Boston University
Department of Electrical and Computer Engineering

ENG EC 414 (Ishwar) Introduction to Machine Learning

HW 3

Issued: Fri 18 Feb 2022 **Due:** 10:55pm Boston time Fri 25 Feb 2022 in Gradescope + Blackboard.
Required reading: Slides on clustering + your notes from lectures and Discussion 4.

Important: Before you proceed, please read the documents pertaining to *Homework formatting and submission guidelines* and the *HW-grading policies* in the Homeworks section of Blackboard, especially guidelines for submitting reports in Gradescope. and code in Blackboard. In particular, for computer assignments *you are prohibited from using any online code or built-in MATLAB functions except as indicated in the problem or skeleton code (when provided).*
In order to receive full credit, all work should be supported by a concise explanation that is clear, relevant, specific, logical, and correct. In particular, for each part, you must clearly outline the key steps and provide proper justification for your calculations.

Note: Problem difficulty = number of coffee cups ☕

Problem 3.1 [18pts] *(Impact of initialization on k-means)* Consider a one-dimensional dataset of $n = 2m+1$ points ($m \geq 1$) spread on the real line with one point at $x = 0$, m points at $x = a$, and m points at $x = (a + b)$ where $0 < a \leq b < 2a$. Now suppose we run the k-means algorithm on this dataset with $k = 2$ and *distinct* initial centers chosen over the range $[0, a + b]$. If a cluster become empty, the algorithm skips the center update step for that cluster and continues until there is no change in any cluster.

(a) [8pts] Determine all distinct solutions that the k-means algorithm could actually converge to.

(b) [4pts] For each solution, provide expressions for (i) centers (ii) WCSS, in terms of m, a, b.

(c) [2pts] Compute the ratio of the largest and smallest WCSS as a function of m, a, b and comment on the implications.

(d) [4pts] ☕☕ Suppose that the two initial centers are chosen independently and uniformly at random over the range $[0, a + b]$. Compute the probability of converging to each possible solution.

Problem 3.2 [44pts] *(k-means implementation)* In this problem we will implement k-means clustering and explore the impact of initialization and number of clusters on one synthetic and one real-world dataset. We will also explore a dataset where k-means will fail to produce meaningful clusters. You are provided skeleton code to assist you in implementing this clustering method. **You are prohibited from using any online code or built-in MATLAB functions that provide a complete standalone implementation of k-means clustering.**

(a) [12pts] *(Synthetic training set generation)* Generate 3 two-dimensional Gaussian clusters of data points having the following mean vectors and covariance matrices: $\boldsymbol{\mu}_1 = [2, 2]^\top$, $\boldsymbol{\mu}_2 = [-2, 2]^\top$, $\boldsymbol{\mu}_3 = [0, -3.25]^\top$, and $\Sigma_1 = 0.02 \cdot I_2$, $\Sigma_2 = 0.05 \cdot I_2$, $\Sigma_3 = 0.07 \cdot I_2$, where I_2 is the 2×2 identity matrix. Let each data cluster have 50 points. Create a scatter plot of the generated Gaussian data. Color the data points in the 1st, 2nd, and 3rd clusters with red, green, and blue colors, respectively. Implement

k-means and test it using $k = 3$ with the following **initialization**: $\boldsymbol{\mu}_1^{\text{initial}} = [3, 3]^\top$, $\boldsymbol{\mu}_2^{\text{initial}} = [-4, -1]^\top$, $\boldsymbol{\mu}_3^{\text{initial}} = [2, -4]^\top$. and the following **stopping criterion:** stop if the derived cluster means become stationary (i.e., do not change by more than a suitable threshold over iterations). Create a separate scatter plot to visualize the clusters produced by your k-means algorithm.

(b) [4pts] *(Effect of different initialization)* Using the same synthetic training dataset from part (a), re-run your k-means algorithm implementation for $k = 3$ using the following (different) **initialization:** $\boldsymbol{\mu}_1^{\text{initial}} = [-0.14, 2.61]^\top$, $\boldsymbol{\mu}_2^{\text{initial}} = [3.15, -0.84]^\top$, $\boldsymbol{\mu}_3^{\text{initial}} = [-3.28, -1.58]^\top$. Create a new scatter plot of the resulting clusters. Discuss what you observe and how it relates to what you learned in class.

(c) [10pts] *(Best of multiple random initializations)* To reduce the possibility selecting an initialization which results in a "bad" clustering (high WCSS), the k-means algorithm is typically run multiple times using different random initializations. The best clustering result, i.e., the one having the smallest WCSS is saved and used as the final output. Run your implementation of the k-means algorithm on the same synthetic training dataset from part (a) for 10 different random initializations. Report the WCSS values for each of the 10 trials. Identify the trial which yields the smallest WCSS value. Report its WCSS value and create a scatter plot of the clustering produced by it.

Note: From this part onwards, whenever you are asked to run your implementation of the k-means algorithm, you should select the best of 10 different random initializations as the final output. We suggest choosing centers either 1) uniformly at random from the dataset or alternatively 2) uniformly distributed over a tight rectangular region encompassing the dataset.

(d) [8pts] *(Heuristic choice of k via "elbow" method)* **Note:** This part is closely related to Problem 3.3. For the same synthetic training dataset from part (a), run your implementation of the k-means algorithm for each $k \in \mathcal{K}_{\text{range}} = \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$. For each value of k, select the best clustering out of 10 different random initializations as the final output for that value of k (as in part(c)). Plot the WCSS values (of the final outputs) against k for all $k \in \mathcal{K}_{\text{range}}$. Discuss what you observe and how it relates to what you learned in class.

