HW03

Yawen Han (yh3069)

Apr 19, 2019

Problem 01: K-Means

Generate Data Points

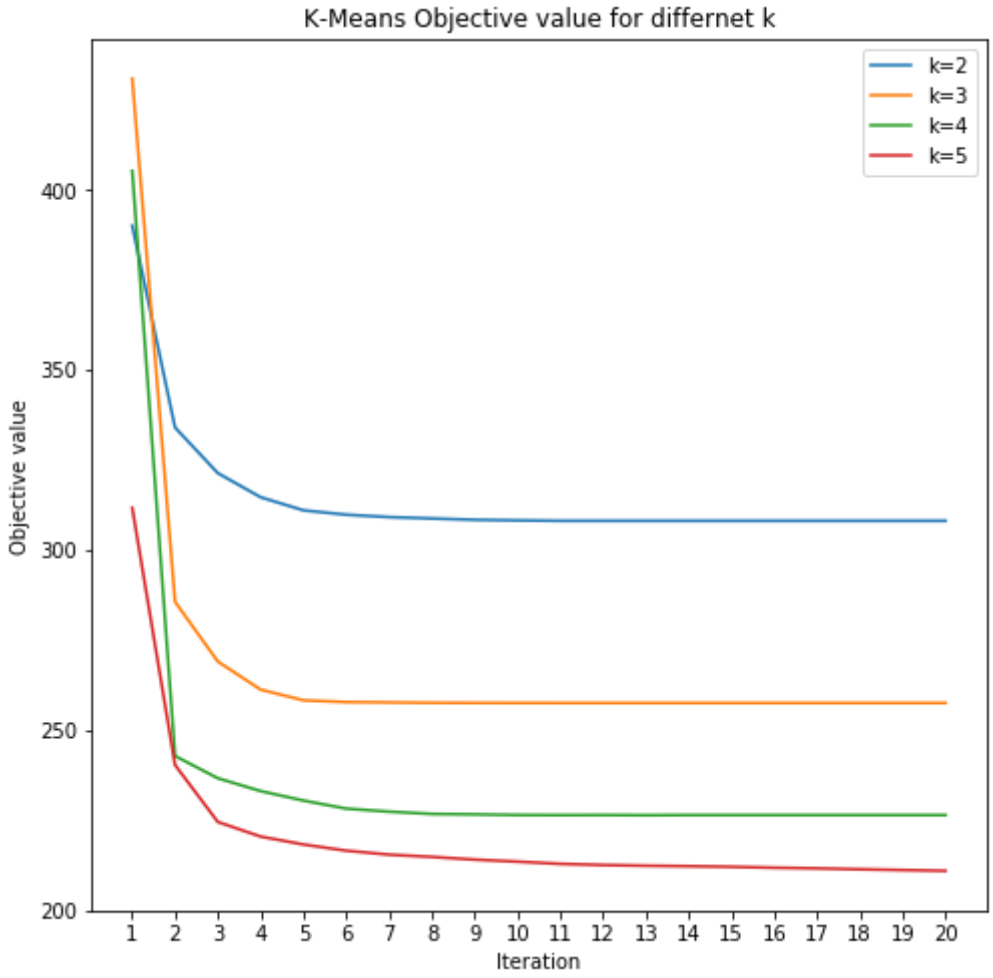
First, generate 500 observations from a mixture of three Gaussians on R2 with mixing weights $\pi = [0.2, 0.5, 0.3]$ and means μ and covariances Σ.

Implement K-Means Algorithm

Then implement K-Means algorithm, with **"num_centroids"** for number of clusters and **"iterations"** for terminal criteria.

