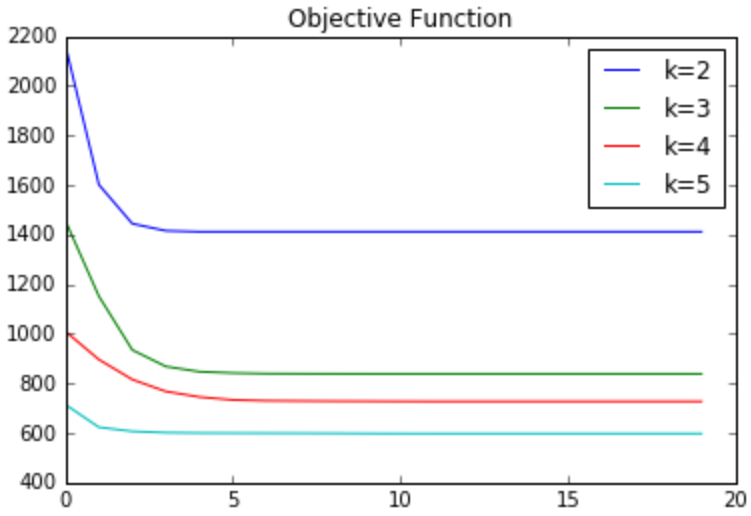
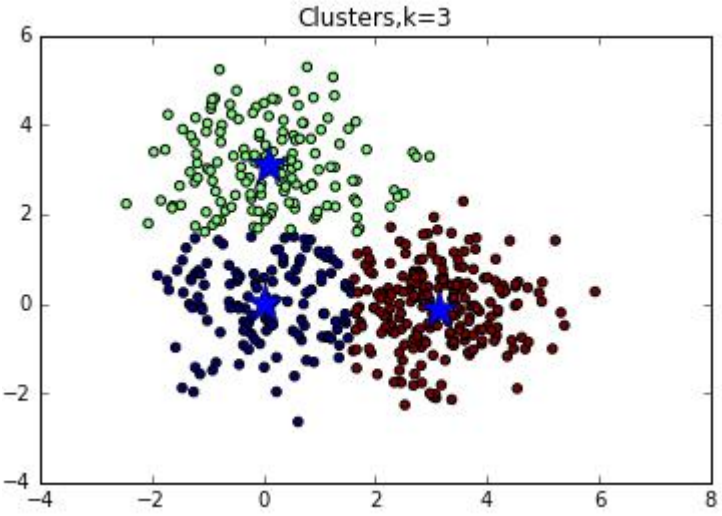
Anubha Bhargava
Homework 4

Problem 1

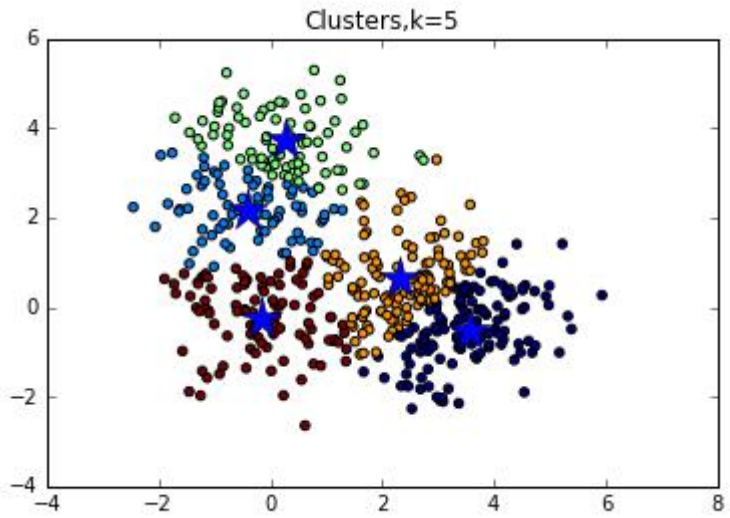
1. For K = 2,3,4,5 plot the value of the K-means objective function per 20 iterations.



