CS6220 Unsupervised Data Mining

HW2 KMEANS and Gaussian Mixtures

Make sure you check the syllabus for the due date. Please use the notations adopted in class, even if the problem is stated in the book using a different notation.

We are not looking for very long answers (if you find yourself writing more than one or two pages of typed text per problem, you are probably on the wrong track). Try to be concise; also keep in mind that good ideas and explanations matter more than exact details.

Submit all code files Dropbox (create folder HW1 or similar name). Results can be pdf or txt files, including plots/tabels if any.

"Paper" exercises: submit using Dropbox as pdf, either typed or scanned handwritten.

DATATSET **: SpamBase: emails (54-feature vectors) classified as spam/nospam**

DATATSET **: 20 NewsGroups : news articles**

DATATSET **: MNIST : 28x28 digit B/W images**

DATATSET **: FASHION : 28x28 B/W images**

https://en.wikipedia.org/wiki/MNIST_database
http://yann.lecun.com/exdb/mnist/
https://www.kaggle.com/zalando-research/fashionmnist

PROBLEM 1: KMeans Theory

Given Kmeans Objective discussed in class with Euclidian distance

$$\min \sum_i \sum_k \pi_{ik} \cdot ||X_i - \mu_k||^2$$

A) prove that E step update on membership (\pi) achieves the minimum objective given the current centroids(\mu)
B) prove that M step update on centroids (\mu) achievess the minimum objective given the current memberships(\pi)
C) Explain why KMeans has to stop (converge), but not necessarily to the global minimum objective value.

PROBLEM 2 : KMeans on data

Using Euclidian distance or dot product similarity (choose one per dataset, you can try other similarity metrics),
A) run KMeans on the MNIST Dataset, try K=10
B) run KMeans on the FASHION Dataset, try K=10
C) run KMeans on the 20NG Dataset, try K=20

For all three datasets, evaluate the KMeans objective for a higher K (for example double) or smaller K(for example half).
For all three datasets, evaluate external clustering performance using data labels and performance metrics Purity

and Gini Index (see [A] book section 6.9.2).

PROBLEM 3 : Gaussian Mixture on toy data

You are required to implemet the main EM loop, but can use math API/functions provided by your language to calculate normal densities, covariance matrix, etc.
A) The gaussian 2-dim data on file 2gaussian.txt has been generated using a mixture of two Gaussians, each 2-dim, with the parameters below. Run the EM algorithm with random initial values to recover the parameters.

```
mean_1 [3,3]); cov_1 = [[1,0],[0,3]]; n1=2000 points
mean_2 =[7,4]; cov_2 = [[1,0.5],[0.5,1]]; ; n2=4000 points
```

You should obtain a result visually like this (you dont necessarily have to plot it)

B) Same problem for 2-dim data on file 3gaussian.txt , generated using a mixture of three Gaussians. Verify your findings against the true parameters used generate the data below.

```
mean_1 = [3,3] ; cov_1 = [[1,0],[0,3]]; n1=2000
mean_2 = [7,4] ; cov_2 = [[1,0.5],[0.5,1]] ; n2=3000
mean_3 = [5,7] ; cov_3 = [[1,0.2],[0.2,1]]    ); n3=5000
```

Additional notes helpful for implementing Gaussian Mixtures:
https://xcorr.net/2008/06/11/log-determinant-of-positive-definite-matrices-in-matlab/
http://andrewgelman.com/2016/06/11/log-sum-of-exponentials/
https://hips.seas.harvard.edu/blog/2013/01/09/computing-log-sum-exp/

PROBLEM 4 : Gaussian Mixture on real data

Run EM to obtain a Gaussian Mixture on FASHION dataset (probably won't work) and on SPAMBASE dataset (should work). Use a library/package (such as scikit-learn) and at first use the option that imposes a diagonal covariance matrix.
Sampling data might be necessary to complete the run.