Question 1-a: Objective vs. Iteration for k=2,3,4,5

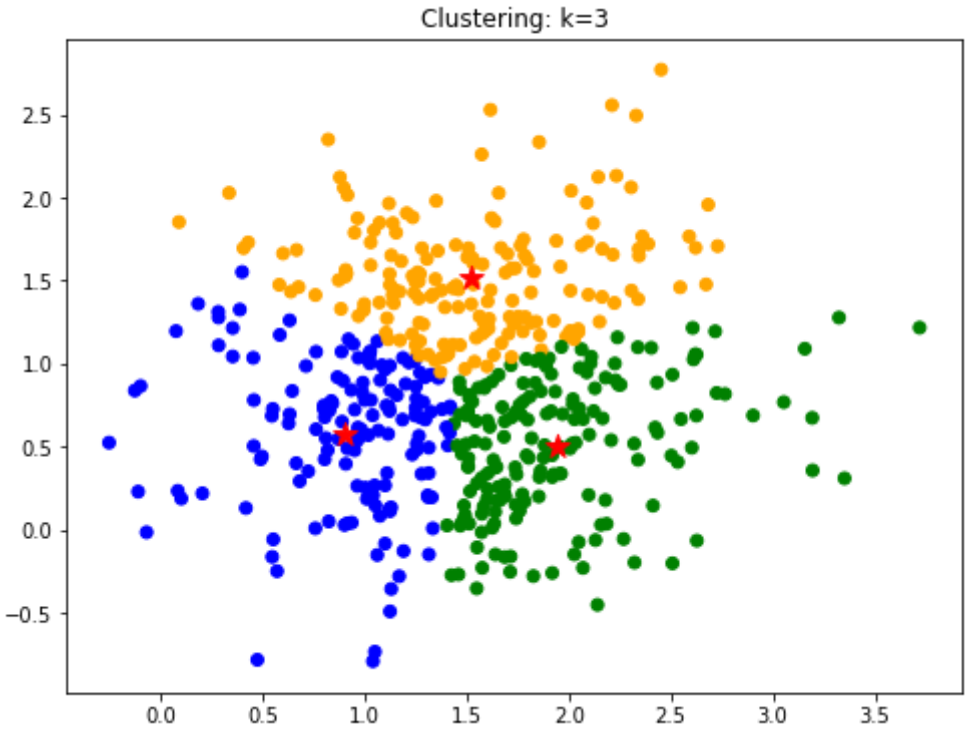
For K = 2, 3, 4, 5, plot the value of the K-means objective function per iteration for 20 iterations



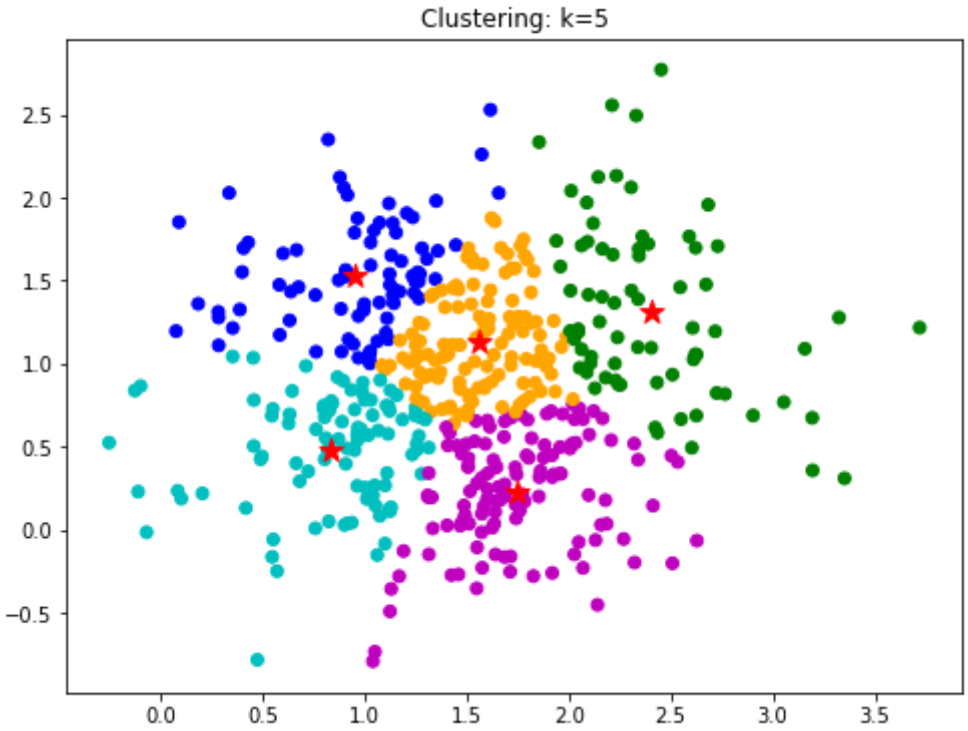
According to the plot above, we can conclude that with the increasing number of k(number of clusters), the objective function get a samller value. This makes sense as the more clusters, the relative distance between each data point and corresponding centroid is smaller.