2. For K = 3,5 plot the 500 data points and indicate the cluster of each final iteration by marking it with a symbol (blue star)

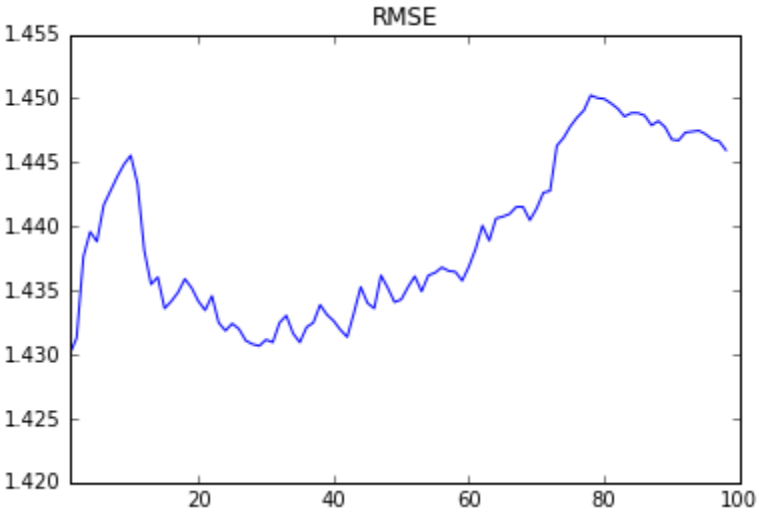


Anubha Bhargava
Homework 4

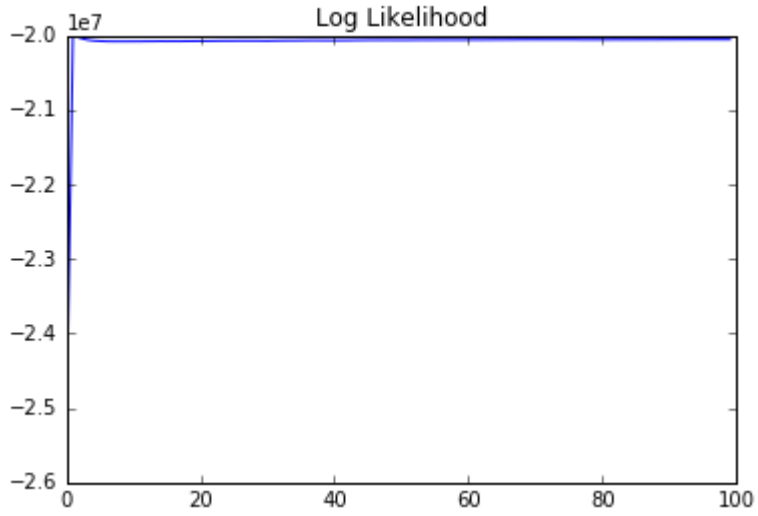


Problem 2:

1. Plot the RMSE for iterations 2 – 100.
2. Show the log joint likelihood for iterations 2 – 100.



Anubha Bhargava
Homework 4



The log likelihood function should ideally gradually reach its value which it plateaus at, rather than spiking to the value as shown above.

3. Pick 3 movies and find the 5 closest movies according to Euclidean distance. List the query movie, five nearest movies and their distances.

```
Toy Story (1995)
[['0.662341248557' '0.733372040639' '0.735039521642' '0.774326307024'
  '0.776536712279']
 ['Sleeper (1973)\n' 'Home Alone (1990)\n' 'Matilda (1996)\n'
  'Right Stuff, The (1983)\n' 'Full Metal Jacket (1987)\n']] Sound of Music,
The (1965)
[['0.903385872257' '0.929495124962' '0.942807933684' '0.985724592066'
  '1.0075351503']
 ['Cinderella (1950)\n' 'My Fair Lady (1964)\n' 'Sabrina (1995)\n'
  "Ulee's Gold (1997)\n" 'Time to Kill, A (1996)\n']] Batman (1989)
[['0.879177912635' '0.896469185623' '0.941425511232' '1.0016365567'
  '1.01760513994']
 ['Young Guns (1988)\n' 'Assassins (1995)\n' 'Mimic (1997)\n'
  'Mask, The (1994)\n' 'Back to the Future (1985)\n']]
```