(e) [6pts] *(Clustering a real-world dataset)* Here we examine a real-world dataset containing National Basketball Association (NBA) statistics from the 2018-2019 season. Read in the NBA data from the "NBA_stats_2018_2019.xlsx" file and plot the Points Per Game (PPG) versus Minutes Per Game (MPG) statistics for all NBA players (a player is represented by a row of the data). The PPG and MPG information form a 2D dataset. Apply your implementation of the k-means algorithm with $k = 10$ selecting the best of 10 different random initializations as the final output. Create a scatter plot of the resulting clusters.

(f) [4pts] *(Failure of k means)* Here we examine the performance of the k-means algorithm on a dataset composed of 3 concentric rings. Use `sample_circle.m` to generate a dataset with $k = 3$ concentric ring clusters and 500 points for each cluster. Create a scatter plot of the dataset. Apply your implementation of the k-means algorithm on this dataset using $k = 3$ and choosing the best of 10 different random initializations. Create a scatter plot of the best clustered results. Discuss what you observe and how it relates to what you learned in class.

Problem 3.3 [10pts] *(Selecting k via k-means WCSS + penalty λk)* From the WCSS versus k plot of Problem 3.2(d) we know that the k-means WCSS cost decreases as k increases. A heuristic method for finding the "correct" number of clusters is to identify a sharp bend in this curve. A more principled method is to add, for each value of k, a *regularization penalty* term λk, for some $\lambda > 0$, to the WCSS of the clustering returned

by the k-means algorithm and then select the value of k for which the sum is minimum. Specifically, let $f(k, \lambda) = \text{WCSS}_{k-\text{means}} + \lambda k$, where $\text{WCSS}_{k-\text{means}}$ is the WCSS of the clustering returned by running the k-means algorithm with the specified value of k and $\lambda > 0$ is a cluster penalty parameter that discourages finding solutions that have too many clusters. For each $\lambda \in \{15, 20, 25, 30\}$, plot $f(k, \lambda)$ as a function of k for $k \in \mathcal{K}_{\text{range}}$ of Problem 3.2(d) Discuss what you observe and how it relates to what you learned in class. In particular, comment on the effect of the value of λ. **Note:** λ is related to the *squared* radius and not radius of clusters because the base cost (without penalty) is the sum of *squared* Euclidean distances.

Problem 3.4 [22pts] *(DP-means implementation)* Here you will implement the DP-means clustering algorithm and apply it to both synthetic and real-world data. You are provided skeleton code to assist you in implementing this clustering method. **Use of online code or any built-in MATLAB functions that provide a complete standalone implementation of DP Means is prohibited.** Use the following **stopping criterion:** stop if the number of existing clusters stays the same **and** the derived cluster means become stationary (i.e., do not change by more than a suitable threshold over iterations). **Note:** In DP means, you should ignore empty clusters and discard them at the very end. The final value of k is the number of non-empty clusters.

(a) [2pts] Describe the role of the λ parameter in the DP-means algorithm.

(b) [15pts] Apply your implementation of the DP-means algorithm to the synthetic 3 Gaussian cluster dataset from Problem 3.2(a) for each $\lambda \in \{0.15, 0.4, 3, 20\}$. For each λ, create a scatter plot of the obtained clusters and identify which λ yields the best clustering performance. (Here, 'best' means the most meaningful clustering as can be observed visually: if you can clearly observe 'k' clusters in the scatter plot of the dataset, for which value of λ do we get a clustering that is close to it?) Discuss what you observe and how it relates to what you learned in class. In particular, comment on the effect of the value of λ.

(c) [5pts] Apply your implementation of the DP-means algorithm to the NBA data from Problem 3.2(e) for each $\lambda \in \{44, 100, 450\}$. For each λ, create a scatter plot of the obtained clusters. Discuss what you observe and how it relates to what you learned in class. In particular, comment on the effect of the value of λ.

Code-submission via Blackboard: Create three "dot m" files. One named <yourBUemailID>_hw3_2.m for Problem 3.2, one named <yourBUemailID>_hw3_3.m for Problem 3.3, and one named <yourBUemailID>_hw3_4.m for Problem 3.4. When run, your scripts should be able to display in the command window whatever you are asked to compute and report in each part. Two skeleton codes are provided for your reference: skeleton_hw3_2.m and skeleton_hw3_4.m. Note that Problem 3.3 relies upon the code for Problem 3.2. Reach-out to the TAs via Piazza and office hours for questions related to coding. When submitting code, please include a 'Readme' text file in your source code directory describing approximate running times of different parts, any additional comments (as needed) explaining how to use your code, and any dependencies between different parts. Please do not include into the directory, any data files that are already provided. Write your code under the assumption that all data files are in the same directory as your source code.

Problem 3.1 [18pts] *(Impact of initialization on k-means)* Consider a one-dimensional dataset of $n = 2m+1$ points ($m \geq 1$) spread on the real line with one point at $x = 0$, m points at $x = a$, and m points at $x = (a + b)$ where $0 < a \leq b < 2a$. Now suppose we run the k-means algorithm on this dataset with $k = 2$ and *distinct* initial centers chosen over the range $[0, a + b]$. If a cluster become empty, the algorithm skips the center update step for that cluster and continues until there is no change in any cluster. ← what I did bb

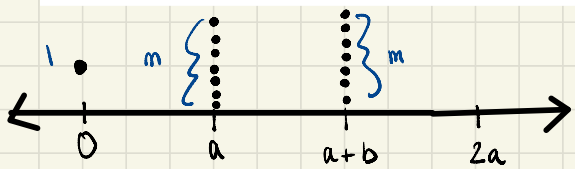
(a) [8pts] Determine all distinct solutions that the k-means algorithm could actually converge to. Are solns not centers?

