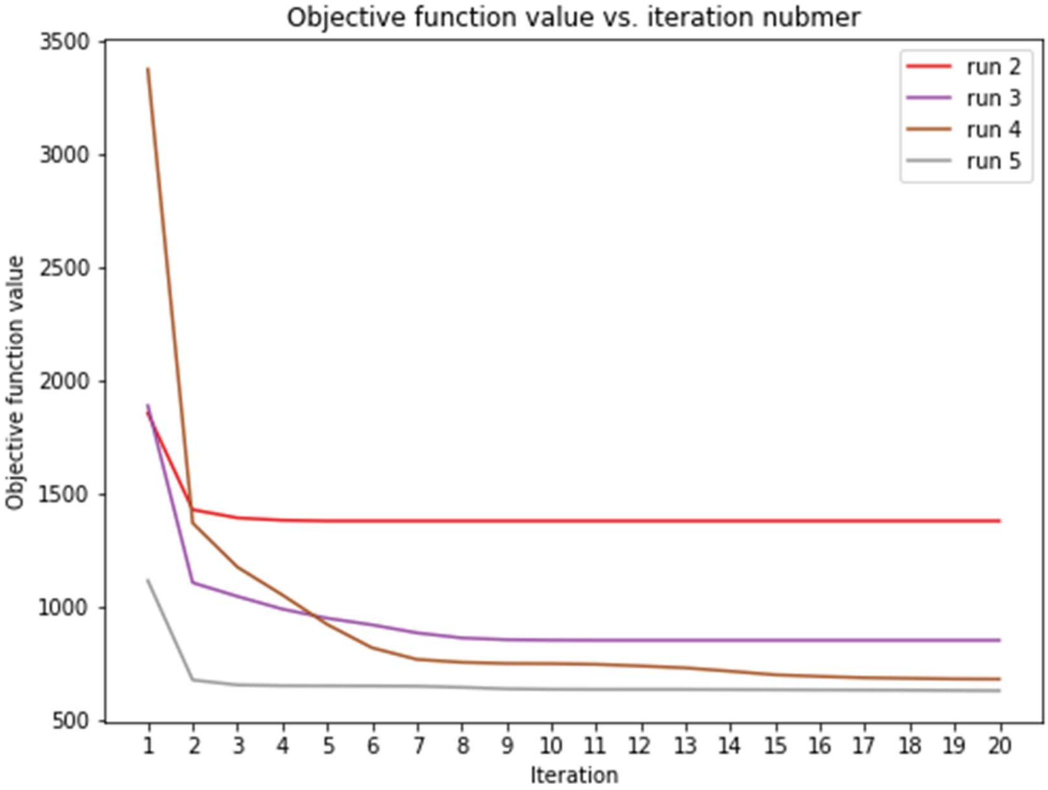
COMS4721 Machine Learning for Data Science Homework 3

Po-Chieh Liu

UNI: pl2441

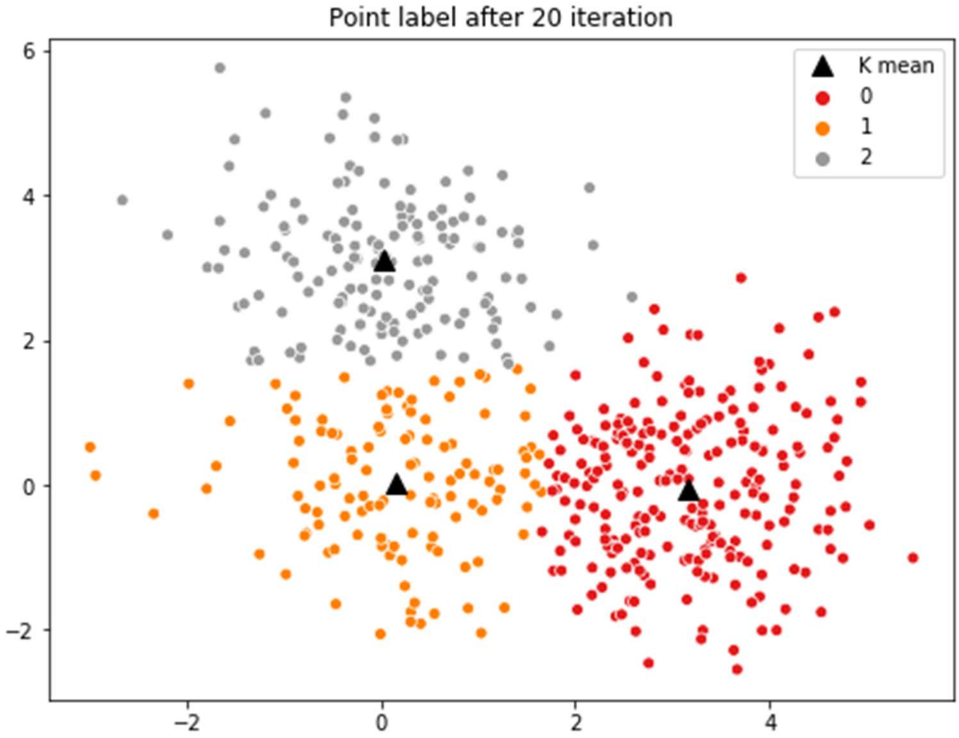
Problem 1

(a) For K = 2; 3; 4; 5, plot the value of the K-means objective function per iteration for 20 iterations (the algorithm may converge before that).

