ELEN 4903: Machine Learning

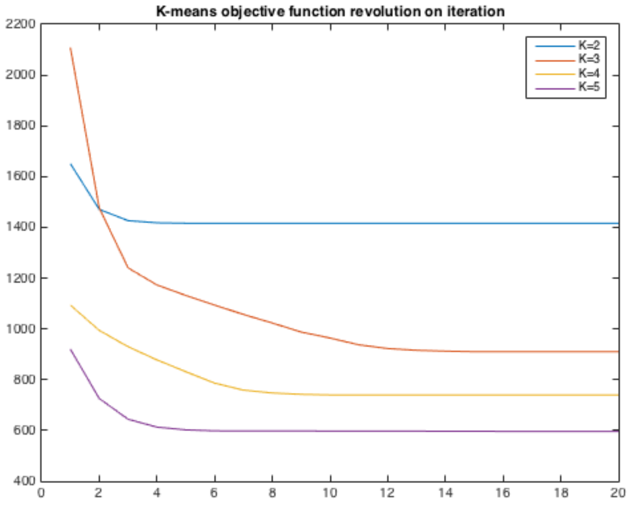
Columbia University, Spring 2016

Homework 4: Due April 15, 2016 by 11:59pm

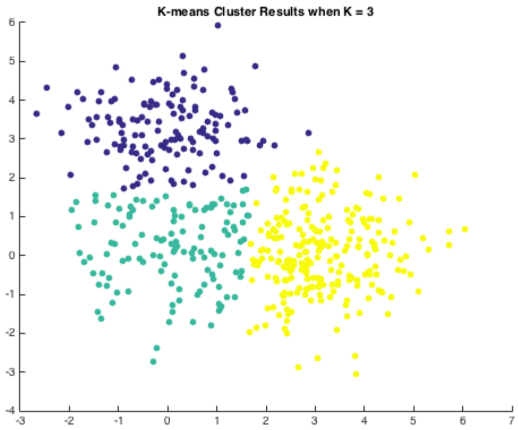
Yuntong Wang
yw2768

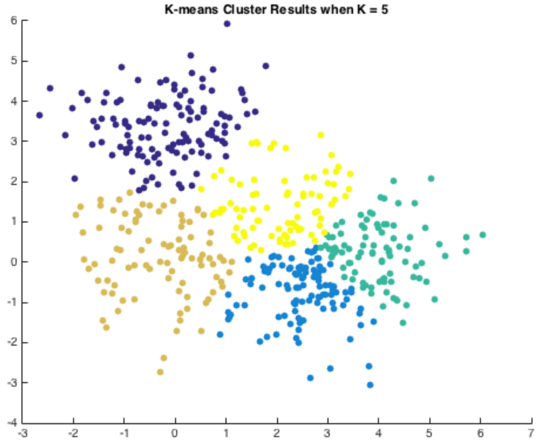
Problem 1.

1. 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).



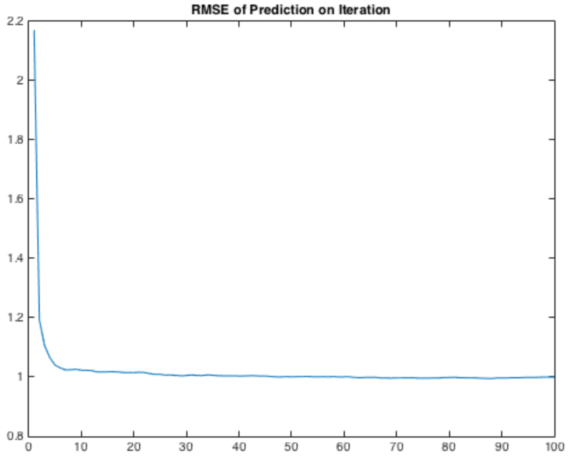
2. For K = 3, 5, plot the 500 data points and indicate the cluster of each for the final iteration by marking it with a color or a symbol.



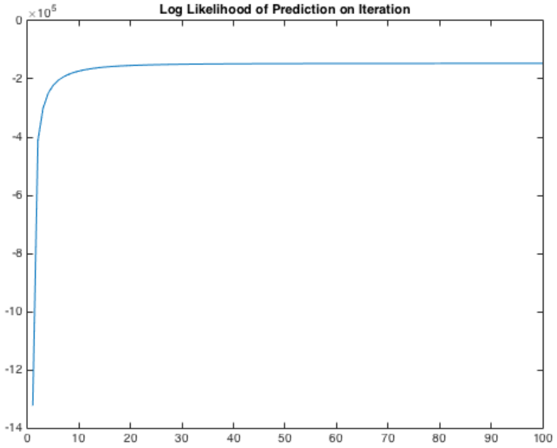


Problem 2.