(b) [4pts] For each solution, provide expressions for (i) centers (ii) WCSS, in terms of m, a, b.

(c) [2pts] Compute the ratio of the largest and smallest WCSS as a function of m, a, b and comment on the implications.

(d) [4pts] Suppose that the two initial centers are chosen independently and uniformly at random over the range $[0, a + b]$. Compute the probability of converging to each possible solution.

a)



Distinct centers in $[0, a+b]$

$x = 0 \rightarrow$ 1 point

$x = a \rightarrow$ m points

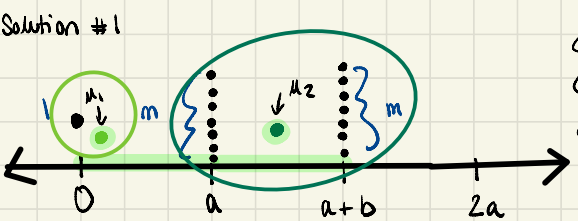
$x = a+b \rightarrow$ m points

All points at 3 distinct points

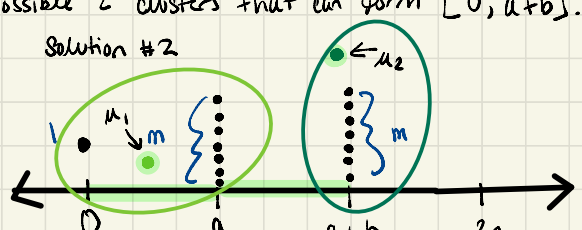
K-means is the average of clusters until stabilized, so what are possible averages?

For $k=2$, we have 2 clusters on the range of $[0, a+b]$. For 2 distinct solutions, we can have all m points in one cluster and the point at $x=0$ in another. Another solution could also be to form another cluster with the point at $x=0$ and all m points at $x=a$ in that cluster, then the other cluster will have all m points at $x=a+b$. These are the possible 2 clusters that can form $[0, a+b]$.

Solution #1



Solution #2



1) (1) $0/1$, $\frac{m(a+(a+b))}{2m} \to 0$, $\frac{2a+b}{2} \to 0$, $a+b/2$

Solution #1: $\mu_1 = 0$, $\mu_2 = a+b/2$

2) $\frac{0(1) + m(a)}{m+1}$, $\frac{m(a+b)}{m}$

$\frac{ma}{m+1}$, $a+b$

Solution #2: $\mu_1 = \frac{ma}{m+1}$, $\mu_2 = a+b$

b) i) Solutions' centers specified in 3.1a

ii) Solution #1 WCSS

Remember $WCSS = \sum_{k=1}^{K} \sum_{j \in C_k} (dist(x_j, \mu_k))^2$

$WCSS_1 = (0-0)^2 + m(a-(a+b/2))^2 + m(a+b-(a+b/2))^2$

$= 1 + m(b/2)^2 + m(-b/2)^2$

$= 1 + m((b/2)^2 + (-b/2)^2)$

$= \frac{mb^2}{2}$

Solution #2: $\mu_1 = ma/m+1$, $\mu_2 = a+b$

$WCSS_2 = (0 - ma/m+1)^2 + m(a-(ma/m+1))^2 + m(a+b-(a+b))^2$

$= (-ma/m+1)^2 + m(a-(ma/m+1))^2$

$= (-(ma/(m+1)))^2 + m\left(\frac{a(m+1)-ma}{m+1}\right)^2$

$= \frac{(ma)^2}{(m+1)^2} + m\left(\frac{am+a-ma}{m+1}\right)^2$

$= \frac{(ma)^2}{(m+1)^2} + m(a/m+1)^2$

$= m^2\left(\frac{a}{m+1}\right)^2 + m\left(\frac{a}{m+1}\right)^2$

$= m^2 + m\left(\frac{a}{m+1}\right)^2$

$= m(m+1)\left(\frac{a}{m+1}\right)^2$

$= m\,\frac{(m+1)}{(m+1)^{2\,1}}\,\frac{a^2}{}$

$= ma^2/m+1$

$\rightarrow m+1 \geq 2$, (b/c $m \geq 1$, so $m+1$, $1+1=2$) ← denominator is larger

$\frac{1}{m+1} \leq 1/2$

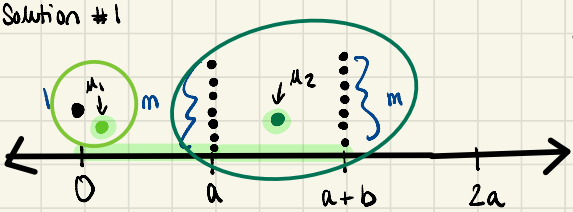
$\rightarrow a \leq b$, $a^2 \leq b^2$ so $WCSS_1$ is bigger

$$\frac{\frac{mb^2/2}{ma^2}}{\frac{}{m+1}} = mb^2/2 \cdot m+1/ma^2 = \frac{mb^2(m+1)}{2ma^2} = \frac{b^2(m+1)}{2a^2}$$

As the # of points increase, WCSS increases. $WCSS_{ratio} = \frac{b^2(m+1)}{2a^2}$

d)

