IERG4300 / ESTR4300/ IEMS5709 Fall 2021 Homework 3

Release date: Nov 3, 2021

Due date: Nov 20, 2021 (Saturday) 11:59pm

The solution will be posted right after the deadline, so no late homework will be accepted!

Every Student MUST include the following statement, together with his/her signature in the submitted homework.

I declare that the assignment submitted on the Elearning system is original except for source material explicitly acknowledged and that the same or related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the website http://www.cuhk.edu.hk/policy/academichonesty/.

Signed (Student__________________________) Date:________________________________

Name___________________________________ SID ________________________________

Submission notice:

- Submit your homework via the elearning system

General homework policies:

A student may discuss the problems with others. However, the work a student turns in must be created COMPLETELY by oneself ALONE. A student may not share ANY written work or pictures, nor may one copy answers from any source other than one's own brain.

Each student **MUST LIST** on the homework paper the **name of every person he/she has discussed or worked with**. If the answer includes content from any other source, the student **MUST STATE THE SOURCE**. Failure to do so is cheating and will result in sanctions. Copying answers from someone else is cheating even if one lists their name(s) on the homework.

If there is information you need to solve a problem but the information is not stated in the problem, try to find the data somewhere. If you cannot find it, state what data you need, make a reasonable estimate of its value and justify any assumptions you make. You will be graded not only on whether your answer is correct but also on whether you have done an intelligent analysis.

Q1 [30 marks]: Parameter Design for Minhash/ Locality-Sensitive Hashing (LSH)

Let ***r*** be the number of rows within each band and ***B*** be the total number of bands within the Minhash signature matrix ***M***. We want to design the system so that:

1) For any pair of items with similarity greater than or equal to ***T1***, the probability that they will be correctly identified as a similar-pair candidate should be at least ***P1***.
2) For any pair of items with similarity below ***T2***, the probability that they will be mistakenly identified as a similar-pair candidate should be no more than ***P2***.

(a) **[10 marks]** Derive the set of inequalities to govern the relationship between T1, T2, P1, P2, r and B so that the aforementioned accuracy/error requirements would be satisfied.

(b) **[10 marks]** By graphing the inequality of part (a) in a 2-D plot. one can determine the corresponding values of r and b so that a given set of accuracy/ error requirements (defined in terms of T1, T2, P1, P2) can be satisfied. Describe and illustrate such an approach.

(c) **[10 marks]** For T1=0.8, T2=0.3, P1=0.95 and P2=0.05, use your results in part (a) to derive a single pair of values for (r, B) so that the aforementioned accuracy/error requirements would be satisfied. In general, there can be multiple feasible solutions. You only need to produce one pair of solutions. Show your steps (and source code if any).

Q2 [50 marks]: k-means Clustering

The MNIST database is a dataset of handwritten digits, comprising 60000 training examples and 10000 test examples. In this question, we will implement the k-means algorithm using the test set of MNIST dataset. The data can be downloaded here: http://yann.lecun.com/exdb/mnist/. The MNIST test set contains 10 various digits, totalling 10000 instances, with representative images shown in Fig. 1. Each of the digits is a 28x28 pixel image, resulting in a 784-dimensional space.

The training set contains two files:

(1) *train-images-idx3-ubyte*: training set images (9912422 bytes)

(2) *train-labels-idx1-ubyte*: training set labels (28881 bytes)

And the testing set contains the following 2 files:

(3) *t10k-images-idx3-ubyte*: testing set images (1648877 bytes)

(4) *t10k-labels-idx1-ubyte*: testing set labels (4542 bytes)

File (1) contains image instances. Rows are images and columns are pixels with values from 0 to 255. File (2) contains the ground truth labels of images in (1). Similarly, File (3) contains the testing image instances and file (4) contains the true labels of images in (3). You can get more detailed information of the data from http://yann.lecun.com/exdb/mnist/

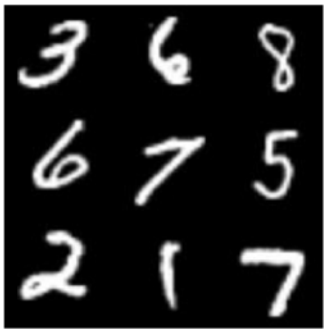

Fig. 1. An image of representative digits from the MNIST dataset.