Question 1-b: Cluster plots for k=3,5

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.



For **k=3**, the data points assignments and centroids are shown above.



For **k=5**, the data points assignments and centroids are shown above. Compared to k=3, some clusters are more sparse than others.

Problem 02: Bayes Classifier Revisited

In this section, the **EM** algorithm for the **Gaussian mixture** model is implemented, with the purpose of using it in a **Bayes classifier**.

EM Algirithm

Implement the EM algorithm by iterating E-step and M-step.

1. LogLikelihood

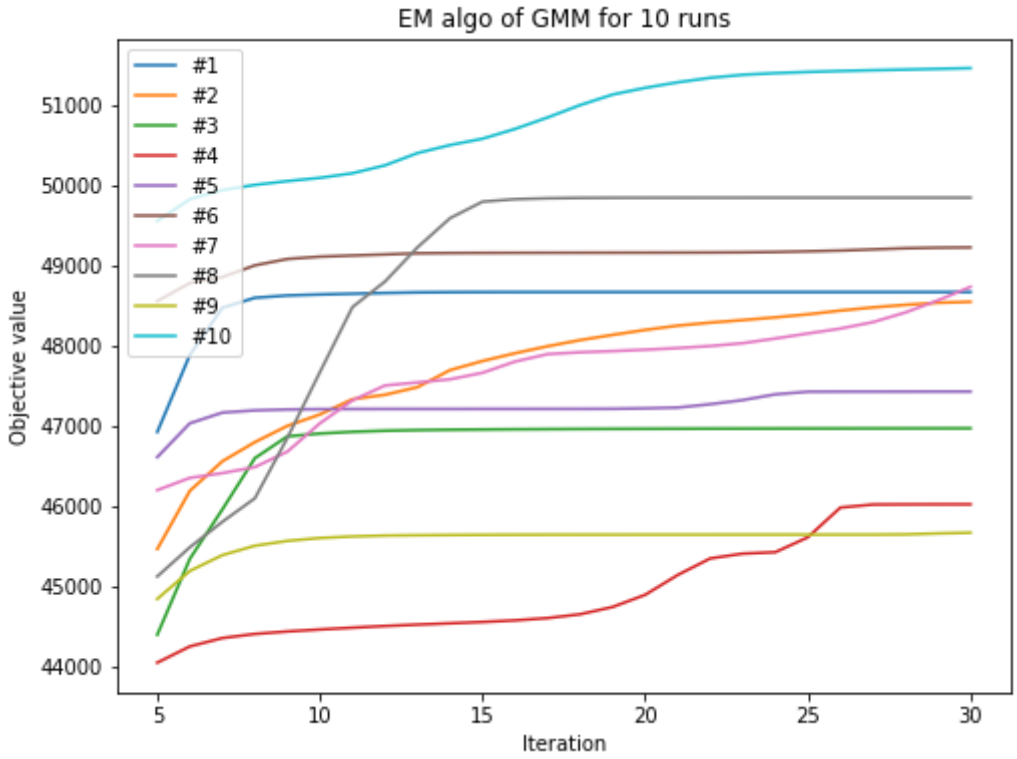
2. E-Step

3.M-Step

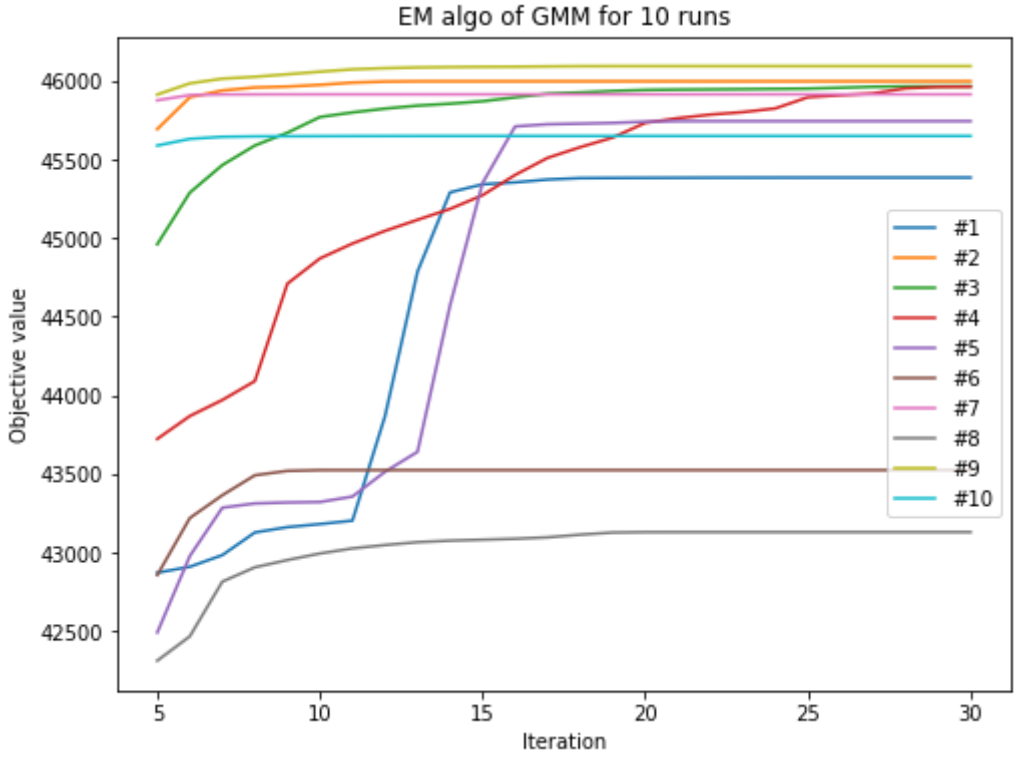
3. EM algo

Question 2-a: GMM with EM

plot the log marginal objective function for a 3-Gaussian mixture model over 10 different runs and for iterations 5 to 30.



The plot above for **class = 0**, demonstrating the log marginal objective function for a 3-Gaussian mixture model over 10 different runs and for iterations 5 to 30.