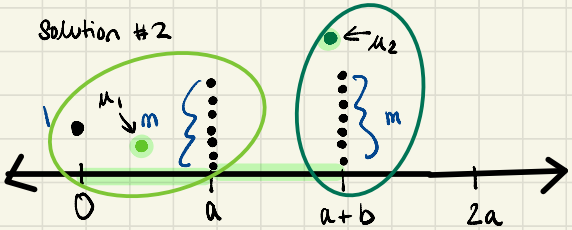
Solution #1



$\mu_1 \in [0, a/2]$, $\mu_2 \in [a, a+b]$

$Pr[soln\ 1] = \frac{[0, a/2] + [a+a+b]}{2}$

Solution #2

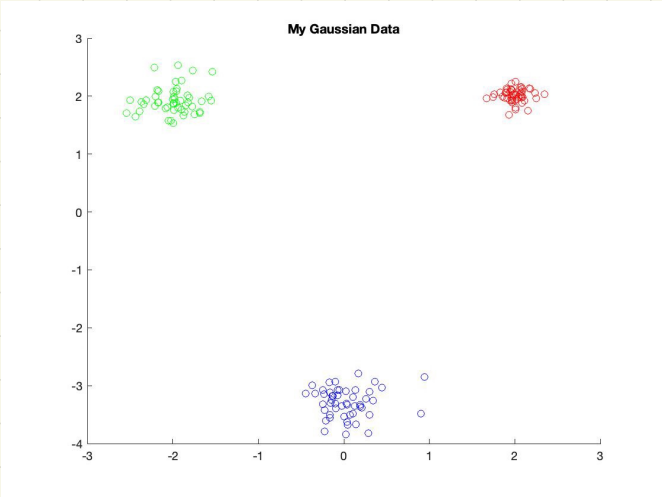


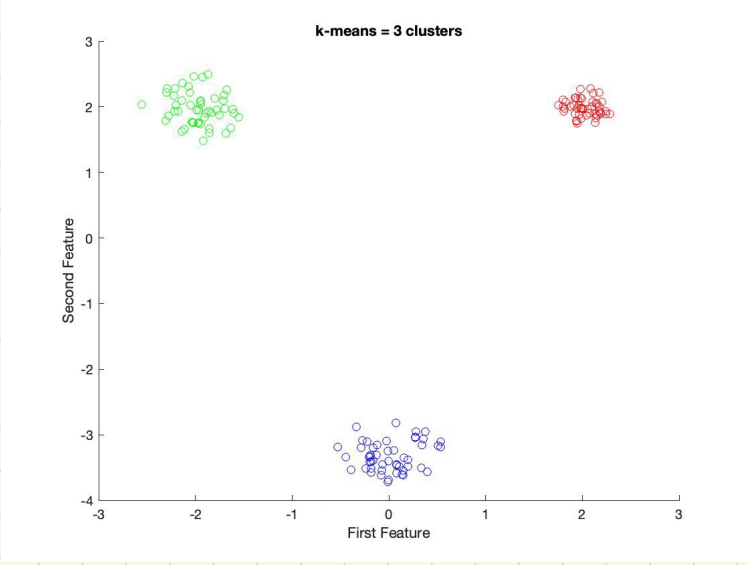
$\mu_1 \in [0, a]$, $\mu_2 \in [3a/2, a+b]$

$Pr[sol'n\ 2] = \frac{[0, a] + [3a/2, a+b]}{2}$

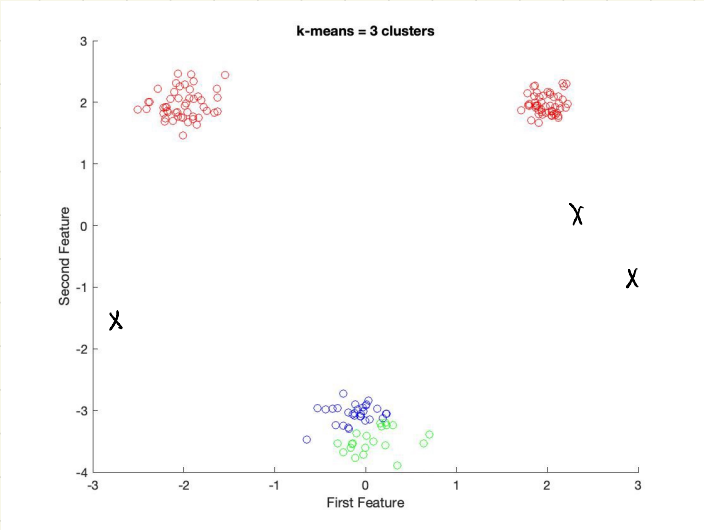
Problem 3.2 [44pts] *(k-means implementation)* In this problem we will implement k-means clustering and explore the impact of initialization and number of clusters on one synthetic and one real-world dataset. We will also explore a dataset where k-means will fail to produce meaningful clusters. You are provided skeleton code to assist you in implementing this clustering method. **You are prohibited from using any online code or built-in MATLAB functions that provide a complete standalone implementation of k-means clustering.**

