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: Unsupervised. It doesn't use any kind of label.

   (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: By adding a constraint on r_{ik}. We would want r_{ik} to be 1 for one, and only one, k for each i. So, our constraints can be written as:

   $$r_{ik} \in \{0, 1\} \quad \forall i \in \{1, 2, \cdots, D\}, \forall r \in \{1, 2, \cdots, K\} \tag{1}$$

   $$\sum_{k=1}^{K} r_{ik} = 1 \quad \forall i \in \{1, 2, \cdots, D\} \tag{2}$$

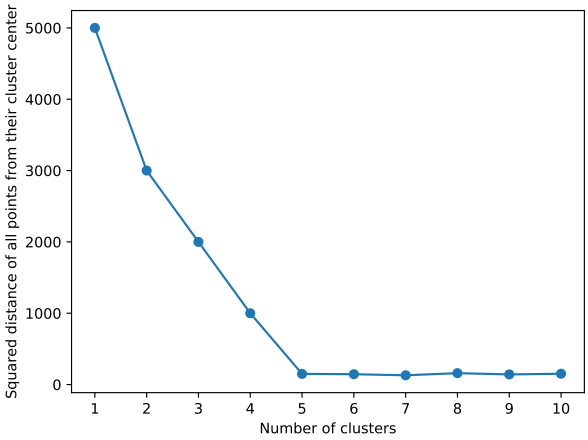
   Additional to those constraints we could also specify a constraint to ensure the active k is the one corresponding to the closest center:

   $$r_{ik} = \delta \left(k = \underset{j}{\operatorname{argmin}} ||x_i - \mu_j||^2 \right)$$

   (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: Replace the integrity constraint $r_{ik} \in \{0, 1\}$ with $r_{ik} \in [0, 1]$ and ignore the δ function.

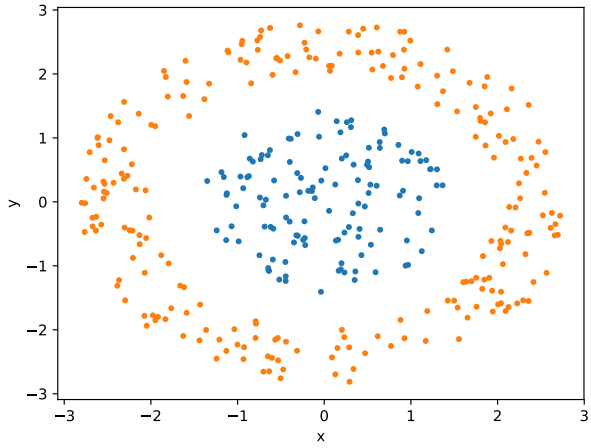
   (d) Looking at the following plot, what is the best choice for number of clusters?



Your answer: Using the Elbow method[a] we can say that $k = 5$ is the best choice because after that the cost decrease is too small.

[a] https://en.wikipedia.org/wiki/Elbow_method_(clustering)

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



Your answer: The basic k-means, using Euclidean distance, is certainly not enough to cluster the provided data. The reason for this is that k-means tries to find non-overlapping spherical clusters. This is a consequence of using Euclidean distance and not of the k-means algorithm per se. But k-means can be used with a whole family of distance functions. There are distance functions that are kernel based that can effectively identify clusters in this data. A Gaussian Kernel based distance is one of them [a], this is sometimes referred as Spectral clustering.

[a] https://sites.google.com/site/dataclusteringalgorithms/kernel-k-means-clustering-algorithm