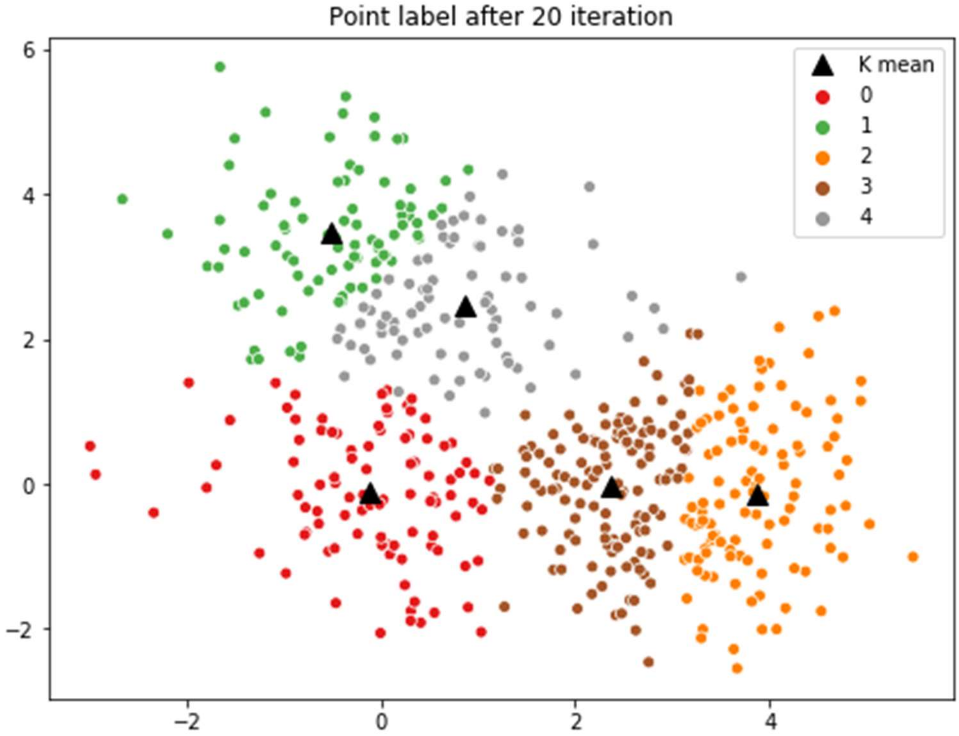


(b) For K = 3; 5, plot the 500 data points and indicate the cluster of each for the final iteration by marking it with a color or a symbol.