(a) [12pts] *(Synthetic training set generation)* Generate 3 two-dimensional Gaussian clusters of data points having the following mean vectors and covariance matrices: $\boldsymbol{\mu}_1 = [2,2]^\top$, $\boldsymbol{\mu}_2 = [-2,2]^\top$, $\boldsymbol{\mu}_3 = [0,-3.25]^\top$, and $\Sigma_1 = 0.02 \cdot I_2$, $\Sigma_2 = 0.05 \cdot I_2$, $\Sigma_3 = 0.07 \cdot I_2$, where I_2 is the 2×2 identity matrix. Let each data cluster have 50 points. Create a scatter plot of the generated Gaussian data. Color the data points in the 1st, 2nd, and 3rd clusters with red, green, and blue colors, respectively. Implement k-means and test it using $k = 3$ with the following **initialization**: $\boldsymbol{\mu}_1^{\text{initial}} = [3,3]^\top$, $\boldsymbol{\mu}_2^{\text{initial}} = [-4,-1]^\top$, $\boldsymbol{\mu}_3^{\text{initial}} = [2,-4]^\top$. and the following **stopping criterion:** stop if the derived cluster means become stationary (i.e., do not change by more than a suitable threshold over iterations). Create a separate scatter plot to visualize the clusters produced by your k-means algorithm.





(b) [4pts] *(Effect of different initialization)* Using the same synthetic training dataset from part (a), re-run your k-means algorithm implementation for $k = 3$ using the following (different) **initialization:** $\mu_1^{\text{initial}} = [-0.14, 2.61]^\top$, $\mu_2^{\text{initial}} = [3.15, -0.84]^\top$, $\mu_3^{\text{initial}} = [-3.28, -1.58]^\top$. Create a new scatter plot of the resulting clusters. Discuss what you observe and how it relates to what you learned in class.

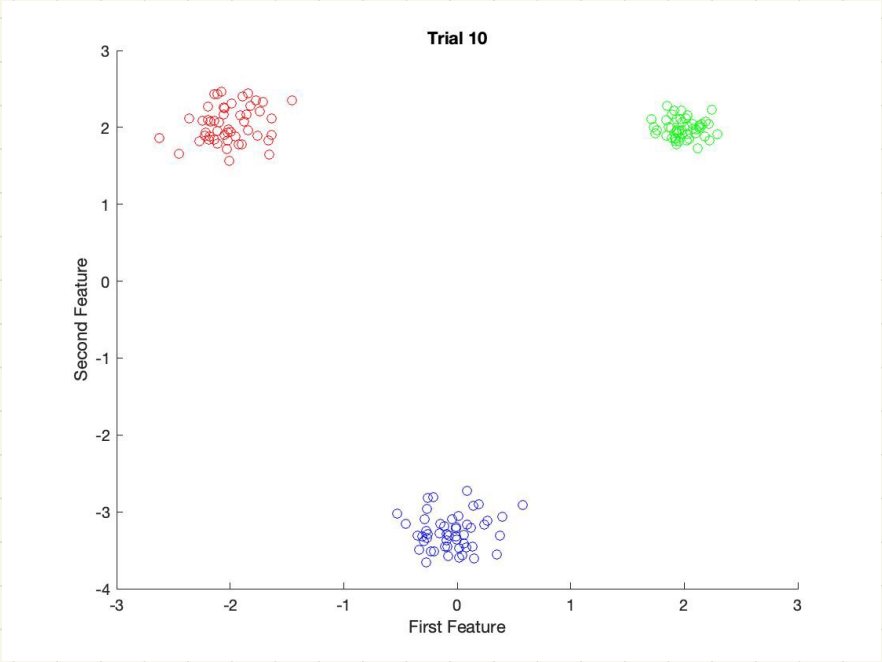


With a different initialization, the clusters are at different places and do not cluster all the points in their appropiate clusters. The new initial centers skew the clusters because some points in a bunch are closer than others.

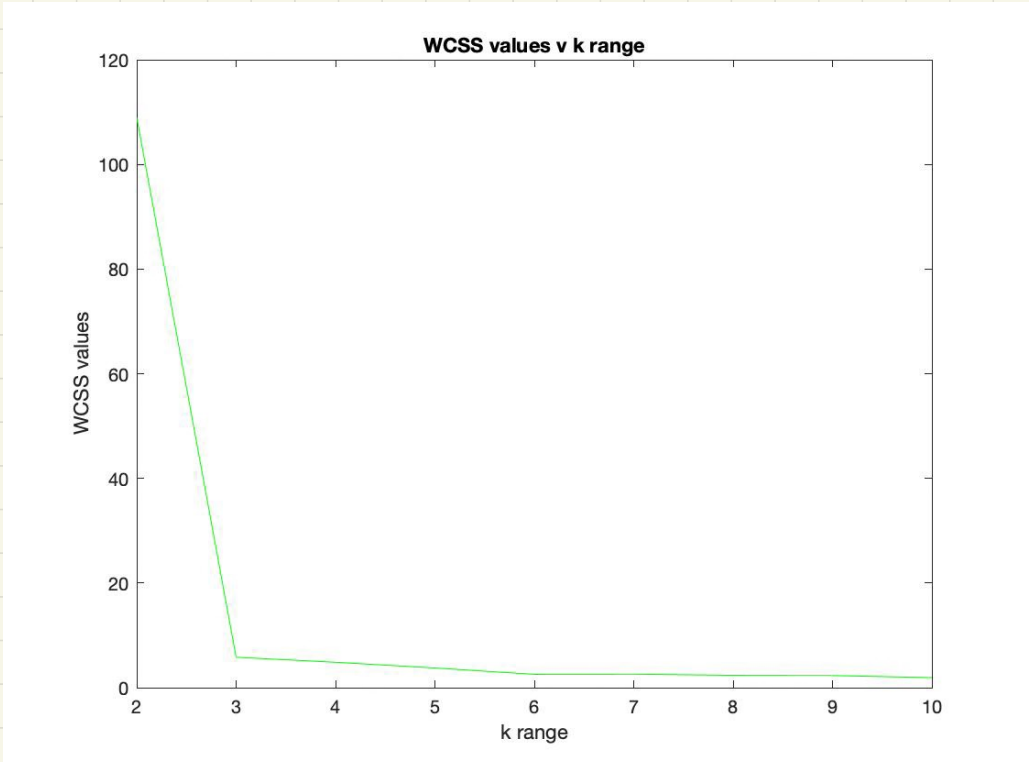
(c) [10pts] *(Best of multiple random initializations)* To reduce the possibility selecting an initialization which results in a "bad" clustering (high WCSS), the k-means algorithm is typically run multiple times using different random initializations. The best clustering result, i.e., the one having the smallest WCSS is saved and used as the final output. Run your implementation of the k-means algorithm on the same synthetic training dataset from part (a) for 10 different random initializations. Report the WCSS values for each of the 10 trials. Identify the trial which yields the smallest WCSS value. Report its WCSS value and create a scatter plot of the clustering produced by it.

