 or LA=89) + P(A=88, SF=88 or LA=88)

$= (3\cdot(\frac{1}{2})^3 \cdot (\frac{1}{2})^3) \cdot 2) + (3 \cdot (\frac{1}{2})^3 \cdot 3 \cdot (\frac{1}{2})^3) \cdot 2 = \frac{3}{32} + \frac{9}{32} = \frac{3}{8}$

2 Maximum Likelihood

2.1 Poisson distribution

The Poisson distribution is defined as

$$\mathrm{P}(x = k) = \tfrac{\lambda^k e^{-\lambda}}{k!} \quad (k = 0, 1, 2, ...)$$

$$\begin{gathered} l\,(\lambda) = log(\textstyle\prod_{i=1}^{N} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}) = \sum_{i=1}^{N} x_i ln\lambda - n\lambda - \sum_{i=1}^{N} ln(x_i!) \\ \text{let } l\,'(\lambda) = -n + \textstyle\sum_{i=1}^{N} \frac{x_i}{\lambda} = 0 \\ \Rightarrow estimator \hat{\lambda} = \textstyle\sum_{i=1}^{N} \frac{x_i}{n} = \overline{x} \end{gathered}$$

(2)

2.2 Multinomial distribution

$$f(x_1, x_2, , ..., x_k; n, \theta_1, \theta_2, ..., \theta_k) = \frac{n!}{x_1!x_2! \cdot \cdot_k!} \prod_{j=1}^{k} {\theta_j}^{x_j}$$

the log-likelihood function is:

$l(\theta_j) = ln(\frac{n!}{x_1!x_2!\cdot\cdot_k!} \prod_{j=1}^{k} {\theta_j}^{x_j}) = ln(\frac{n!}{x_1!x_2!\cdot\cdot_k!}) + \sum_{j=1}^{k} x_j ln(\theta_j)$

introduce a Lagrange multiplier λ :

$\mathrm{L}(\theta_j) = l(\theta_j) + \lambda(1 - \sum_{j=1}^{k} \theta_j) = ln(\frac{n!}{x_1!x_2!\cdot\cdot_k!}) + \sum_{j=1}^{k} ln(\theta_j) - \lambda(\sum_{j=1}^{k} \theta_j - 1)$

$$let \quad \frac{\partial L(\theta_j)}{\partial \theta_j} = \frac{x_j}{\theta_j} - = 0, \frac{\partial L(\theta_j)}{\partial} = 1 - \sum_{j=1}^{k} \theta_j = 0$$

$$\Rightarrow the \quad estimator \quad \theta_j = \frac{x_j}{n}$$

2.3 Gaussian normal distribution

$$N(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp^{-(x-\mu)^2/2\sigma^2}$$

$$l(x;\mu,\sigma^2) = log(N(x;\mu,\sigma^2)) = -\frac{n}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

let

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0, \quad \frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\Rightarrow estimator \quad \hat{\mu} = \frac{x_i}{n}, \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$$

3 Principal Component Analysis

(a) What is the assignment of z_j^n for $j = 1, ..., M$ minimizing J?

answer:

$$J = \frac{1}{N}\sum_{n=1}^{N} \parallel x^n - \hat{x}^n \parallel^2 \ = \frac{1}{N}\sum_{n=1}^{N}((x^n)^T x^n - 2(x^n)^T \hat{x}^n)^T + (\hat{x}^n)^T \hat{x}^n)$$

$$\frac{\partial J}{\partial z_j^n} = \frac{\partial}{\partial z_j^n}\frac{1}{N}(-2(x^n)^T \hat{x}^n)^T + (\hat{x}^n)^T \hat{x}^n)$$

$let \quad \frac{\partial}{\partial z_j^n}\frac{1}{N}(-2(x^n)^T \hat{x}^n)^T + (\hat{x}^n)^T \hat{x}^n)$

$= \frac{\partial}{\partial z_j^n}(-2 z_j^n (x^n)^T \mu_j + \sum_{i,j}^{M} z_i^n z_j^n \mu_i^T \mu_j + 2\sum_{i=1}^{M}\sum_{j=M+1}^{D} z_i^n b_j \mu_i^T \mu_j)$

$= 0$,
Since $\mu_i^T \mu_j = 0$ for $i \neq j$, and $\mu_i^T \mu_j = 1$ for $i = j$,

$$\frac{\partial}{\partial z_j^n}(-2z_j^n(x^n)^T\mu_j + \sum_{i,j}^{M} z_i^n z_j^n \mu_i^T \mu_j + 2\sum_{i=1}^{M}\sum_{j=M+1}^{D} z_i^n b_j \mu_i^T \mu_j)$$

$$= \frac{\partial}{\partial z_j^n}(-2z_j^n(x^n)^T\mu_j + (z_i^n)^2) = 0$$

so

$$z_j^n = (x^n)^T\mu_j, j = 1, ..., M$$

(b) What is the assignment of b_j for $j = M+1, ..., D$ minimizing J?
answer:

$$\frac{\partial J}{\partial b_j^n} = \frac{1}{N}\sum_{n=1}^{N}((x^n)^T x^n - 2(x^n)^T \hat{x}^n)^T + (\hat{x}^n)^T \hat{x}^n)$$

$$= \frac{\partial}{\partial b_j^n}\frac{-2}{N}((x^n)^T b_j)^T + (\hat{x}^n)^T \hat{x}^n)$$

$$= -2\hat{x}\mu_j + \frac{1}{N}\frac{\partial}{\partial b_j^n}\sum_{n=1}^{N}(\sum_{i,j}^{M} z_i^n z_j^n \mu_i^T \mu_j + 2\sum_{i=1}^{M}\sum_{j=M+1}^{D} z_i^n b_j \mu_i^T \mu_j + \sum_{i=M+1,j=M+1}^{D} b_i b_j \mu_i^T \mu_j) = 0$$