K = 3

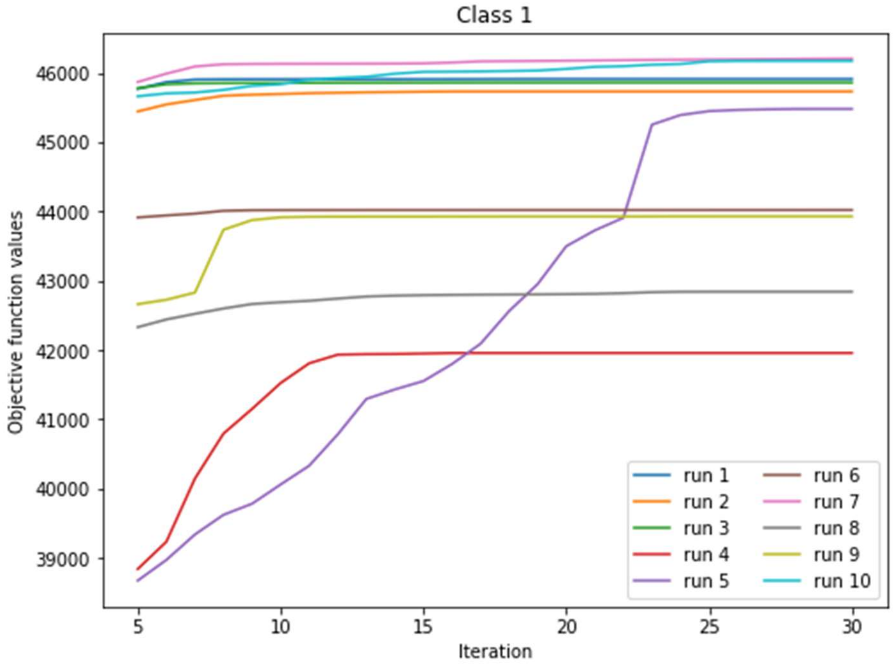


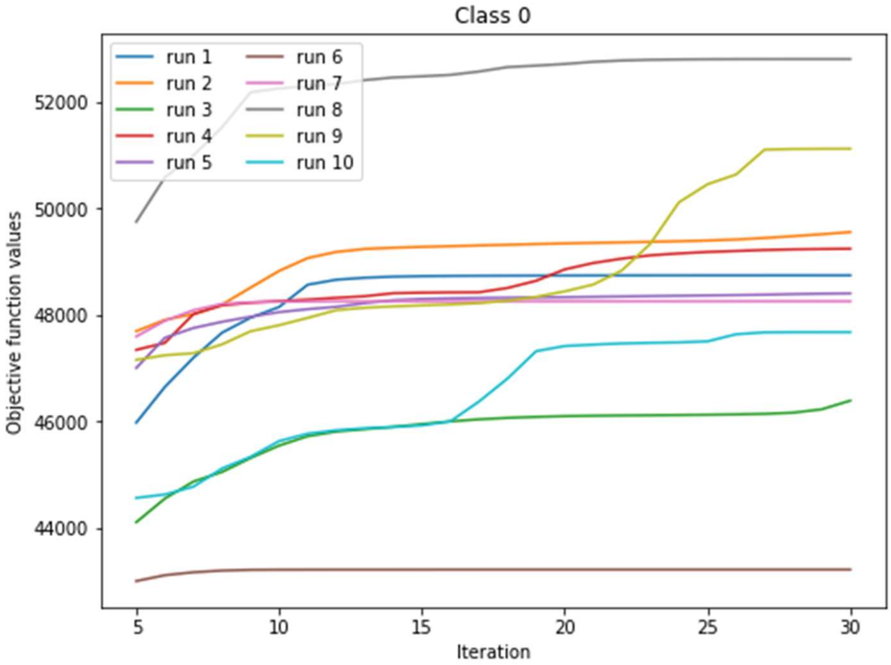
K=5



Problem 2

(a) Implement the EM algorithm for the GMM described in class. Using the training data provided, for each class separately, plot the log marginal objective function for a 3-Gaussian mixture model over 10 different runs and for iterations 5 to 30. There should be two plots, each with 10 curves.





(b) Using the best run for each class after 30 iterations, predict the testing data using a Bayes classifier and show the result in a 2 2 confusion matrix, along with the accuracy percentage.

Confusion Matrix for using prior from part (a):

Confusion matrix	Labeled 0	Labeled 1
Predicted 0	208	10
Predicted 1	70	172

Accuracy: 82.61%

Repeat this process for a 1-, 2-, 3- and 4-Gaussian mixture model. Show all results nearby each other, and don't repeat Part (a) for these other cases. Note that a 1-Gaussian GMM doesn't require an algorithm, although your implementation will likely still work in this case.

Confusion matrix for 1-, 2-, 3-, 4-Gaussian prior:

1-Gaussian with Accuracy 77.39

Confusion matrix	Labeled 0	Labeled 1
Predicted 0	180	6
Predicted 1	98	176

2-Gaussian with Accuracy 79.13

Confusion matrix	Labeled 0	Labeled 1
Predicted 0	191	9
Predicted 1	87	173

3-Gaussian with Accuracy 81.96

Confusion matrix	Labeled 0	Labeled 1
Predicted 0	202	7
Predicted 1	76	175

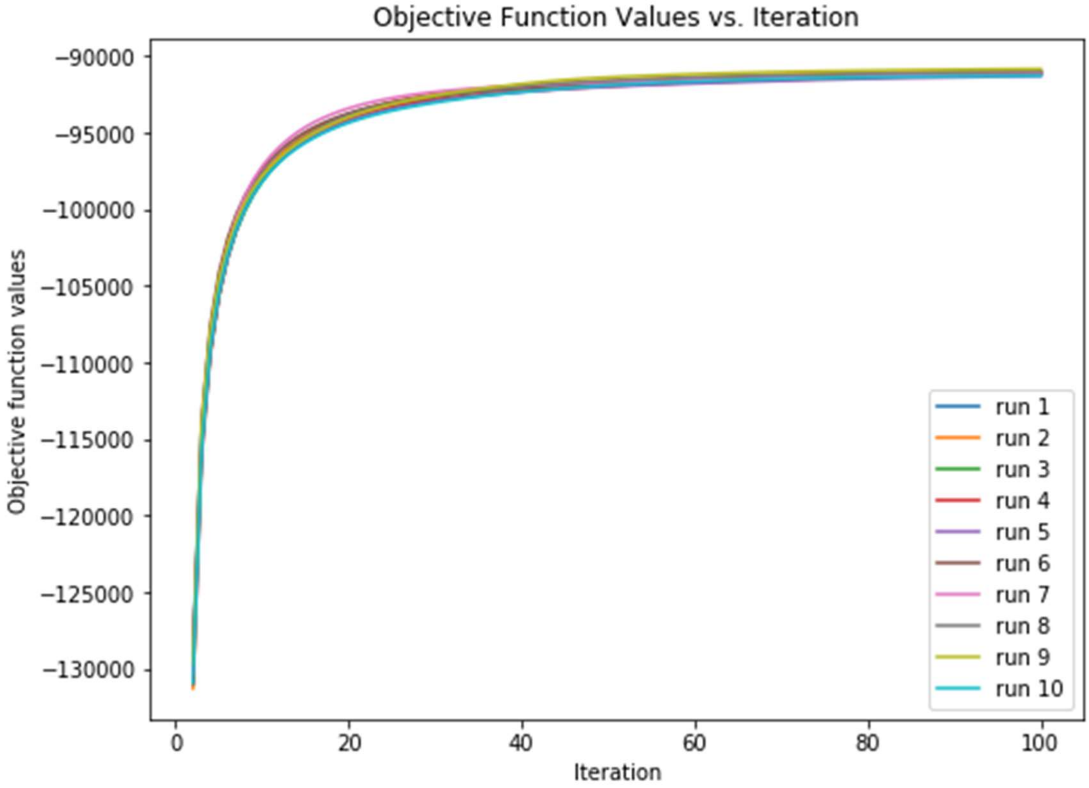
4-Gaussian with Accuracy 80.22

Confusion matrix	Labeled 0	Labeled 1
Predicted 0	192	5
Predicted 1	86	177

Problem 3

(a) On a single plot, show the log joint likelihood for iterations 2 to 100 for each run. In a table, show in each row the final value of the training objective function next to the RMSE on the testing set. Sort these rows according to decreasing value of the objective function.

PLOT



TABLE

run #	Obj Value	RMSE	Rank
9	-90823.8473	0.673369	1
8	-90976.9997	0.673904	2
2	-90977.8629	0.673725	3
3	-90978.2374	0.674018	4
1	-91084.3616	0.674189	5
4	-91141.5272	0.674347	6
6	-91168.0194	0.674507	7
7	-91173.0251	0.674398	8
10	-91261.4854	0.674695	9
5	-91288.9654	0.674875	10

(b) For the run with the highest objective value, pick the movies “Star Wars” “My Fair Lady” and “Goodfellas” and for each movie find the 10 closest movies according to Euclidean distance using their respective locations vj . List the query movie, the ten nearest movies and their distances. A mapping from index to movie is provided with the data.

Star Wars (1977)

MOVIE ID	DISTANCE	NAME
171	0.099517	Empire Strikes Back, The (1980)
173	0.254445	Raiders of the Lost Ark (1981)
180	0.405606	Return of the Jedi (1983)
172	0.618292	Princess Bride, The (1987)
428	0.704261	Day the Earth Stood Still, The (1951)
612	0.779787	My Man Godfrey (1936)
209	0.786532	Indiana Jones and the Last Crusade (1989)
193	0.809872	Sting, The (1973)
1006	0.853839	Waiting for Guffman (1996)
152	0.857303	Fish Called Wanda, A (1988

My Fair Lady

MOVIE ID	DISTANCE	NAME
418	0.494611	Mary Poppins (1964)
98	0.792115	Snow White and the Seven Dwarfs (1937)
417	0.868884	Cinderella (1950)
419	0.898732	Alice in Wonderland (1951)
142	0.899821	Sound of Music, The (1965)
968	0.904088	Winnie the Pooh and the Blustery Day (1968)
1146	0.961298	My Family (1995)
431	1.008023	Fantasia (1940)
601	1.03174	American in Paris, An (1951)
150	1.144325	Willy Wonka and the Chocolate Factory (1971)

GoodFellas

MOVIE ID	DISTANCE	NAME
692	0.317113	Casino (1995)
187	0.414579	Full Metal Jacket (1987)
176	0.514571	Good, The Bad and The Ugly, The (1966)
55	0.796721	Pulp Fiction (1994)
645	0.809231	Once Upon a Time in the West (1969)
182	0.810928	Alien (1979)
522	0.913397	Cool Hand Luke (1967)
503	1.031519	Bonnie and Clyde (1967)
184	1.032881	Psycho (1960)
186	1.03688	Godfather: Part II, The (1974)