Note: From this part onwards, whenever you are asked to run your implementation of the k-means algorithm, you should select the best of 10 different random initializations as the final output. We suggest choosing centers either 1) uniformly at random from the dataset or alternatively 2) uniformly distributed over a tight rectangular region encompassing the dataset.

```
Converged.
Trial 10 had the best wcss with a value of 5
```

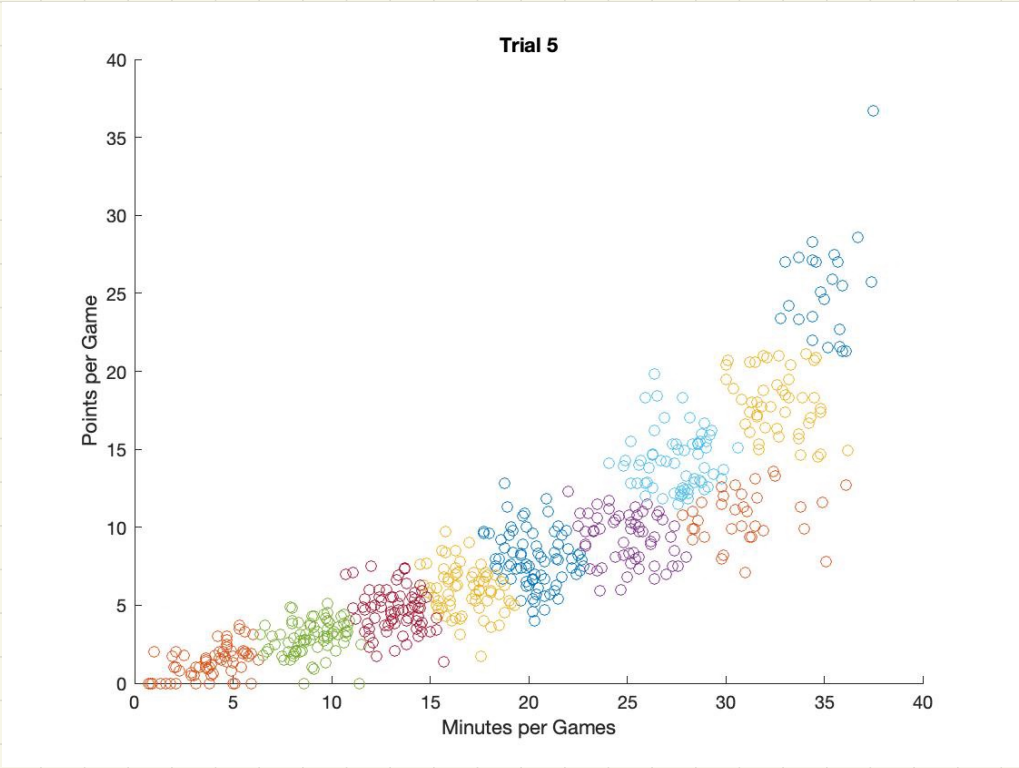


(d) [8pts] *(Heuristic choice of k via "elbow" method)* **Note:** This part is closely related to Problem 3.3. For the same synthetic training dataset from part (a), run your implementation of the k-means algorithm for each $k \in \mathcal{K}_{\text{range}} = \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$. For each value of k, select the best clustering out of 10 different random initializations as the final output for that value of k (as in part(c)). Plot the WCSS values (of the final outputs) against k for all $k \in \mathcal{K}_{\text{range}}$. Discuss what you observe and how it relates to what you learned in class.