Toy Story (1995)	Sleeper (1973) – 0.662341248557 Home Alone (1990) – 0.7333720406039 Matilda (1996) – 0.735039521642 Right Stuff, The (1983) – 0.774326307024 Full Metal Jacket (1987) – 0.776536712279
Sound of Music, The (1965)	Cinderella (1950) – 0.903385872257 My Fair Lady (1964) – 0.929495124962 Sabrina (1995) – 0.942807933684 Ulee's Gold (1997) – 0.985724592066 Time to Kill, A (1996) – 1.007535103
Batman (1989)	Young Guns (1988) – 0.879177912635

	Assassins (1995) - 0.896469185623 Mimic (1997) – 0.941425511232 Mask, The (1994) – 1.0016365567 Back to the Future (1985) – 1.01760513994

4. Pick 5 centroids corresponding to 5 clusters for vector u. Give number of users allocated to that cluster. List the 10 movies with the largest dot product with that centroid.

There are 20 centroids in total. The first output array shows the number of users that occurred the most frequently. Centroid 11 had the most number of users. It had 89 users.

```
5 centroids that have the most data in u: [[ 5 15 19 13 11]
 [66 70 72 76 89]]
```

The following shows the output from Python for the 10 movies with the largest dot product for the 5 centroids specified above for vector u.

The ten movies with the largest dot product	
Centroid: 5	Raiders of the Lost Ark (1981) – 4.21334026982 Princess Bride, The (1987) – 4.21502904662 Cinema Paradiso (1988) – 4.22534049365 Empire Strikes Back, The (1980) – 4.27258013918 Return of the Jedi (1983) – 4.27522133099 Wallace & Gromit: The Best of Aardman Animation (1996) – 4.31887207014 Casablanca (1942) – 4.26835616301 Star Wars (1977) – 4.55081205496 Wrong Trousers, The (1993) – 4.56486639397 **Close Shave, A (1995)** – 4.56553899311
Centroid: 15	Reservoir Dogs (1992) – 4.25519446023 Wallace & Gromit: The Best of Aardman Animation (1996) – 4.28460364839 Shawshank Redemption, The (1994) – 4.29240322605 Apt Pupil (1998) – 4.32417341718 Godfather, The (1972) – 4.34531554921 Blade Runner (1982) – 4.40219865924 Usual Suspects, The (1995) – 4.43373179536 Star Wars (1977) – 4.51285005199 L.A. Confidential (1997) – 4.53919638966 **Pulp Fiction (1994) – 4.5490330934**
Centroid: 19	Star Wars (1977) – 4.50477497358 Rear Window (1954) – 4.53543283715 Graduate, The (1967) – 4.54488964911 Lawrence of Arabia (1962) – 4.54713936899 Boot, Das (1981) – 4.56551566226 Pulp Fiction (1994) – 4.59899005421 Citizen Kane (1941) – 4.62532944274 Godfather: Part II, The (1974) – 4.63693672953 Casablanca (1942) – 4.69245258647

	Godfather, The (1972) – 4.89525022316
Centroid: 13	Usual Suspects, The (1995) – 4.37778591903 Casablanca (1942) – 4.38445931604 To Kill a Mockingbird (1962) – 4.38653449772 When We Were Kings (1996) – 4.41221975674 Schindler's List (1993) – 4.5254829671 Shawshank Redemption, The (1994) – 4.53460068398 Close Shave, A (1995) – 4.53719400686 Lawrence of Arabia (1962) – 4.54560066306 **Godfather, The (1972) – 4.57283618075**
Centroid: 11	Rear Window (1954) – 4.36560165062 Three Colors: Red (1994) – 4.38432092534 Amistad (1997) – 4.38497650364 Casablanca (1942) – 4.41628095478 Secrets & Lies (1996) – 4.42623765429 Kolya (1996) – 4.43792299547 One Flew Over the Cuckoo's Nest (1975) – 4.45337452692 Fargo (1996) – 4.45731845132 Dead Man Walking (1995) – 4.50567323574 **Schindler's List (1993) – 4.52529714987**

5. Pick 5 centroids corresponding to 5 clusters for vector v. Give number of movies allocated to that cluster. List the 10 movies with the smallest Euclidean distance to that centroid.

