CS 4641/7641 **Full Name:** _____
Fall 2019
Midterm Exam
10/10/2019
Time Limit: 75 Minutes GT Username _____

This exam contains 7 pages (including this cover page) and 5 questions.
Total of points is 100.

It is an open book exam. Electronic devices and Internet usage are not allowed in this exam.

Grade Table (for instructor use only)

Question	Points	Score
1	15	
2	15	
3	25	
4	25	
5	20	
Total:	100	

1. Information Theory and Probability.

   Sentiment analysis is a popular topic in the natural language processing area. For a Yelp review classification task, we assume that in our data set, the positive and negative reviews are balanced. Thus, the probability of a review being positive is $p(Y) = 0.5$ and being negative is $p(N) = 0.5$. For Yelp reviews, **restaurant** is a common keyword. We also suppose that 20% of positive reviews have the keyword restaurant and 4% of negative reviews have the keyword restaurant. R refers to a review with the keyword restaurant. Thus, $p(R|Y) = 0.2$ and $p(R|N) = 0.04$.

   (a) (10 points) Calculate the probability $p(Y|R)$ which refers if a review has the keyword restaurant, that review must be positive. (**Hint:** Use product rule and Bayes' rule)

   Answer

   $$p(R) = p(R|Y)p(Y) + p(R|N)p(N) = (0.2)(0.5) + (0.04)(0.5) = 0.12$$

   $$p(Y|R) = \frac{p(R|Y)p(Y)}{p(R)} = \frac{(0.2)(0.5)}{(0.12)} = 0.83$$

   (b) (5 points) If you know that a review is positive, how much information do you gain (in bits) by learning that it also has the keyword restaurant?

   Answer

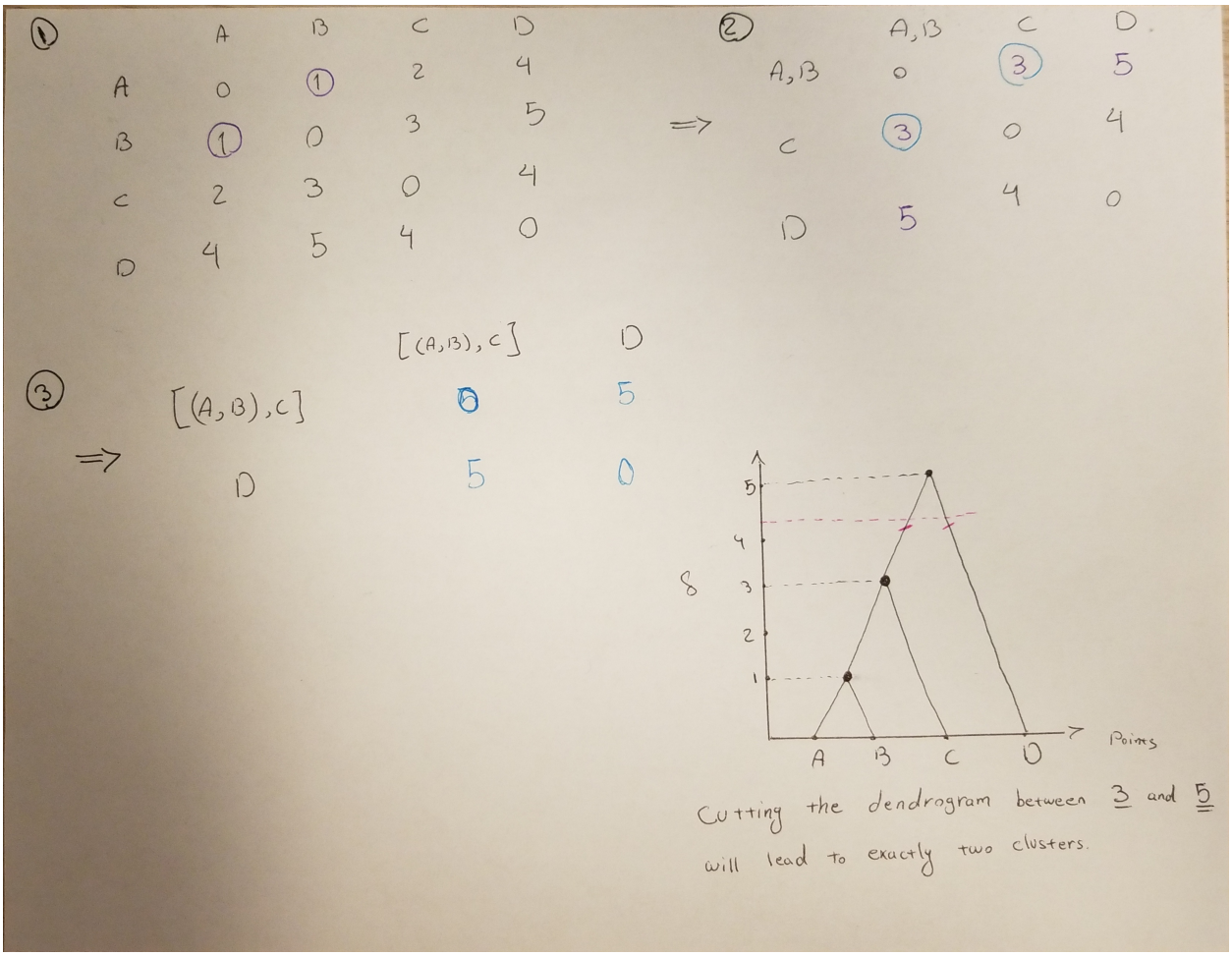
   The information gained from an event is $-log2$ of its probability. Thus the information gained from learning that a positive review has the keyword restaurant, since $p(R|Y) = 0.2$, is 2.32 bits.

