**Machine Learning**

**CSCI 567 Spring 2021**

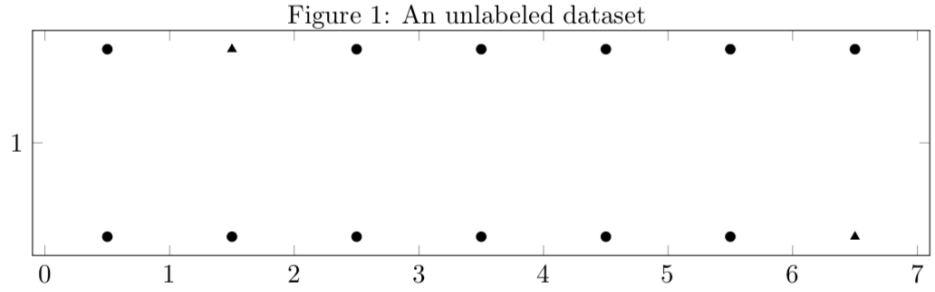
Discussion: K-means, Naive Bayes

Date: Mar 26th, 2021

1. Consider the following dataset. All points are unlabeled and part of the same set. The triangles are used to distinguish two points later.

Suppose we run the K-means algorithm on this dataset with $K = 2$ and the two points indicated by triangles as the initial centroids. When the algorithm converges, there will be two clearly separated clusters. Directly on Figure 1, draw a straight line that separates these two clusters, as well as the centroids of these two clusters.

Figure 1: An unlabeled dataset




Next, find two other different sets of initialize centroids that will converge to the exact same result if we apply K-means. Please follow the instructions below

- the initialize centroids have to be points of the dataset;
- these two sets of points can overlap with each other, but of course cannot be the same;
- similarly these sets can overlap with the initialization of Figure 1 but cannot be the same;
- do not pick those that lead to ambiguous results due to different ways of breaking ties.

2. Suppose we have the following training data. Each data point has three features (Weather, Emotion, Homework), where Weather $\in \{Sunny, Cloudy\}$, Emotion $\in \{Happy, Normal, Unhappy\}$, Homework $\in \{Much, Little\}$. The label PlayBasketball indicates whether it is suitable to play basketball. You are asked to build a naive Bayes classifier. Recall the naive Bayes assumption is

$P$(Weather, Emtion, Homework|PlayBasketball) =
$P$(Weather|PlayBasketball) $\times$ $P$(Emotion|PlayBasketball) $\times$ $P$(Homework|PlayBasketball).

(a) Write down the MLE for the following parameters (you only need to provide the final number):

| Weather | Emotion | Homework | PlayBasketball |
|---|---|---|---|
| Sunny | Happy | Little | Yes |
| Sunny | Normal | Little | Yes |
| Cloudy | Happy | Much | Yes |
| Cloudy | Unhappy | Little | Yes |
| Sunny | Unhappy | Little | No |
| Cloudy | Normal | Much | No |

- $P(\text{PlayBasketball} = No)$
- $P(\text{Weather} = Cloudy|\text{PlayBasketball} = Yes)$
- $P(\text{Emotion} = Unhappy|\text{PlayBasketball} = No)$
- $P(\text{Homework} = Little|\text{PlayBasketball} = Yes)$

(b) Given a new data instance $\boldsymbol{x} = (\text{Weather} = Cloudy, \text{Emotion} = Unhappy, \text{Homework} = Little)$, compute $P(\text{PlayBasketball} = No|\boldsymbol{x})$ (show your work).