(a) **[20 marks]** Refer to the lecture notes, you will need to implement K-means using MapReduce to perform clustering for the training-set. For this question, the number of clusters is 10. Please output the vector representation of each centroid and the number of digit image assigned to each cluster like the following example:

[Centroid ID]: [Centroid Vector Representation], [Number of Digit Images]

Centroid 0: [784-dimension variable], 1000

Centroid 1: [784-dimension variable], 2000

……

Centroid 10: [784-dimension variable], 3000

Submit both your code and results.

Hint:

1. You should use multiple rounds MapReduce to implement the K-means algorithm. For each round, each mapper processes part of the training data points and stores all the current centroids. The reducers will update the centroids based on “partial sum” information transferred from all the mappers.
2. Before the last round of the training process, there is actually no need to remember (or store) the cluster-membership assignment of each training data point. You just need to keep track of enough information to enable the computation of the centroid at the reducer (e.g., the partial sum of the training data points in a mapper for each cluster, and the number of training data points in a mapper that is assigned to each cluster, etc.). You need to store/ output the cluster membership assignment for each data point after k-means has converged. Or you can implement another program to assign each training data point to the final clusters produced by k-means.

(b) **[20 marks]** In the simplest k-means algorithm, we initialize the centroids randomly. However, we know that bad choice of centroids will lead to suboptimal clustering. One approach to avoid this situation is to test out different centroid initializations and then choose an initialization that shows the best performance.

We will run k-means 3 times in this question and each operation with different random centroid initialization. By utilizing the clustering results in these 3 times together with the ground truth label of the training set, we can calculate the accuracy of the clustering results (**training sets**). The ground truth labels of training images are stored in the file train*-labels-idx1-ubyte,* so you can compare the results with the labels.

1. Find the ground truth label of each image from file *train-labels-idx1-ubyte.*
2. Select the major label with most images of each cluster as the label of the cluster.
3. Calculate the ratio of correctly clustered images to total images: if the ground truth label of an image is the same with its cluster label, then it is correctly clustered. Otherwise, the image is clustered incorrectly.

4. Report the accuracy performance in the tables below with different “centroid initialization”.
5. Compare the results in Table 1 to Table 3, determine the best random seed. Explain your choice.
6. The submitted result should be in the same format as Table 1 to Table 3.

[Note: For part (b), you can implement the program using a single machine or MapReduce.]

Submit both your codes and results.

Table. 1. The Accuracy of Clustering Performance with Random Seed 1

Cluster Number	# Train images belongs to the cluster	Major Label of central images	# Correctly clustered images	Classification Accuracy (%)
0				
1				
……				
9				
Total Set		NA		

Table. 2. The Accuracy of Clustering Performance with Random Seed 2

Cluster Number	# Train images belongs to the cluster	Major Label of central images	# Correctly clustered images	Classification Accuracy (%)
0				
1				
……				
9				
Total Set		NA		

Table. 3. The Accuracy of Clustering Performance with Random Seed 3

Cluster Number	# Train images belongs to the cluster	Major Label of central images	# Correctly clustered images	Classification Accuracy (%)
0				
1				
……				

9				
Total Set		NA		

(c) **[10 marks]** In classification works, researchers always utilize n-fold cross validation to further evaluate the performance of the model. As a result, in this part of the question, you are to perform five-fold cross validation to evaluate the accuracy of the classification results based on the outcome of k-means clustering. Perform each five - fold cross validation exercise as follows:

1. Merge the original training-set and testing set of the images in part (a) into one single data-set.
2. Randomly split the merged data-set into n equal-sized partitions of data points.
3. Choose one of the n partitions from Step 2 as the testing set while merging the remaining (n-1) partitions to become a new training data-set and then re-do part (a) and (b) to compute the corresponding classification accuracy. You **ONLY** use the best result of three times which you found in part (b), in determining the label of each cluster produced by k-means.
4. Repeat Step 3 for (n-1) times. Each time, you use a different partition (out of the n partitions) as a new testing set while merging the remaining (n-1) partitions as a new training set.
5. The submitted result should be in the same format as Table 4.

Submit both your code and results.