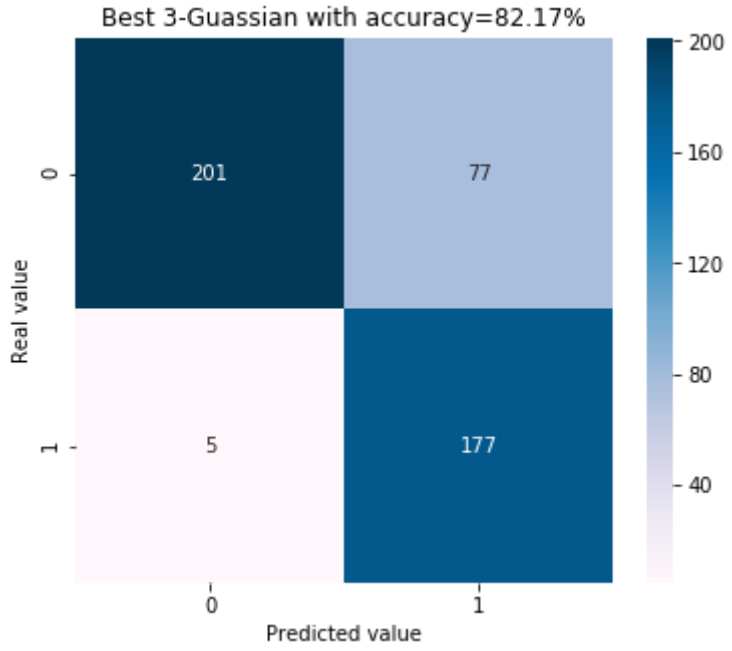
The plot above for **class = 1**, demonstrating the log marginal objective function for a 3-Gaussian mixture model over 10 different runs and for iterations 5 to 30.

Question 2-b: Bayes Classifier

1. The best run prediction for (a)

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

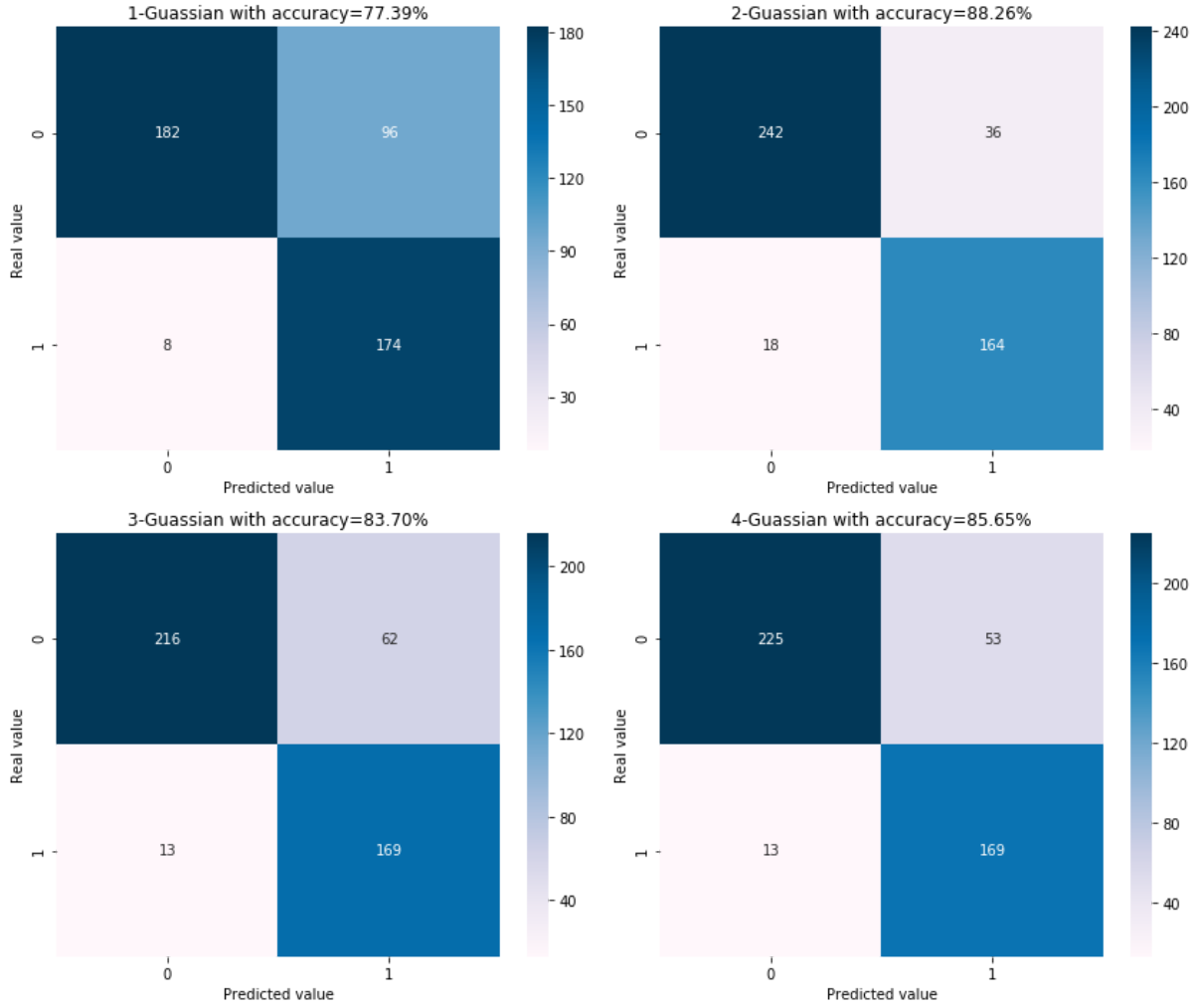
```
Text(0.5, 1.0, 'Best 3-Guassian with accuracy=82.17%')
```



