Midterm 2: CS 6375
Fall 2020

- **NAME** _____
- **UTD-ID** _____

Question	Points	Score
Optimization: Gradient Descent and EM	40	
Learning Theory	20	
Bayesian networks: Variable Elimination	20	
Hidden Markov Models	10	
Clustering	10	
Total:	100	

Question 1: Optimization: Gradient Descent and EM (40 points)

Consider a probabilistic mixture model which is defined over three binary variables X, Y and Z where Z is the mixture variable. Assume that each variable (including Z) takes values from the domain $\{0, 1\}$. The mixture model thus has two components, one corresponding to $Z = 0$ and the other corresponding to $Z = 1$. We will assume that the model makes the following assumptions:

$$P(X, Y|Z = 0) = P(X|Z = 0)P(Y|Z = 0) \text{ and } P(X, Y|Z = 1) = P(X|Z = 1)P(Y|Z = 1)$$

and has the following parameters:

$$\begin{aligned} P(Z = 0) &= \alpha \\ P(X = 0|Z = 0) &= \beta_0 \\ P(X = 0|Z = 1) &= \beta_1 \\ P(Y = 0|Z = 0) &= \lambda_0 \\ P(Y = 0|Z = 1) &= \lambda_1 \end{aligned}$$

Assume that you are given the following dataset having m examples.

$$\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$$

Your task in this question is to derive a gradient descent algorithm for computing the maximum likelihood estimates of the parameters α, β_0, β_1, λ_0 and λ_1

(a) (10 points) Write the expression for log-likelihood of the data. (Note that Z is a hidden variable).

Solution:

$$LL = \sum_{i=1}^{m} \ln P(x^{(i)}, y^{(i)}) = \sum_{i=1}^{m} \ln \Big[P(x^{(i)}, y^{(i)}, Z = 0) + P(x^{(i)}, y^{(i)}, Z = 1)\Big]$$

where $P(x^{(i)}, y^{(i)}, Z = 0)$ is given by

$$\begin{aligned} P(x^{(i)}, y^{(i)}, Z = 0) &= P(Z = 0)P(x^{(i)}|Z = 0)P(y^{(i)}|Z = 0) \\ &= \alpha \left(\beta_0(1 - x^{(i)}) + (1 - \beta_0)x^{(i)}\right) \left(\lambda_0(1 - y^{(i)}) + (1 - \lambda_0)y^{(i)}\right) \end{aligned}$$

and $P(x^{(i)}, y^{(i)}, Z = 1)$ is given by

$$\begin{aligned} P(x^{(i)}, y^{(i)}, Z = 1) &= P(Z = 1)P(x^{(i)}|Z = 1)P(y^{(i)}|Z = 1) \\ &= (1 - \alpha) \left(\beta_1(1 - x^{(i)}) + (1 - \beta_1)x^{(i)}\right) \left(\lambda_1(1 - y^{(i)}) + (1 - \lambda_1)y^{(i)}\right) \end{aligned}$$

(b) (10 points) Compute the gradient of the log-likelihood w.r.t. the parameters α, β_0, β_1, λ_0 and λ_1 and write down the expressions below:

Gradient w.r.t. $\alpha =$

Solution:

$$\sum_{i=1}^{m} \frac{1}{P(x^{(i)}, y^{(i)})} \Big\{ \Big(\beta_0(1-x^{(i)}) + (1-\beta_0)x^{(i)}\Big) \Big(\lambda_0(1-y^{(i)}) + (1-\lambda_0)y^{(i)}\Big) - \Big(\beta_1(1-x^{(i)}) + (1-\beta_1)x^{(i)}\Big) \Big(\lambda_1(1-y^{(i)}) + (1-\lambda_1)y^{(i)}\Big) \Big\}$$

Gradient w.r.t. $\beta_0 =$

Solution:

$$\sum_{i=1}^{m} \frac{\alpha \left(1-2x^{(i)}\right) \left(\lambda_0(1-y^{(i)}) + (1-\lambda_0)y^{(i)}\right)}{P(x^{(i)}, y^{(i)})}$$