1. Plot the RMSE of your predictions on the held out test set provided for iteration number 2 to 100.



2. On a separate plot, show the log joint likelihood for iterations 2 to 100.



3. After 100 iterations, pick three reasonably well-known movies from the list provided and for each movie find the 5 closest movies according to Euclidean distance using their respective locations v_j . List the query movie, the five nearest movies and their distances. A mapping from index to movie is provided with the data.

```
Query movie: Pulp Fiction (1994)
No. 1 closest movie: Deer Hunter, The (1978), distance: 0.791783
No. 2 closest movie: Raising Arizona (1987), distance: 0.826980
No. 3 closest movie: Jackie Brown (1997), distance: 0.907795
No. 4 closest movie: Trainspotting (1996), distance: 0.918652
No. 5 closest movie: Swingers (1996), distance: 0.921128

Query movie: Forrest Gump (1994)
No. 1 closest movie: Dances with Wolves (1990), distance: 0.537926
No. 2 closest movie: Field of Dreams (1989), distance: 0.658093
No. 3 closest movie: Client, The (1994), distance: 0.709732
No. 4 closest movie: Glory (1989), distance: 0.748582
No. 5 closest movie: Stand by Me (1986), distance: 0.759418

Query movie: Titanic (1997)
No. 1 closest movie: Client, The (1994), distance: 0.867284
No. 2 closest movie: Fugitive, The (1993), distance: 0.869061
No. 3 closest movie: Great Expectations (1998), distance: 0.888748
No. 4 closest movie: Jurassic Park (1993), distance: 0.894321
No. 5 closest movie: Apollo 13 (1995), distance: 0.914937
```

4. After 100 iterations, perform K-means on the vectors $u_1,\ldots,u_{N_1}$ learned by your algorithm. Set K = 20, which is an arbitrary number. The centroids can be interpreted as personality types (as far as movies are concerned).