The confusion matrix for the best run in (a) is shown above, the accuracy is **82.17%**.

2. Repeat the process for 1-4 Guassian Mixture Model

Repeat this process for a 1-, 2-, 3- and 4-Gaussian mixture model.



The confusion matrixes for for a 1-, 2-, 3- and 4-Gaussian mixture model are shown above, with accuarcy shown in the title.

#-Guassian Mixture	Accuracy
1	77.39%
2	88.26%
3	83.7%
4	85.65%

The accuarcy for each Gussian mixture model is also summarized in the table above.

Problem 03: Matrix Factorization

In this section, the MAP inference algorithm for the matrix completion problem is implemented.

Before the implemention, let's have a look at the dataset:

	user_id	movie_id	rating
0	196	242	-0.53039
1	186	302	-0.53039
2	244	51	-1.53040
3	166	346	-2.53040
4	298	474	0.46961
5	115	265	-1.53040
6	253	465	1.46960
7	305	451	-0.53039
8	6	86	-0.53039
9	62	257	-1.53040

The dataset is summarized to exclude the missing values. In order to implement the algorithm, it will be transformed to the sparse matrix first.

```
There are 943 unique users
There are 1682 unique movies
```

```
percentage of user-movies that have a rating: 5.99%
```

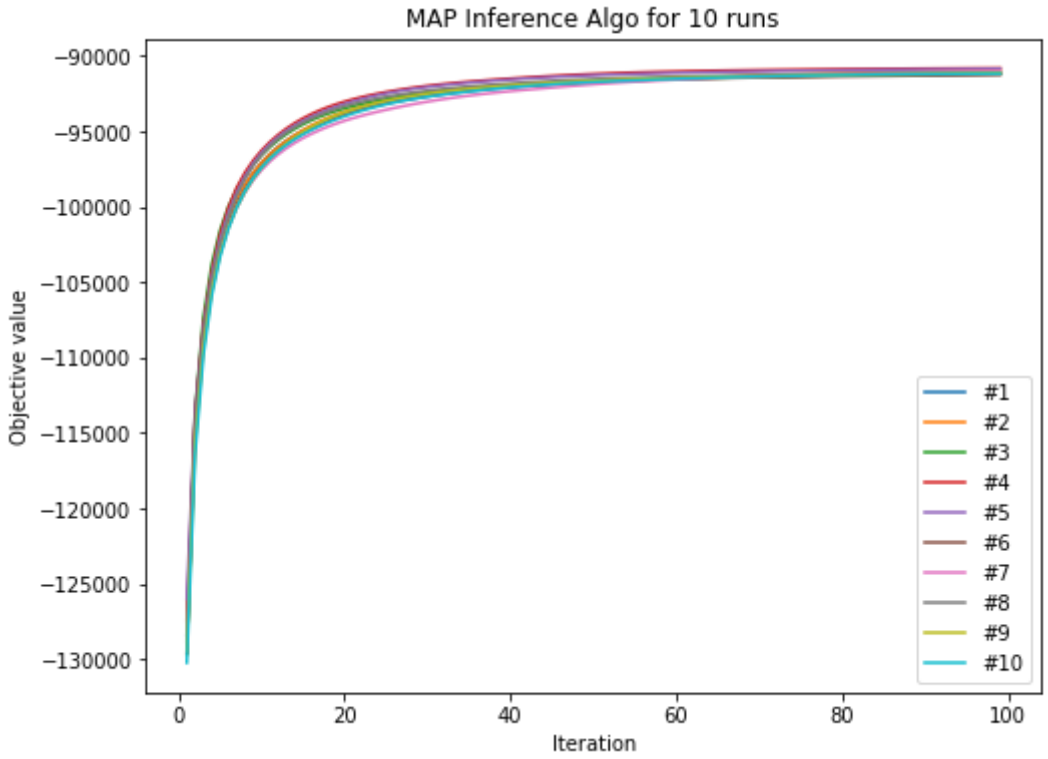
Matrix Factorization

Implement the algorithm by optimizing the MAP, and update the u_i and v_j locations iteratively.

```
Shape of P:  (10, 943)
Shape of Q:  (10, 1682)
```

Question 3-a: MAP Inference Algo for Matrix Factorization

Run your code 10 times. For each run, initialize your ui and vj vectors as N (0, I) random vectors.



On the plot above, it shows the the log joint likelihood for iterations 2 to 100 for each run. The objective function value is improved after 100 iterations, and reach the stable stage.