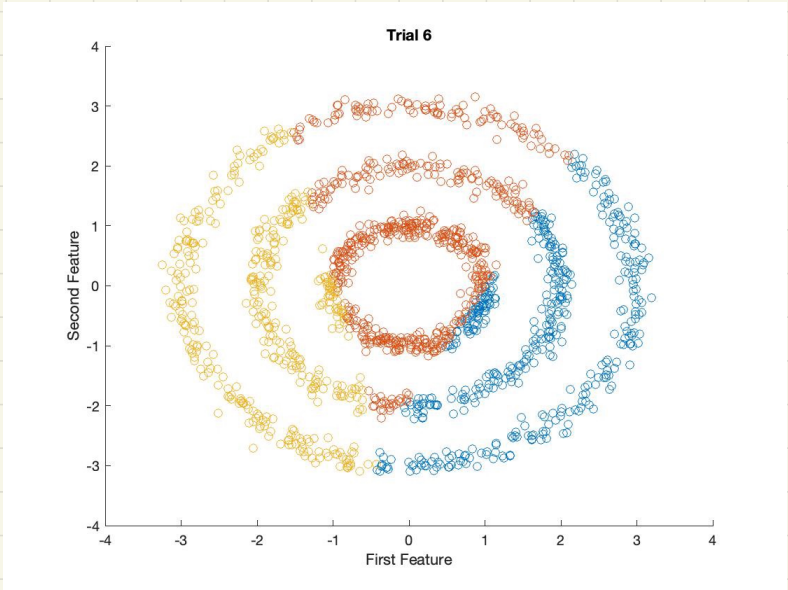
I see that as K increases the cost of WCSS decreases. When applying Kmeans we need to choose the best K that minimizes the sum. Using the elbow method, we can choose the best K where the slope is no longer decreasing exponentially and begins to decrease substantially slow like a line.

(e) [6pts] *(Clustering a real-world dataset)* Here we examine a real-world dataset containing National Basketball Association (NBA) statistics from the 2018-2019 season. Read in the NBA data from the "NBA_stats_2018_2019.xlsx" file and plot the Points Per Game (PPG) versus Minutes Per Game (MPG) statistics for all NBA players (a player is represented by a row of the data). The PPG and MPG information form a 2D dataset. Apply your implementation of the k-means algorithm with $k = 10$ selecting the best of 10 different random initializations as the final output. Create a scatter plot of the resulting clusters.




```
Converged.
Trial 5 had the best wcss with a value of 1127
```

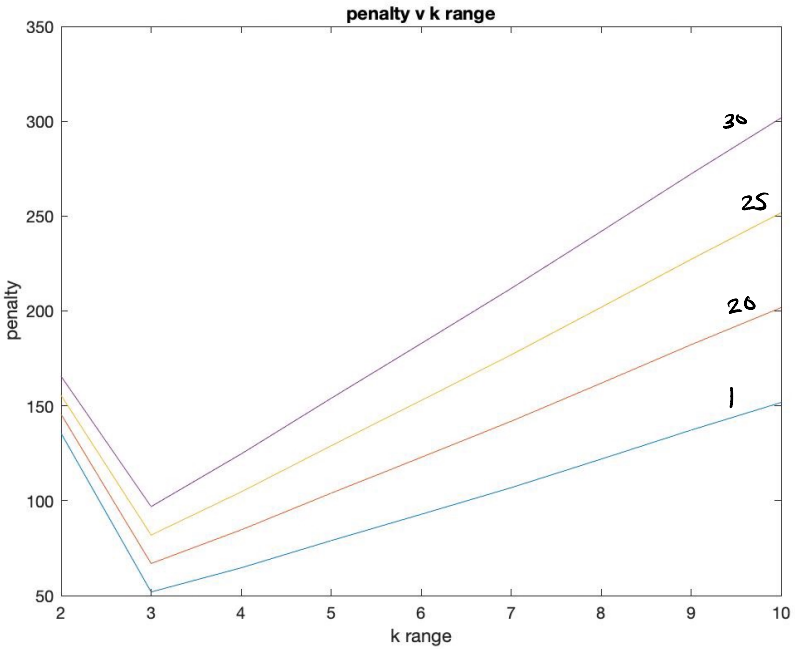
(f) [4pts] *(Failure of k means)* Here we examine the performance of the k-means algorithm on a dataset composed of 3 concentric rings. Use `sample_circle.m` to generate a dataset with $k = 3$ concentric ring clusters and 500 points for each cluster. Create a scatter plot of the dataset. Apply your implementation of the k-means algorithm on this dataset using $k = 3$ and choosing the best of 10 different random initializations. Create a scatter plot of the best clustered results. Discuss what you observe and how it relates to what you learned in class.




```
Converged.
Trial 6 had the best wcss with a value of 960
>>
```

The K-means algorithm fails. There are 3 distinct ringed groups, however the K-means can't detect the individual groups.

Problem 3.3 [10pts] *(Selecting k via k-means WCSS + penalty λk)* From the WCSS versus k plot of Problem 3.2(d) we know that the k-means WCSS cost decreases as k increases. A heuristic method for finding the "correct" number of clusters is to identify a sharp bend in this curve. A more principled method is to add, for each value of k, a *regularization penalty* term λk, for some $\lambda > 0$, to the WCSS of the clustering returned by the k-means algorithm and then select the value of k for which the sum is minimum. Specifically, let $f(k, \lambda) = \text{WCSS}_{k-\text{means}} + \lambda k$, where $\text{WCSS}_{k-\text{means}}$ is the WCSS of the clustering returned by running the k-means algorithm with the specified value of k and $\lambda > 0$ is a cluster penalty parameter that discourages finding solutions that have too many clusters. For each $\lambda \in \{15, 20, 25, 30\}$, plot $f(k, \lambda)$ as a function of k for $k \in \mathcal{K}_{\text{range}}$ of Problem 3.2(*d*) Discuss what you observe and how it relates to what you learned in class. In particular, comment on the effect of the value of λ. **Note:** λ is related to the *squared* radius and not radius of clusters because the base cost (without penalty) is the sum of *squared* Euclidean distances.



As k increases, the greater effect of the λ penalty.