- Pick the 5 centroids corresponding to the 5 clusters that have the most data. For each cluster selected, give the number of users allocated to that cluster.
- For each of these 5 centroids, list the 10 movies with the largest (most positive) dot product with that centroid. Also give the value of the dot product next to the movie.

```
No. 1 popular centroid:    0.5746    0.4743   -0.5477    0.9900   -0.5629   -0.0244   -0.0163    0.6061
1.1627    0.5796

The number of users allocated to the centroid: 82
No. 1 popular movie of this centroid: Godfather, The (1972), distance: 4.885317
No. 2 popular movie of this centroid: Schindler's List (1993), distance: 4.632431
No. 3 popular movie of this centroid: Casablanca (1942), distance: 4.572252
No. 4 popular movie of this centroid: Lawrence of Arabia (1962), distance: 4.544643
No. 5 popular movie of this centroid: One Flew Over the Cuckoo's Nest (1975), distance: 4.530663
No. 6 popular movie of this centroid: Godfather: Part II, The (1974), distance: 4.504323
No. 7 popular movie of this centroid: Citizen Kane (1941), distance: 4.474424
No. 8 popular movie of this centroid: Shawshank Redemption, The (1994), distance: 4.453958
No. 9 popular movie of this centroid: Pulp Fiction (1994), distance: 4.443999
No. 10 popular movie of this centroid: Silence of the Lambs, The (1991), distance: 4.440085

No. 2 popular centroid:    0.5595    0.7722   -0.1509    0.8152   -0.8636   -0.6476    0.2387    0.2116
0.7342    0.4623

The number of users allocated to the centroid: 80
No. 1 popular movie of this centroid: L.A. Confidential (1997), distance: 4.448202
No. 2 popular movie of this centroid: Star Wars (1977), distance: 4.427577
No. 3 popular movie of this centroid: Good Will Hunting (1997), distance: 4.374810
No. 4 popular movie of this centroid: As Good As It Gets (1997), distance: 4.351784
No. 5 popular movie of this centroid: Shawshank Redemption, The (1994), distance: 4.302475
No. 6 popular movie of this centroid: Titanic (1997), distance: 4.273922
No. 7 popular movie of this centroid: Apt Pupil (1998), distance: 4.238456
No. 8 popular movie of this centroid: Wag the Dog (1997), distance: 4.224733
No. 9 popular movie of this centroid: Fargo (1996), distance: 4.210373
No. 10 popular movie of this centroid: Usual Suspects, The (1995), distance: 4.206115

No. 3 popular centroid:    0.4313    0.2342   -0.0764    0.9156   -0.6074   -0.6554   -0.0091   -0.0536
1.1474    0.7807

The number of users allocated to the centroid: 68
No. 1 popular movie of this centroid: Good Will Hunting (1997), distance: 4.642554
No. 2 popular movie of this centroid: As Good As It Gets (1997), distance: 4.557078
No. 3 popular movie of this centroid: Schindler's List (1993), distance: 4.499874
No. 4 popular movie of this centroid: Apt Pupil (1998), distance: 4.461279
No. 5 popular movie of this centroid: Dead Man Walking (1995), distance: 4.447042
No. 6 popular movie of this centroid: Shawshank Redemption, The (1994), distance: 4.369652
No. 7 popular movie of this centroid: English Patient, The (1996), distance: 4.354396
No. 8 popular movie of this centroid: Amistad (1997), distance: 4.341302
No. 9 popular movie of this centroid: Full Monty, The (1997), distance: 4.303735
No. 10 popular movie of this centroid: Fargo (1996), distance: 4.262216

No. 4 popular centroid:    0.5158    1.0516   -0.0904    1.0969   -0.5621   -0.3835   -0.1553    0.4263
1.1462    0.4184
```

```
The number of users allocated to the centroid: 67
No. 1 popular movie of this centroid: Wrong Trousers, The (1993), distance: 4.641704
No. 2 popular movie of this centroid: Godfather, The (1972), distance: 4.634404
No. 3 popular movie of this centroid: Pulp Fiction (1994), distance: 4.634344
No. 4 popular movie of this centroid: Shawshank Redemption, The (1994), distance: 4.605539
No. 5 popular movie of this centroid: Godfather: Part II, The (1974), distance: 4.603703
No. 6 popular movie of this centroid: Chasing Amy (1997), distance: 4.586262
No. 7 popular movie of this centroid: Wallace & Gromit: The Best of Aardman Animation (1996),
distance: 4.572984
No. 8 popular movie of this centroid: Citizen Kane (1941), distance: 4.554002
No. 9 popular movie of this centroid: Schindler's List (1993), distance: 4.548498
No. 10 popular movie of this centroid: Apt Pupil (1998), distance: 4.534231

No. 5 popular centroid:    0.8662    0.8539   -0.1447    0.9822   -0.7224    0.0542   -0.2617   -0.0805
0.8966    0.6004

The number of users allocated to the centroid: 65
No. 1 popular movie of this centroid: Star Wars (1977), distance: 4.571571
No. 2 popular movie of this centroid: Close Shave, A (1995), distance: 4.571103
No. 3 popular movie of this centroid: Wrong Trousers, The (1993), distance: 4.535521
No. 4 popular movie of this centroid: Wallace & Gromit: The Best of Aardman Animation (1996),
distance: 4.505720
No. 5 popular movie of this centroid: Casablanca (1942), distance: 4.463125
No. 6 popular movie of this centroid: Beautiful Thing (1996), distance: 4.397647
No. 7 popular movie of this centroid: Cold Comfort Farm (1995), distance: 4.380283
No. 8 popular movie of this centroid: Usual Suspects, The (1995), distance: 4.373759
No. 9 popular movie of this centroid: Princess Bride, The (1987), distance: 4.358973
No. 10 popular movie of this centroid: 12 Angry Men (1957), distance: 4.354052
```

5. After 100 iterations, perform K-means on the vectors $v_1,\ldots,v_{N_2}$ learned by your algorithm. Set K = 20, which is an arbitrary number.