Table. 4. The accuracy of k-means clustering performance under 5-fold cross validation:

Testing set	Classification Accuracy (%)
Part 1	
Part 2	
……	
Part 5	
Average	

Q3 [20 marks + 30 bonus]: Bernoulli Mixture Models

A Bernoulli Mixture Model (BMM) is a probabilistic model that assumes data points are sampled from a mixture of multi-dimensional Bernoulli distributions. In what follows, we will derive the optimal parameters for BMM which would maximize its log-likelihood function:

Consider a set of binary variables x_i, where $i = 1,..., D$, each of which is governed by a Bernoulli distribution with parameter q_i, so that

$$p(x|q) \ = \prod_{i=1}^{D} q_i^{x_i}(1 - q_i)^{(1-x_i)}$$

where **x** = $(x_1, x_2,... x_D)^T$ and **q** =$(q_1,..., q_D)^T$. In other words, **x** follows a D-dimensional Bernoulli distribution where each variable x_i is independent of each other and Prob (x_i =1) = q_i which is the i-th element of **q**. Consider a mixture of K of such D-dimensional Bernoulli distributions with its density function given by:

$$p(x|q, \pi) \ = \sum_{k=1}^{K} \pi_k p(x|qk) \qquad\qquad (*)$$

where **q** = {**q**$_1$, ..., **q**$_K$} and $\pi \ = \ (\pi_1, \pi_2,..., \pi_K)$. Now, consider a data set **X** = {**x$_1$, ..., x$_N$**} which is generated by the Bernoulli Mixture model of (*).

Now, we show that the log-likelihood of p(**X**) is given by: $\sum_{n=1}^{N} log \sum_{k=1}^{K} \pi_k p(x_n|q_k)$. First note that

$$P(X) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k p(x_n \mid q_k)$$

By taking the logarithm of the above expression, we have:

$$\begin{aligned}\log P(X) &= \log \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k p(x_n \mid q_k) \\ &= \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p(x_n \mid q_k)\end{aligned}$$

Define a variable $\gamma(z_{nk}) \ = \frac{\pi_k p(x_n|q_k)}{(\sum_{j=1}^{K} \pi_j p(x_n|q_k))}$, which represents the "responsibility" of the $k - th$ cluster (i.e. the $k - th$ component of the Bernoulli mixture) for the data point (vector) **x$_n$**. Now prove that the best π_k after the first round of the EM-algorithm is

$\frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}$ and **q**

$\frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}$ and $\mathbf{q_k}$ = $\frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$. In order to maximize with respect to π_k , we need to introduce a Lagrange multiplier to ensure that $\sum_k \pi_k = 1$. As a result, we now maximize the following quantity:

$$\log P(X) + \lambda(\sum_{k=1}^{K} \pi_k - 1) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p(x_n \mid q_k) + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

Taking derivative the above expression with respect to π_k , we have

$$\begin{aligned}
\frac{\partial}{\partial \pi_k}\left(\log P(X) + \lambda(\sum_{k=1}^{K} \pi_k - 1)\right) &= \frac{\partial}{\partial \pi_k}\left(\sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p(x_n \mid q_k) + \lambda(\sum_{k=1}^{K} \pi_k - 1)\right) \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \pi_k} \log \sum_{k=1}^{K} \pi_k p(x_n \mid q_k) + \lambda \\
&= \sum_{n=1}^{N} \frac{p(x_n \mid q_k)}{\sum_{k=1}^{K} \pi_k p(x_n \mid q_k)} + \lambda \\
&= \sum_{n=1}^{N} \frac{\gamma(z_{nk})}{\pi_k} + \lambda
\end{aligned}$$

Set it to 0, multiply both side by π_k and then sum over *k,* we have

$$\begin{aligned}
\sum_{n=1}^{N} \frac{\gamma(z_{nk})}{\pi_k} + \lambda &= 0 \\
\sum_{n=1}^{N} \gamma(z_{nk}) + \lambda \pi_k &= 0 \\
\sum_{k=1}^{K}\left(\sum_{n=1}^{N} \gamma(z_{nk}) + \lambda \pi_k\right) &= 0 \\
N + \lambda &= 0 \\
\lambda &= -N
\end{aligned}$$

