Homework 5 (100 points)
CS 6375: Machine Learning

Spring 2015

Due date: April 29, 2015

1 EM algorithm [50 points]

- Download the data from the class web page.
- Implement the EM algorithm for general Gaussian mixture models (assume that the data is an array of long doubles). Use the algorithm to cluster the given data Remember that the data is 1-D. I recommend that you run the algorithm multiple times from a number of different initialization points (different θ^0 values) and pick the one that results in the highest log-likelihood (since EM in general only finds local maxima). One heuristic is to select r different randomly-chosen initialization conditions. For example, for each start, select the initial K Gaussian means by randomly selecting K initial data points, and select the initial K variances as all being some multiple of the overall data variance – the selection of initial covariances is not as critical as the initial means. Another option for initialization is to randomly assign class labels to the training data points and then calculate θ^0 based on this initial random assignment. Another option is to begin the iterations by executing a single M-step.

  Report the parameters you get for different initializations. What initialization strategy did you use? How sensitive was the performance to the initial settings of parameters.
- Now assume that variance equals 1.0 for all the three clusters and you only have to estimate the means of the three clusters using EM. Report the parameters you get for different initializations. Which approach worked better, this one or the previous one. Why? Explain your answer.

What to turn in for this part:

- Your code. EM for general GMMs and EM for GMMs with known variance.
- A report containing answers to the questions above.

2 Bayesian Networks (50 points)

Q1 (10 points) Consider the Bayesian network BN1 given below:

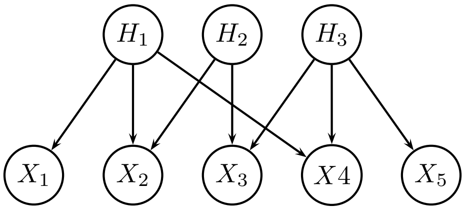

Figure 1: BN1

Assume that all variables are binary and take values from the domain $\{0,1\}$ where 0 is false and 1 is true.

- You observe that X_1 X_2 and X_5 are assigned to true (evidence). Show that marginal probability distribution over H_1, H_2, H_3 given evidence, namely$P(H_1, H_2, H_3|X_1 \wedge X_2 \wedge X_5)$ can be computed from the Bayesian network BN2 given below:

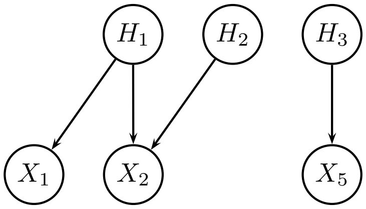

Figure 2: BN2

BN2 and BN1 have the same conditional probability tables for all the common variables. **Hint: Think about the factor that will be generated when you eliminate X_3 and X_4 before other variables**.

Q2 (20 points) Consider the Bayesian network given below:

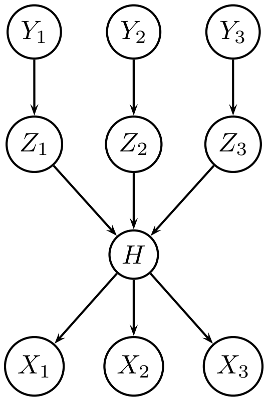

Figure 3: Bayesian network

- (10 points) Give the Bayesian network obtained by eliminating the hidden variable H. This Bayesian network represents the marginal probability distribution over the following set of variables:
$$\{Y_1, Y_2, Y_3, Z_1, Z_2, Z_3, X_1, X_2, X_3\}$$
- (10 points) Notice that you can learn the parameters of the Bayesian network containing the hidden variable H (see **Figure 3**) using the expectation maximization (EM) algorithm. On the other hand, as discussed in class, you can easily learn the parameters of the Bayesian network obtained by eliminating H in closed form. Which of the two approaches will you prefer: the EM-based approach or the closed-form solution and why? Explain your answer. No credit without proper explanation.

Q3 (10 points) Consider the Bayesian network given below:

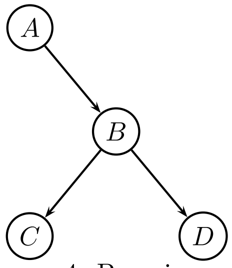

Figure 4: Bayesian network

All variables in the Bayesian network are binary and take values from the set $\{0, 1\}$. The parameters are as follows: $P(A = 1) = 0.2$, $P(B = 1|A = 1) = 0.5$, $P(B = 1|A = 0) = 0.8$, $P(C = 1|B = 1) = 0.3$, $P(C = 1|B = 0) = 0.4$, $P(D = 1|B = 1) = 0.1$ and $P(D = 1|B = 0) = 0.5$.

- Use the elimination ordering D, C, A to compute $P(B = 1)$. What is the time and space complexity of the algorithm.

Q4 (10 points)

- Consider a naive Bayes model where $X_1, \ldots, X_n$ are the attributes and Y is the class variable. Assume that all variables are binary and take values from the set $\{0, 1\}$. Suppose that D is a data set having m examples in which variable Y is hidden and variables $X_1, \ldots, X_n$ are observed in all cases. Show that if EM is applied to this problem with parameters having uniform initial values (namely all parameters are initialized to 0.5), then it will converge in one step, returning the following estimates:

$$P(Y = 1) = 0.5$$

$$P(X_i = 1|Y = 1) = P(X_i = 1|Y = 0) = \frac{D(X_i = 1)}{m} \text{ for } i = 1, \ldots, n$$

  where $D(X_i = 1)$ is the number of examples in the data set D that contain $X_i = 1$ and m is the total number of examples.

- Is the above Naive Bayes model useful if Y is always hidden.