- Pick the 5 centroids corresponding to the 5 clusters that have the most data. For each cluster selected, give the number of movies allocated to that cluster.
- For each of these 5 centroids, list the 10 movies with the smallest Euclidean distance to that centroid. Also give the value of the Euclidean distance next to the movie.

```
No. 1 popular centroid:    0.7427    0.6447    0.0555    0.9406   -0.8937   -0.3586    0.1290    0.3633
0.7193    0.4757

The number of movies allocated to the centroid: 72
```

```
No. 1 closest movie: Kiss the Girls (1997), distance: 0.688779
No. 2 closest movie: Murder in the First (1995), distance: 0.720754
No. 3 closest movie: Truth About Cats & Dogs, The (1996), distance: 0.728488
No. 4 closest movie: Now and Then (1995), distance: 0.733360
No. 5 closest movie: Apollo 13 (1995), distance: 0.736436
No. 6 closest movie: Mrs. Doubtfire (1993), distance: 0.744605
No. 7 closest movie: Client, The (1994), distance: 0.749637
No. 8 closest movie: MatchMaker, The (1997), distance: 0.765032
No. 9 closest movie: Rainmaker, The (1997), distance: 0.786197

No. 2 popular centroid:    0.2605    0.9607   -0.2746    1.0030   -0.6895   -0.3111   -0.2235    0.4509
1.1258    0.4874

The number of movies allocated to the centroid: 69
No. 1 closest movie: Smoke (1995), distance: 0.680785
No. 2 closest movie: Pink Floyd - The Wall (1982), distance: 0.737760
No. 3 closest movie: Trust (1990), distance: 0.740496
No. 4 closest movie: Horseman on the Roof, The (Hussard sur le toit, Le) (1995), distance: 0.789407
No. 5 closest movie: Night on Earth (1991), distance: 0.799507
No. 6 closest movie: Three Colors: White (1994), distance: 0.800610
No. 7 closest movie: Go Fish (1994), distance: 0.834110
No. 8 closest movie: Fear of a Black Hat (1993), distance: 0.839640
No. 9 closest movie: Nobody's Fool (1994), distance: 0.841593

No. 3 popular centroid:    0.7887    0.8500   -0.5039    0.7010   -0.6417   -0.2164   -0.3670   -0.1270
0.8966    0.7125

The number of movies allocated to the centroid: 69
No. 1 closest movie: Muriel's Wedding (1994), distance: 0.809618
No. 2 closest movie: Dead Man Walking (1995), distance: 0.828093
No. 3 closest movie: Ridicule (1996), distance: 0.841611
No. 4 closest movie: Big Night (1996), distance: 0.849971
No. 5 closest movie: Bottle Rocket (1996), distance: 0.927907
No. 6 closest movie: Remains of the Day, The (1993), distance: 0.929740
No. 7 closest movie: Cinema Paradiso (1988), distance: 0.930294
No. 8 closest movie: In the Bleak Midwinter (1995), distance: 0.932149
No. 9 closest movie: Kolya (1996), distance: 0.934061

No. 4 popular centroid:    0.3056    0.1305   -0.4629    0.9211   -0.7288   -0.5465    0.1047    0.0528
1.2089    0.5377
```

```
The number of movies allocated to the centroid: 65
No. 1 closest movie: Horseman on the Roof, The (Hussard sur le toit, Le) (1995), distance: 0.747991
No. 2 closest movie: Postino, Il (1994), distance: 0.828093
No. 3 closest movie: Six Degrees of Separation (1993), distance: 0.839267
No. 4 closest movie: Shine (1996), distance: 0.863237
No. 5 closest movie: Nobody's Fool (1994), distance: 0.866704
No. 6 closest movie: Birdcage, The (1996), distance: 0.870074
No. 7 closest movie: Three Colors: White (1994), distance: 0.870972
No. 8 closest movie: Trust (1990), distance: 0.875270
No. 9 closest movie: Streetcar Named Desire, A (1951), distance: 0.961867

No. 5 popular centroid:    0.9440    0.7565   -0.3897    0.8711   -0.4700   -0.0514   -0.2698    0.5625
1.3979    0.1243

The number of movies allocated to the centroid: 63
No. 1 closest movie: Beavis and Butt-head Do America (1996), distance: 0.906344
No. 2 closest movie: Frighteners, The (1996), distance: 0.987641
No. 3 closest movie: Austin Powers: International Man of Mystery (1997), distance: 1.021540
No. 4 closest movie: In the Line of Duty 2 (1987), distance: 1.043768
No. 5 closest movie: Tales from the Crypt Presents: Bordello of Blood (1996), distance: 1.099267
No. 6 closest movie: Tetsuo II: Body Hammer (1992), distance: 1.118190
No. 7 closest movie: April Fool's Day (1986), distance: 1.128640
No. 8 closest movie: Heidi Fleiss: Hollywood Madam (1995) , distance: 1.150258
No. 9 closest movie: Hurricane Streets (1998), distance: 1.153277
```