	#run	Objective_val	RMSE
3	4	-90803.097293	1.104995
4	5	-90839.031060	1.101811
1	2	-90963.316661	1.123528
6	7	-91009.267676	1.104033
7	8	-91107.594851	1.107406
8	9	-91123.845278	1.113834
9	10	-91161.350968	1.167587
2	3	-91180.545171	1.123942
0	1	-91186.572792	1.091722
5	6	-91243.471745	1.094620

In the table above, it 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.

Question 3-b: Find Closest Movies

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

`Top10 closest movies for Star Wars :`

	Name	dist
171	Empire Strikes Back	0.325537
173	Raiders of the Lost Ark (1981)	0.602483
180	Return of the Jedi (1983)	0.705874
209	Indiana Jones and the Last Crusade (1989)	0.912302
172	Princess Bride	0.914045
194	Terminator	0.959145
95	Terminator 2: Judgment Day (1991)	0.963688
0	Toy Story (1995)	1.047298
519	Great Escape	1.052359
11	Usual Suspects	1.060604

`Top10 closest movies for My Fair Lady :`

	Name	dist
142	Sound of Music	0.495255
418	Mary Poppins (1964)	0.643078
416	Parent Trap	0.853166
419	Alice in Wonderland (1951)	0.896519
98	Snow White and the Seven Dwarfs (1937)	0.901292
131	Wizard of Oz	0.907173
587	Beauty and the Beast (1991)	0.907759
450	Grease (1978)	0.924008
417	Cinderella (1950)	0.946477
713	Carrington (1995)	0.968409

```
Top10 closest movies for GoodFellas :
```

	Name	dist
692	Casino (1995)	0.800468
187	Full Metal Jacket (1987)	0.829972
503	Bonnie and Clyde (1967)	0.836691
134	2001: A Space Odyssey (1968)	0.867620
179	Apocalypse Now (1979)	0.870649
176	Good	0.882256
645	Once Upon a Time in the West (1969)	0.932872
186	Godfather: Part II	0.935551
468	Short Cuts (1993)	0.970953
788	Swimming with Sharks (1995)	1.061499

Toggle code