CS 446: Machine Learning
Homework

1. [**10 points**] K-Means

   (a) Mention if K-Means is a supervised or an un-supervised method.

   > Your answer: K-Means is un-supervised method.

   (b) Assume that you are trying to cluster data points x_i for $i \in \{1, 2 \ldots D\}$ into K clusters each with center μ_k where $k \in \{1, 2, \ldots K\}$. The objective function for doing this clustering involves minimizig the euclidean distance between the points and the cluster centers. It is given by

   $$\min_{\mu} \min_{r} \sum_{i \in D} \sum_{k=1}^{K} \frac{1}{2} r_{ik} ||x_i - \mu_k||_2^2$$

   How do you ensure hard assignemnt of one data point to one and only one cluster at a given time? Note: By hard assignment we mean that your are 100 % sure that a point either belongs or not belongs to a cluster.

   > Your answer: $r_{ik} \in \{0, 1\}$ This will make any data point x_i belong to ($r_{ik} = 1$) or not belong to ($r_{ik} = 0$) centroid μ_k
   >
   > $$r_{ik} = \begin{cases} 1, & \text{if } ||x_i - \mu_k||^2 < ||x_i - \mu_l||^2, \forall l \neq k \\ 0, & \text{otherwise} \end{cases}$$

   (c) What changes must you do in your answer of part b, to make the hard assingment into a soft assignment? Note: By soft assignment we mean that your are sure that a point either belongs or not belongs to a cluster with some probability.

   > Your answer: $r_{ik} \in [0, 1]$ This will make i^{th} line of matrix R the probability vector of data point x_i considering a Softmax Function.
   >
   > $$r_{ik} = \frac{e^{-\beta ||x_i - \mu_k||^2}}{\sum\limits_{l=1}^{k} e^{-\beta ||x_i - \mu_l||^2}}$$
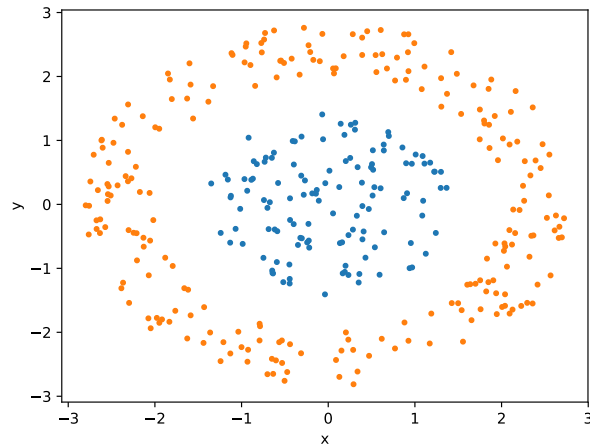   >
   > where the parameter $\beta > 0$ is called stiffness parameter.

   (d) Looking at the following plot, what is the best choice for number of clusters?
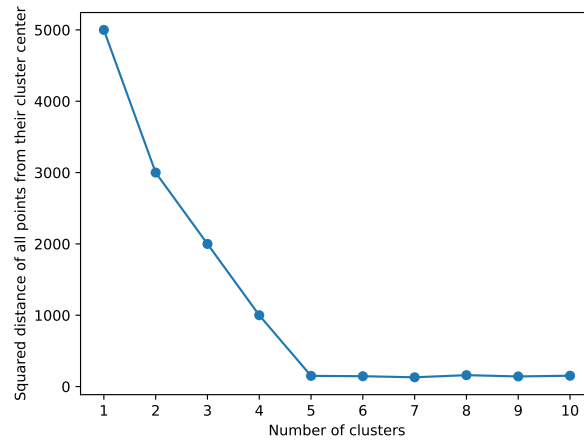
Your answer: The number of clusters should be 5 because it is the elbow of the plot.

(e) Would K-Means be an effecient algorithm to cluster the following data? Explain your answer in a couple of lines.

Your answer: No. Assume the first two initial centers we initialize both on the outer circle but one is on the right, the other is on the left. After running K-Means algorithm, we will get two cluster which contains both inner and outer semi-circle respectively. It is not what we expected.