COGS 109 Fall 2015
Due: 11/08/2015, 11:59Pm
Grade: ___out of 100 points

Homework assignment 4: K-means and mean-shift clustering

In the questions below, we will guide you through the basics of performing k-means and mean-shift algorithm. Please download file "hw4_data.mat" at:
https://sites.google.com/site/ucsdcogs109fall2015/assignments/assignment4/hw4_data.mat

Matrix "kmeandata" includes 560 sample points with each sample being a two-dimensional row vector.

Note:

- Copy your code and insert the corresponding figures, if any, to answer each question below.
- If you follow the hints, it should be generally straightforward to write your code.
- If you are not sure how to call a particular function in matlab, remember to use the matlab command "help" followed by the function name to learn about its argument specification. You can also find a lot of information by Googling your queries.

1. (**20 points**) Step-by-step K-means. Let k = 2 and initialize the two means (cluster centers) at [-2,-2] and [4,4] respectively.

   (a). Initialize the means (cluster centers)
   Hint: `centers = [-2,-2;4,4];`

   (b). Calculate the distance between each data point to each center.

   Hint: For example, create a "for loop" with j indexing the cluster id.

   ```
   dist(:,j) = sqrt((kmeandata(:,1) - centers(j,1)).^2 +
   (kmeandata(:,2) - centers(j,2)).^2);
   ```

   (c). Assign each data point with the id of its nearest cluster center.

   Hint: `[v g_ind] = min(dist,[],2);`

   (d). Plot clustered data points using different colors for different clusters (points in the same cluster should have the same color and remember to use "hold on").

   Hint: e.g. for cluster 1

   ```
   scatter(kmeandata(g_ind == 1,1),kmeandata(g_ind ==
   1,2),20,'r','filled');
   scatter(centers(1,1),centers(1,2),80,'Marker','x','MarkerE
   dgeColor',[0.6 0 0],'LineWidth',4);
   ```

   (e). Update the means.

   Hint: .e.g for cluster 1

   ```
   centers(1,:) = mean(kmeandata(g_ind == 1,:));
   ```

   (f). Repeat this procedure for 4 times.

Create a figure of 2x2 subplots with each subplot showing result of step (d).

Hint: Create a for loop to run steps (b)-(e) for 4 iterations. Use function "`title(sprintf('Iteration %d', i))`" to show which iteration the subplot corresponds to.

2. (**20 points**) Create another figure of 2 by 2 subplots under the same procedure as described in question 1. However, this time, use the initial centers at [0,-1] and [0,5]. Attach your figure here and describe your observations.

3. (**20 points**) Create another figure of 2 by 2 subplots under the same procedure as described in question 1. However, this time, use k = 3 and initialize the cluster centers at [2,-2],[-2,-2],and[2,2]. Attach your figure here and describe your observations.

   Hint: You have three clusters in this case; adjust your code from question1 accordingly.

4. (**20 points**) Create another figure of 2 by 2 subplots under the same procedure as described in question 1. However, this time, use k = 4 and initialize the cluster centers at [0,-3], [0,-1],[0, 1], and [0,4]. Attach your figure here and describe your observations.

   Hint: You have four clusters in this case; adjust your code from question1 accordingly.

5. (**20 points**) Implement mean-shift algorithm and let k=1 for simplicity. Use a uniform kernel of radius 1.5.

   (a) Initialize the mean at [4,4] and set radius to be 1.5

   Hint:

```
center = [4,4];
radius = 1.5;
```

   (b) Plot all the points and use "hold on"

   Hint:

```
scatter(kmeandata(:,1),kmeandata(:,2),20,'b','filled');
hold on;
```

   (c) plot a disk of radius 1.5 centered at the current mean using color "green"
   Hint:

```
plot(center(:,1) + radius*cos([0:0.01:2*pi]),center(:,2) +
radius*sin([0:0.01:2*pi]),'--','Color',[0,1,0],'LineWidth',1);
```

(d) update the mean (cluster center) using mean-shift algorithm
Hint:

```
dist(:,1) = sqrt((kmeandata(:,1) - center(1,1)).^2
+((kmeandata(:,2) - center(1,2)).^2));
ind = dist <= radius;
center = mean(kmeandata(ind,1:2));
```

(e) repeat (c) and (d) for 10 times
Hint: remember to plot all disks in the same figure.

(f) plot the final cluster mean after 10 iterations (step (e)) using red "*".

6. (**15 Bonus points**) Combine the PCA and K-means.

In homework assignment #3, we performed PCA on the facemat data. Now let's perform k-means (k=4) clustering on the principal component space.

(a) Follow the instructions in assignment #3 but use the top 3 principal components. You should now deal with a matrix of 500x3 (or 3x500 if column vector) after the projection.

(b) Implement the k-means algorithm as outlined in question 1 with k=4. You can randomly pick four out of the 500 faces as initialized cluster means.

(d) Run 10 iterations of the k-means algorithm and then plot the final result, similar to question 1d (only one plot is needed to show the final result). Attach your figure here (you only need to show the 3D points after the projection instead of the original faces).

Hint: You will need to use the scatter3 instead of scatter to generate 3D scatter plots.