2. (15 points) Using Hierarchical clustering and Complete Linkage method, plot the dendrogram. In order to have exactly **two** clusters, we should cut the dendrogram in a specific distance range. Please specify that range. [NOTE: **pairwise distance matrix** is provided here.]

	A	B	C	D
A	0	1	2	4
B	1	0	3	5
C	2	3	0	4
D	4	5	4	0

Table 1: Pairwise distance matrix for the points.

Answer



3. K-Means.

   (a) (10 points) K-Means is a hard-assignment clustering task. Please explain why? Also name two other hard-assignment clustering algorithms?

   Answer

   There is no probability associated with each data point and each point is absolutely part of a cluster.

   (b) (10 points) Expectation-Maximization (EM) approach is conceptually employed in K-Means algorithm. Which part of K-Means algorithm covers Expectation and which part covers Maximization?

Answer

The E-step, where each object is assigned to the centroid such that it is assigned to the most likely cluster. The M-step, where the model (=centroids) are recomputed (= least squares optimization).

(c) (5 points) Why different input parameters initializations in K-Means may lead to different answers? Please provide the mathematical minimization problem in K-Means and explain what we try to minimize?

Answer

K-Means does not converge to a global optimum, it converges into a local optimum.

$$\min \frac{1}{n} \sum_{i=1}^{n} ||x^i - c^{\pi(i)}||^2$$

. We minimize the averaged square distance from each data point to its respective cluster center.

4. Gaussian Mixture Model (GMM).

(a) (5 points) Is GMM a Generative or Discriminative model, why?

Answer

Generative. We are using the joint probability to solve a GMM model.

$$p(x) = \sum_{k} p(x, z_{k=1})$$

(b) (5 points) What parameters are unknown in a GMM model?

Answer

$$\mu, \Sigma, \pi$$

(c) (10 points) Consider four mixture components are employed to calculate a GMM model. Please find the parametric solution of the first mixture component responsibility $p(z_{nk}|x)$ for $k = 1$ (hint: a similar task was done in the class)?

Answer

$$p(x) = \sum_{k} p(x, z_{nk})$$

$$p(x) = p(x|z_{n1})p(z_{n1}) + p(x|z_{n2})p(z_{n2}) + p(x|z_{n3})p(z_{n3}) + p(x|z_{n4})p(z_{n4})$$

$$\sum_k p(x, z_{nk}) = \pi_1 N_1(\mu, \Sigma) + \pi_2 N_2(\mu, \Sigma) + \pi_3 N_3(\mu, \Sigma) + \pi_4 N_4(\mu, \Sigma)$$

Now, let's normalize each component by dividing it over the summation of all the components:

$$\frac{\pi_1 N_1(x|\mu, \Sigma)}{\pi_1 N_1(x|\mu, \Sigma) + \pi_2 N_2(x|\mu, \Sigma) + \pi_3 N_3(x|\mu, \Sigma) + \pi_4 N_4(x|\mu, \Sigma)} =$$

$$= \frac{p(x|z_{n1})p(z_{n1})}{p(x|z_{n1})p(z_{n1}) + p(x|z_{n2})p(z_{n2}) + p(x|z_{n3})p(z_{n3}) + p(x|z_{n4})p(z_{n4})}$$

$$= \frac{p(x, z_{n1})}{\sum_k p(x, z_k)} = p(z_{n1}|x)$$

The above result is also called inferring which gives responsibility value for the first mixture component.