Since $\mu_i^T \mu_j = 0$ for $i \neq j$, and $\mu_i^T \mu_j = 1$ for $i = j$,

$$\frac{\partial J}{\partial b_j^n} = -2\hat{x}\mu_j + \frac{1}{N}\frac{\partial J}{\partial b_j^n}\sum_{n=1}^{N}(b_j^2) = -2\hat{x}\mu_j + 2b_j = 0$$

$$b_j = \hat{x}^T\mu_j, j = M+1, ..., D$$

(c) Express optimal $\hat{X}^n$ and $X^n - \hat{X}^n$ using your answer for (a) and (b).
answer:

$$\hat{X}^n = \sum_{i=1}^{M}((x^n)^T\mu_i)\mu_i + \sum_{i=M+1}^{D}(\hat{x}^T\mu_i)\mu_i$$

$$x^n - \hat{X}^n = \sum_{i=M+1}^{D}((x_n - \hat{x})^T\mu_i)\mu_i$$

(d) What should be the μ_i for $i = 1, ..., D$ to minimize J ?
answer:
According to (b) and (c),
we can get

$$J = \frac{1}{N}\sum_{n=1}^{N} \parallel x^n - \hat{x}^n \parallel^2 = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=M+1}^{D} ((x_n - \hat{x})^T \mu_i)^T ((x_n - \hat{x})^T \mu_i)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{i=M+1}^{D} {\mu_i}^T (x_n - \hat{x})(x_n - \hat{x})^T \mu_i = \sum_{i=M+1}^{D} \mu_i^T S \mu_i$$

Using Lagrange multiplier, we can consider the minimization of

$$\hat{J} = \sum_{i=M+1}^{D} \mu_i^T S \mu_i + \sum_{i=M+1}^{D} (1 - \mu_i^T \mu_i)$$

$$let \quad \frac{\partial \hat{J}}{\partial \mu_i} = 0,$$

we can get

$$S\mu_i = \lambda_i \mu_i$$

μ_i should the eigenvector corresponding to the largest eigenvalue.

4 Clustering

(a) Proof:
According to K-mean clustering, the distortion function is defined as

$$J = \sum_{n=1}^{N} r^{nk} (x^n - \mu^k)^2$$

to minimize function J,

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

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

(b) answer:

K-means algorithm is optimized between the following two steps:
(1) Fix μ^k, minimize objective function J (assign points to closest centers);
(2) Fix r^{nk}, minimize objective function J (recompute the center means).
Since there are only finite assignments to r^{nk} and μ^k, there will be finite steps for K-means algorithm to converge.

(c) answer:
Average linkage will most likely result in clusters most similar to those given by K-means when applied on hierarchical clustering. Because average linkage calculate the mean distance between all pair of points from every two clusters, it is used in the centroid adjustment step in K-means clustering.

(d) answer:
Single linkage will successfully separate the two moons. We can notice that the data within one moon is very compact and close to one another, as a result, the initial clustering will start clustering between points pairs in the same moon first when using single linkage. If K =2 is chosen, single linkage clustering will separate the two moons.

5 Programming: Image compression

Report

1. My K-medoids framework:
First, I select K data points randomly from the pixels vector as my initial centroids. Then I assigned each data point to the closest centroid by computing the minimum of distance. The data point which had the minimum distance to one centroid was chosen as the representative of this centroid. The iteration stopped when the centroids not changing or the iteration step is achieved to maximum (100).
I also used Euclidean distance, Chebyshev distance and Manhattan distance to test the performance in my algorithms.

2. The picture I used is attached as ”my_image.jpg” (627 x 347).

3. I run K-means and K-medoids with K = 2, 4, 8, 16, 32, respectively. The output figures are shown below and the elapsed time is summarized in table 5-1. The results demonstrated that the performance of both K-means and K-medoids increase as number of K increase. However, it takes longer time to converage

for larger K's.
4. I set an intentional poor assignment manully by changing the randomly selected initial centroid to pixels [1,K,3] (K = 4). It affect the result of K-means, but does not affect the result of K-medois.
5. The quality beween K-means and K-medois are similar, however when K is large ($K > 4$), the running time of K-medoids is shorter (see table 5-1). And I also compare the results of choosing different types of distances among 'eluclidean', 'chebychev' and 'manhattan' distances. There is no big difference between these different methods.

Note: Please check figures and tables in "Figures and table.pdf"!!!