For vector v, Centroid 3 had the most movies allocated to the cluster. Centroid 3 has 127 movies allocated to the cluster. The results outputted from Python are shown below.

```
5 centroids that have the most data in v: [[ 18   0  14  19   3]
 [100 107 113 114 127]]
```

The following shows the output from Python for the 10 movies with the smallest Euclidean distance for the 5 centroids specified above for vector v.

The ten movies with the smallest Euclidean Distance	
Centroid: 18	**Mother Night (1996) - 0.416528898954** Maya Lin: A Strong Clear Vision (1994) – 0.4723410534 Nobody's Fool (1994) - 0.49002060303 Walkabout (1971) – 0.511358545584 Looking for Richard (1996) - 0.512298260982 Horseman on the Roof, The (Hussard sur le toit, Le) (1995) – 0.516410023009 Monty Python's Life of Brian (1979) – 0.529649335251 Four Days in September (1997) – 0.53562361054 A Chef in Love (1996) – 0.547933180386 Thirty-Two Short Films About Glenn Gould (1993) – 0.559248772608
Centroid: 0	**Maybe, Maybe Not (Bewegte Mann, Der) (1994) – 0.2621994108998** Sex, Lies, and Videotape (1989) – 0.31568528517 Month by the Lake, A (1995) – 0.351576068039

Anubha Bhargava
Homework 4

	Before the Rain (Pred dozhdot) (1994) – 0.393186638808 Passion Fish (1992) – 0.413948385431 Purple Noon (1960) – 0.449186080725 Foreign Correspondent (1940) – 0.46557051656 Seventh Seal, The (Sjunde inseglet, Det) (1957) – 0.46912187097 Man of the Year (1995) – 0.494571808504 Farewell to Arms, A (1932) – 0.49639500417
Centroid: 14	**Police Story 4: Project S (Chao ji ji hua) (1993) – 0.170070570577** King of New York (1990) – 0.177642156433 Amityville: Dollhouse (1996) – 0.187509275757 Venice/Venice(1992) – 0.187509216 Further Gesture, A (1996) – 0.18772160211 Temptress Moon (Feng Yue) (1996) – 0.19143654737 Getting Away With Murder (1996) – 0.193799630514 Babyfever (1994) – 0.198225403288 Somebody to Love (1994) – 0.200101089357 Boys in Venice (1996) – 0.200116471322
Centroid: 19	**Underneath, The (1995) – 0.293868066471** Wooden Man's Bride, The (Wu Kui) (1994) – 0.321748705614 I Can't Sleep (J'ai pas sommeil) (1994) – 0.33599512417 Hollow Reed (1996) – 0.3572669881 Wild Bill (1995) – 0.361405448362 American Dream – 0.374913924186 Monie, La (1995) – 0.38829283358 Slingshot, The (1993) – 0.38936785073 Brothers in Trouble (1995) – 0.39096471392 Butcher Boy, The (1998) – 0.390964713912
Centroid: 3	**Twilight (1998) – 0.254685062162** Talking About Sex (1994) – 0.273204709009 Mamma Roma (1962) – 0.2759442616 Outlaw, The (1943) – 0.309415687107 Jerky Boys, The (1994) – 0.318351476039 Stars Fell on Henrietta, The (1995) – 0.323319930892 Suture (1993) – 0.327115393694 Substance of Fire, The (1996) – 0.39192562731 Three Lives and Only One Death (1996) – 0.349008739228 All Things Fair (1996) – 0.3571574982