Note $\lambda \equiv$ squared radius of cluster

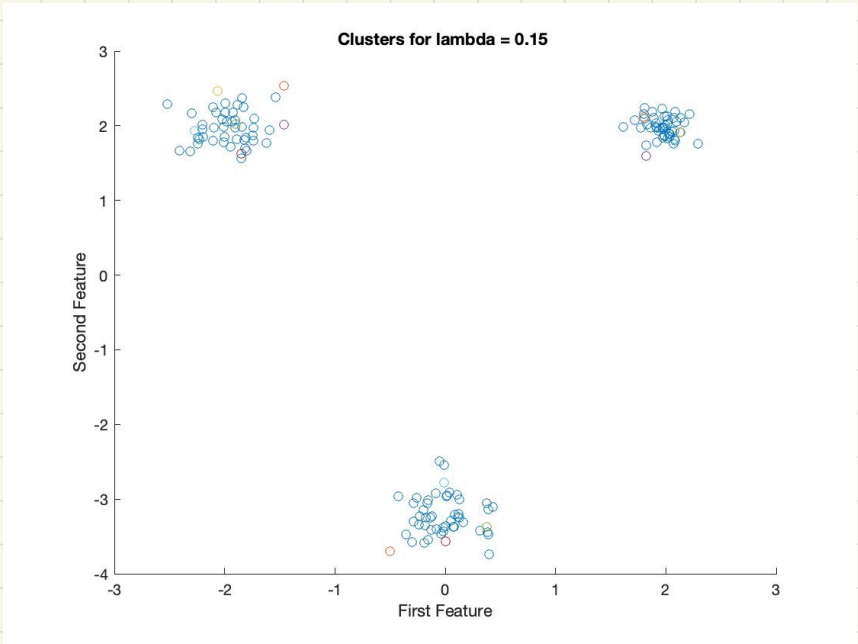
Problem 3.4 [22pts] *(DP-means implementation)* Here you will implement the DP-means clustering algorithm and apply it to both synthetic and real-world data. You are provided skeleton code to assist you in implementing this clustering method. **Use of online code or any built-in MATLAB functions that provide a complete standalone implementation of DP Means is prohibited.** Use the following **stopping criterion:** stop if the number of existing clusters stays the same **and** the derived cluster means become stationary (i.e., do not change by more than a suitable threshold over iterations). **Note:** In DP means, you should ignore empty clusters and discard them at the very end. The final value of k is the number of non-empty clusters.

(a) [2pts] Describe the role of the λ parameter in the DP-means algorithm.

(b) [15pts] Apply your implementation of the DP-means algorithm to the synthetic 3 Gaussian cluster dataset from Problem 3.2(a) for each $\lambda \in \{0.15, 0.4, 3, 20\}$. For each λ, create a scatter plot of the obtained clusters and identify which λ yields the best clustering performance. (Here, 'best' means the most meaningful clustering as can be observed visually: if you can clearly observe 'k' clusters in the scatter plot of the dataset, for which value of λ do we get a clustering that is close to it?) Discuss what you observe and how it relates to what you learned in class. In particular, comment on the effect of the value of λ.

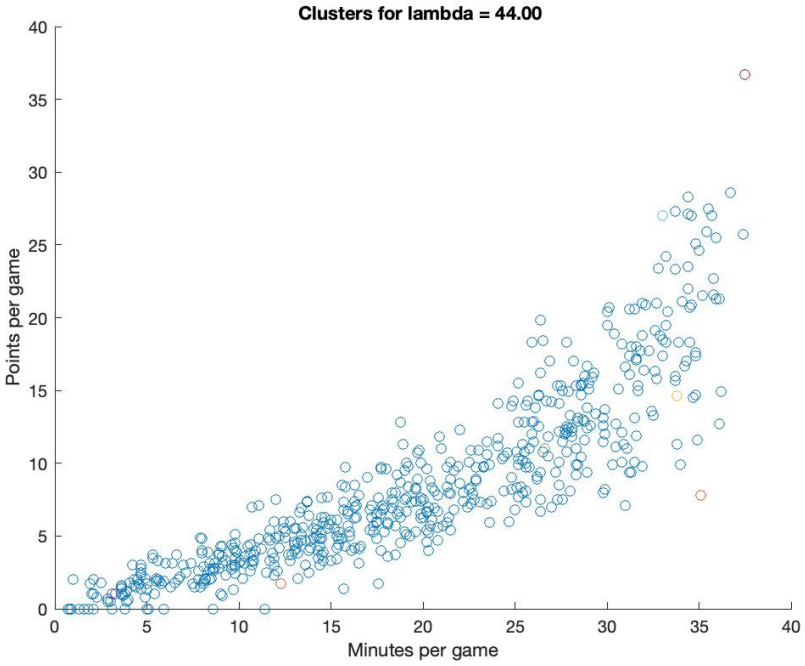
a) λ is the cluster penalty parameter. It helps determine the # of clusters to form in the algorithm.

b)



λ really determines the # of clusters. depending on the squared radius λ needs to be adjusted so the data can cluster properly. λ too big for data will cause less clusters and less effective K-means.

(c) [5pts] Apply your implementation of the DP-means algorithm to the NBA data from Problem 3.2(e) for each $\lambda \in \{44, 100, 450\}$. For each λ, create a scatter plot of the obtained clusters. Discuss what you observe and how it relates to what you learned in class. In particular, comment on the effect of the value of λ.



Similar to 3.4b, λ needs to be appropiately adjusted for the squared radius of data to be applied correctly and to form proper clusters. λ is much larger for the NBA dataset as the magnitude of data is greater.