CSCI 57800 ML Fall 2023 Homework 4

November 20, 2023

Instructions

We will be using Canvas to collect your assignments. Please read the following instructions to prepare your submission.

1. Submit your solution in a pdf file and a zip file (<yourLastName_FirstName>.pdf/zip). Your write-up must be in pdf. Your code must be in the zip file.
2. In your pdf file, the solution to each problem should start on a new page.
3. Latex is strongly encouraged to write your solutions, e.g., using Overleaf (https://www.overleaf.com/). Neither scanned handwritten copies nor hard copies are acceptable.
4. You need to add screen captures of your code and the output in your write-up.
5. You may discuss the problems and potential directions for solving them with another student. However, you need to write your own solutions and code separately, and not as a group activity. Please list the students you collaborated with on your submission.

Problem 1 (20 points)

We are given a Hidden Markov Model with the parameters below:

$S = \{N, M, V\}$ $K = \{$"*Tom*", "*Joe*", "*can*", "*will*", "*see*", "*spot*"$\}$

	π
N	0.8
M	0.1
V	0.1

Table 1: Initial probabilities in our HMM model

	N	M	V
N	0.1	0.4	0.5
M	0.3	0.1	0.6
V	0.8	0.1	0.1

Table 2: Transition probabilities in our HMM model

	Tom	Joe	can	will	see	spot
N	0.2	0.3	0.1	0.1	0.1	0.2
M	0	0	0.5	0.5	0	0
V	0	0	0.2	0.1	0.4	0.3

Table 3: Emission probabilities in our HMM model

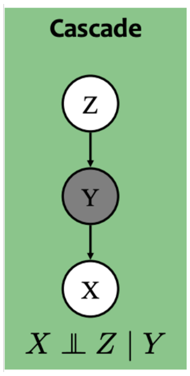
Answer the following questions.

(a) (10 pts) Based on the model, how likely a sentence "Joe can see Tom" occur? Use the Forward algorithm. Show your work.

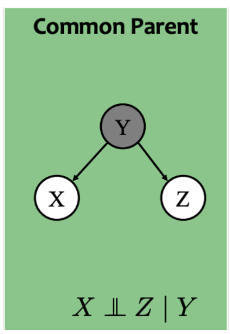
(b) (10 pts) Based on the model, find the most likely tag sequence of "will Joe spot Tom". Use the Viterbi algorithm. Show your work.

Problem 2 (40 points)

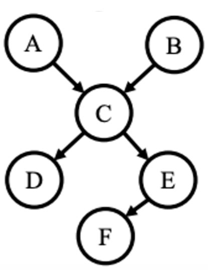
(a) (10 pts) Prove the following conditional Independence.



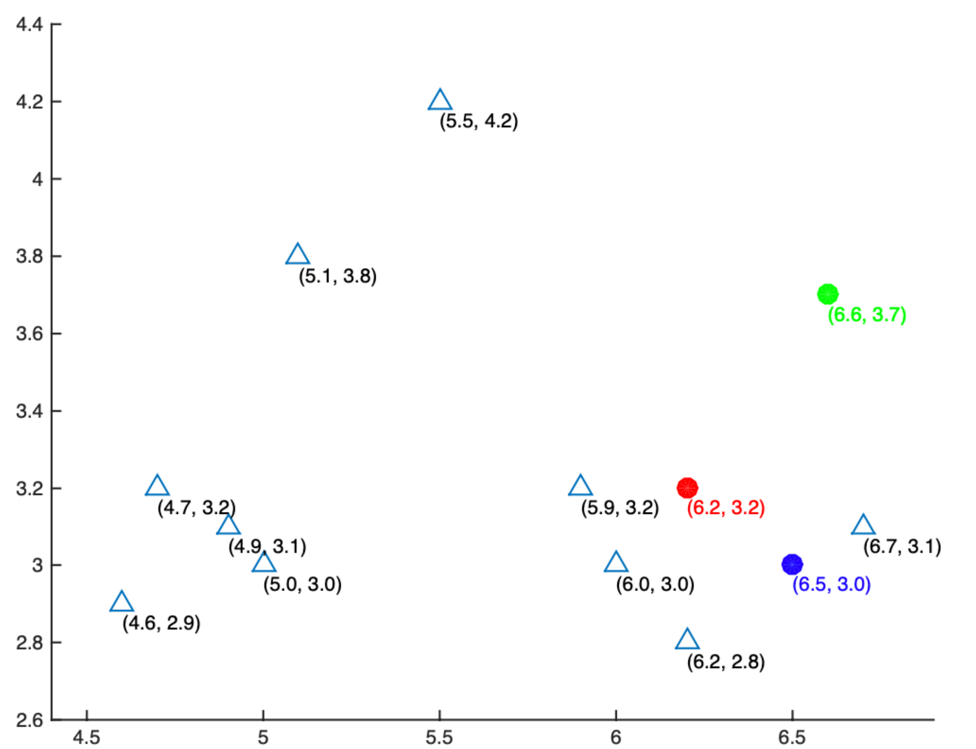
(b) (10 pts) Prove the following conditional Independence.



(c) (20 pts) Given the following graphical model, determine whether $P(E|ACD) = P(E|C)$ is True or False. Demonstrate this by proving the following two conditions:
1. Are E and A conditionally independent, given C? AND
2. Are E and D conditionally independent, given C?
From these two conditions, establish whether $P(E|ACD) = P(E|C)$ is True or False. Show your work.



Problem 3 (40 points)



Given the matrix $\mathbf{X}$ whose rows represent different data points, you are asked to perform a k-means clustering on this dataset using the Euclidean distance as the distance function. Here k is chosen as 3. The Euclidean distance d between a vector $\mathbf{x}$ and a vector $\mathbf{y}$ both in R^p is defined as $d = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$.

All data in $\mathbf{X}$ were plotted in the figure above. The centers of 3 clusters were initialized as $\mu_1 = (6.2, 3.2)$ (red), $\mu_2 = (6.6, 3.7)$ (green), $\mu_3 = (6.5, 3.0)$ (blue).

$$\mathbf{X} = \begin{bmatrix} 5.9 & 3.2 \\ 4.6 & 2.9 \\ 6.2 & 2.8 \\ 4.7 & 3.2 \\ 5.5 & 4.2 \\ 5.0 & 3.0 \\ 4.9 & 3.1 \\ 6.7 & 3.1 \\ 5.1 & 3.8 \\ 6.0 & 3.0 \end{bmatrix} \tag{1}$$

Answer the following questions. Show your work.

(a) (10 pts) What's the center of the first cluster (red) after one iteration? (Answer in the format of $[\mathbf{x1}, \mathbf{x2}]$, round your results to three decimal places, same as problems (b) and (c))

(b) (10 pts) What's the center of the second cluster (green) after two iteration?

(c) (10 pts) What's the center of the third cluster (blue) when the clustering converges?

(d) (10 pts) How many iterations are required for the clusters to converge?

Three bonus points will be given if your homework is easy to review.