(d) (5 points) How do we check the convergence of EM in GMM? Please also provide the mathematical term.

Answer

Log likelihood.

$$L(X|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)$$

5. Clustering Evaluation

In this section, we focus on evaluating the performance of the clustering results using different internal measurements. To answer the following questions, you may require the pairwise distance matrix as shown in Table 2 and cluster assignments in Table 3. Your answer should contain the intermediate computation steps, and partial credits will be given even if your final answer is wrong. Your answer is expected to match the true answer up to 2 digits after the decimal point (or you can keep the fraction, if necessary).

	A	B	C	D	E
A	0.0	1.0	1.0	2.0	3.0
B	1.0	0.0	1.5	2.5	3.5
C	1.0	1.5	0.0	1.0	2.0
D	2.0	2.5	1.0	0.0	1.0
E	3.0	3.5	2.0	1.0	0.0

Table 2: Pairwise distance matrix for the points.

Solution 1

point index	A	B	C	D	E
cluster index	0	0	0	1	1

Solution 2

point index	A	B	C	D	E
cluster index	0	0	1	1	1

Table 3: Two cluster assignment solutions for the points.

(a) (10 points) What are the Beta-CV measures for the two solutions, respectively? And according to this measure, which solution is better?

Let's see an example for how to calculate the Beta-CV for Solution 1.

$$W_{in} = \frac{1}{2}\sum_{i=0}^{1} W(C_i, C_i) = \frac{1}{2}[W(C_0, C_0) + w(C_1, C_1)] =$$

$$[(0.0 + 0.0 + 0.0 + 1.0 + 1.0 + 1.5) + (0.0 + 0.0 + 1.0)] = 4.5$$

$$W_{out} = \frac{1}{2}\sum_{i=0}^{1} W(C_i, \overline{C_i}) = 2.0 + 3.0 + 2.5 + 3.5 + 1.0 + 2.0 = 14.0$$

$$N_{in} = \sum_{i=1}^{2} \binom{n_i}{2} = \binom{3}{2} + \binom{2}{2} = 4$$

$$N_{out} = \sum_{i=0}^{0}\sum_{j=1}^{1} n_i \cdot n_j = 3 \times 2 = 6$$

Then, we can use

$$\text{BetaCV} = \frac{W_{in}/N_{in}}{W_{out}/N_{out}} = \frac{N_{out}}{N_{in}} \cdot \frac{W_{in}}{W_{out}}.$$

Answer

For the first solution $N_{in} = 4$ $N_{out} = 6$ $W_{in} = 4.5$ $W_{out} = 14.0$,
so $BetaCV_{solution_1} = 0.48$

For the second solution $N_{in} = 4$ $N_{out} = 6$ $W_{in} = 5.0$ $W_{out} = 13.5$,
so $BetaCV_{solution_2} = 0.56$.
The first one is better.

(b) (10 points) What are the Silhouette Coefficients for the two solutions, respectively? And according to this measure, which solution is better?

An example of how to calculate the Silhouette Coefficient for a data-point (A) in the first clustering solution: $\mu_{out}^{min} = (2.0 + 3.0)/2.0 = 2.5$, $\mu_{in} = (1.0 + 1.0)/2.0 = 1.0$, thus $s_A = (2.5 - 1.0)/2.5 = 0.6$.

Answer

For solution 1:

$$SC = \frac{1}{5}\sum_{i=1}^{5} S_i = \frac{0.6 + 0.58 + 0.17 + 0.45 + 0.65}{5} = 0.49$$

For solution 2:

$$SC = \frac{1}{5}\sum_{i=1}^{5} S_i = \frac{0.5 + 0.6 + 0.56 - 0.17 + 0.54}{5} = 0.41$$

The first one is better.