University of Illinois Spring 2018

CS 446: Machine Learning
Homework

Due on Tuesday, April 3, 2018, 11:59 AM Central Time

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

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

   Your answer: K-means is an un-supervised method that attempts to find hidden structure in a set of data.

   (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 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:

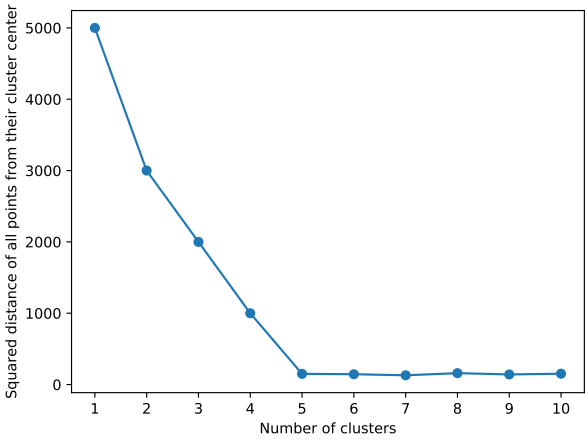
   $$r_{ik} \in \{0, 1\} \qquad \forall i \in D, k \in K$$
   $$\sum_{K} r_{ik} = 1 \qquad \forall i \in D$$

   (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:

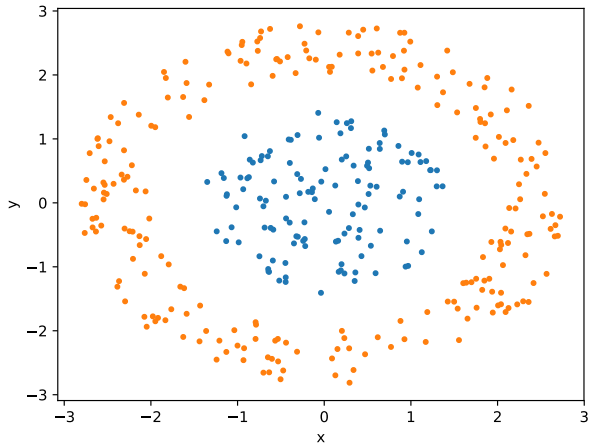
   $$r_{ik} \in [0, 1] \qquad \forall i \in D, k \in K$$
   $$\sum_{K} r_{ik} = 1 \qquad \forall i \in D$$

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



Your answer: The best number of clusters for this set of data is 5, as it has the minimum euclidean distance between cluster centers and doesn't introduce unnecessary clusters.

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



Your answer: No, K-means would not be an efficient algorithm to cluster the data. K-means attempts to find cluster centers with neat hyper-spheres around them; this is due to desire to minimize the intra-cluster sum of squares. Since the data doesn't follow this presumed shape (both clusters share the same apparent cluster center) the algorithm fails to cluster it well.