Substituting $\lambda = -N$,

$$\sum_{n=1}^{N}\frac{\gamma(z_{nk})}{\pi_k}+\lambda=0$$

$$\sum_{n=1}^{N}\frac{\gamma(z_{nk})}{\pi_k}-N=0$$

$$\pi_k=\frac{\sum_{n=1}^{N}\gamma(z_{nk})}{N}$$

Finding q_k **(Here we consider xn and qk as scalar, and the conclusion is also true is xn and qk are vectors)**

$$\begin{aligned}\frac{\partial}{\partial q_k}p(x_n\mid q_k)&=x_n q_k^{x_n-1}(1-q_k)^{1-x_n}-(1-x_n)q_k^{x_n}(1-q_k)^{-x_n}\\&=q_k^{x_n-1}(1-q_k)^{-x_n}\left(x_n(1-q_k)-(1-x_n)q_k\right)\\&=q_k^{x_n-1}(1-q_k)^{-x_n}(x_n-q_k)\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial q_k}\log P(X)&=\frac{\partial}{\partial q_k}\sum_{n=1}^{N}\log\sum_{k=1}^{K}\pi_k p(x_n\mid q_k)\\&=\sum_{n=1}^{N}\frac{\pi_k}{\sum_{k=1}^{K}\pi_k p(x_n\mid q_k)}\cdot\frac{\partial}{\partial q_k}p(x_n\mid q_k)\\&=\sum_{n=1}^{N}\frac{\pi_k}{\sum_{k=1}^{K}\pi_k p(x_n\mid q_k)}\left(q_k^{x_n-1}(1-q_k)^{-x_n}(x_n-q_k)\right)\\&=\sum_{n=1}^{N}\gamma(z_{nk})\cdot\frac{\left(q_k^{x_n-1}(1-q_k)^{-x_n}(x_n-q_k)\right)}{p(x_n\mid q_k)}\\&=\sum_{n=1}^{N}\frac{\gamma(z_{nk})(x_n-q_k)}{q_k(1-q_k)}\end{aligned}$$

Set $\frac{\partial}{\partial q_k}\log P(X)=0$,

$$\sum_{n=1}^{N}\frac{\gamma(z_{nk})(x_n-q_k)}{q_k(1-q_k)}=0$$

$$\sum_{n=1}^{N}\gamma(z_{nk})(x_n-q_k)=0$$

$$\sum_{n=1}^{N}\gamma(z_{nk})x_n-\sum_{n=1}^{N}\gamma(z_{nk})q_k=0$$

$$q_k=\frac{\sum_{n=1}^{N}\gamma(z_{nk})x_n}{\sum_{n=1}^{N}\gamma(z_{nk})}$$

(a) **[20 marks]** Provide the pseudo code (MapReduce is NOT required) to estimate the parameters of a BMM model based on maximum-likelihood arguments using the Expectation-Maximization algorithm. The pseudo code should include a detailed description on the list of input/ intermediate variables used and how each of them get updated during the E-step and M-step of each iteration.

The following Part (b) of Q3 is an OPTIONAL for IERG4300 and IEMS5709 but MANDATORY for ESTR4300:

(b) **[30 bonus marks]** Consider the images of the handwritten digits in the MNIST database described in Q2 of this homework. After binarization of the original grey-scale image to a bi-level black-and-white one, the color (black or white) of each pixel in an 28x28 image can then be considered as the outcome of a binary variable. As such, each image of a handwritten digit can be represented as a 784-dimension data point generated from a mixture of k multi-dimensional Bernoulli components where each component is a 784-dimension Bernoulli variable.

Using the BMM and the EM-algorithm as discussed in parts (a) of Q3 to perform clustering of the training set and use the result to classify the samples in the test-set of the MNIST database described in Q2(a) of this homework. Note that preprocessing is required to convert the grey-scale images to bi-level black and white ones. You may implement the BMM-EM algorithm EITHER in the form of a standalone, sequential program in the language of your choice, e.g. C, C++, Matlab, Java, etc OR under the MapReduce framework.

[Additional 10 extra marks will be given for a correct MapReduce implementation.]

Report the accuracy of your classification.

You need to submit BOTH your code and the classification performance results.