University of Illinois Spring 2018

CS 446: Machine Learning
Homework

Due on Tuesday, April 3, 2018, 11:59 a.m. Central Time

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

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

   Solution:
   (1 point) Un-supervied. It is a clustering method.

   (b) Assume that you are trying to cluster data points x_i for $i \in \{1, 2 \dots D\}$ into K clusters each with center μ_k where $k \in \{1, 2, \dots K\}$. 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 and only one cluster at a given time?

   Solution:
   (2 points) Contraints:

   $$r_{ik} \in \{0, 1\} \quad \forall i, k$$

   $$\sum_{k=1}^{K} r_{ik} = 1 \quad \forall i$$

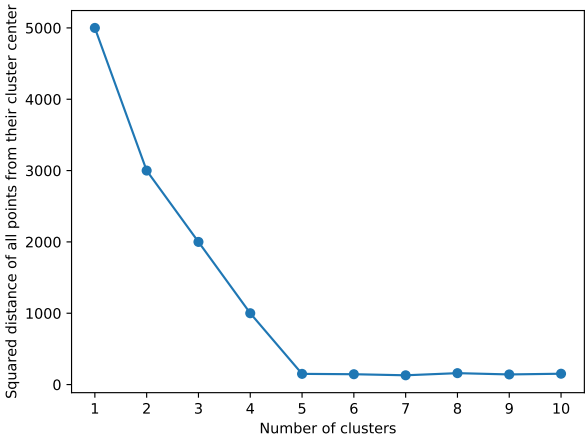
   (c) What changes must you do in your answer of part b, to make the hard assingment into a soft assignment?

   Solution:
   (2 points)

   $$r_{ik} \in [0, 1] \quad \forall i, k$$

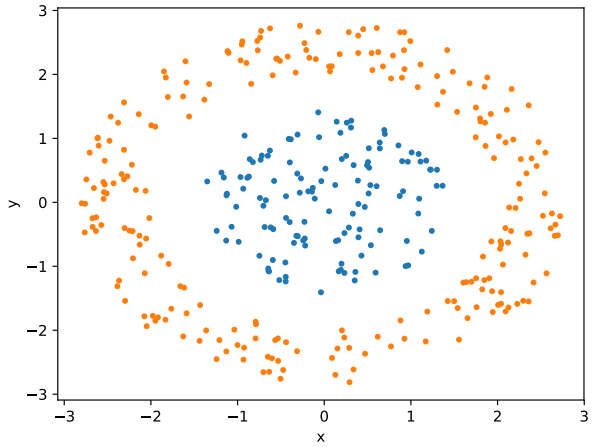
   $$\sum_{k=1}^{K} r_{ik} = 1 \quad \forall i$$

   (d) Looking at the following plot, what is the best choice for number of clusters?



Solution:
(2 point) 5

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



Solution:
(3 points)
No, K-Means will be an ineffecient algorithm for clustering this data as it would cluster the points below the 45° diagonal to one class and the points above the diagonal to a separate class.