Gradient w.r.t. $\beta_1 =$

Solution:

$$\sum_{i=1}^{m} \frac{(1-\alpha) \left(1-2x^{(i)}\right) \left(\lambda_1(1-y^{(i)}) + (1-\lambda_1)y^{(i)}\right)}{P(x^{(i)}, y^{(i)})}$$

Gradient w.r.t. $\lambda_0 =$

Solution:

$$\sum_{i=1}^{m} \frac{\alpha \left(1-2y^{(i)}\right) \left(\beta_0(1-x^{(i)}) + (1-\beta_0)x^{(i)}\right)}{P(x^{(i)}, y^{(i)})}$$

Gradient w.r.t. $\lambda_1 =$

Solution:

$$\sum_{i=1}^{m} \frac{(1-\alpha) \left(1-2y^{(i)}\right) \left(\beta_1(1-x^{(i)}) + (1-\beta_1)x^{(i)}\right)}{P(x^{(i)}, y^{(i)})}$$

EM Algorithm. In this part, your task is to estimate the parameters α, β_0, β_1, λ_0 and λ_1 using the EM algorithm.

(c) (10 points) Recall that in the E-Step, we complete the data and then assign a weight (conditional probability) to each possible completion of each example. Consider an arbitrary example $(x^{(j)}, y^{(j)})$. This example can be completed in two ways: $(x^{(j)}, y^{(j)}, Z = 0)$ and $(x^{(j)}, y^{(j)}, Z = 1)$. Let f_{j0} and f_{j1} be the weights assigned to $(x^{(j)}, y^{(j)}, Z = 0)$ and $(x^{(j)}, y^{(j)}, Z = 1)$ respectively. Write down the precise expression for f_{j0} and f_{j1} such that $f_{j0} + f_{j1} = 1$ in terms of the current parameters: α, β_0, β_1, λ_0 and λ_1.

Solution:

$$g_{j0} = P(x^{(j)}, y^{(j)}, Z = 0)$$
$$g_{j1} = P(x^{(j)}, y^{(j)}, Z = 1)$$

where $P(x^{(i)}, y^{(i)}, Z = 0)$ is given by

$$\begin{aligned} P(x^{(i)}, y^{(i)}, Z = 0) &= P(Z = 0)P(x^{(i)}|Z = 0)P(y^{(i)}|Z = 0) \\ &= \alpha \left(\beta_0(1 - x^{(i)}) + (1 - \beta_0)x^{(i)}\right) \left(\lambda_0(1 - y^{(i)}) + (1 - \lambda_0)y^{(i)}\right) \end{aligned}$$

and $P(x^{(i)}, y^{(i)}, Z = 1)$ is given by

$$\begin{aligned} P(x^{(i)}, y^{(i)}, Z = 1) &= P(Z = 1)P(x^{(i)}|Z = 1)P(y^{(i)}|Z = 1) \\ &= (1 - \alpha) \left(\beta_1(1 - x^{(i)}) + (1 - \beta_1)x^{(i)}\right) \left(\lambda_1(1 - y^{(i)}) + (1 - \lambda_1)y^{(i)}\right) \end{aligned}$$

$$f_{j0} = \frac{g_{j0}}{g_{j0} + g_{j1}}$$
$$f_{j1} = \frac{g_{j1}}{g_{j0} + g_{j1}}$$

(d) (10 points) Recall that in the M-step, we update the parameters such that they maximize the log-likelihood of the weighted data. In this question, you will derive and write down the expressions for the following new or updated parameters (you can use f_{j0} and f_{j1} defined in part (a) to simplify notation):

$new\ \alpha =$

Solution:

$$\frac{\sum_{j=1}^{m} f_{j0}}{m}$$

$new\ \beta_0 =$

Solution:

$$\frac{\sum_{j=1}^{m} f_{j0}(1 - x^{(j)})}{\sum_{j=1}^{m} f_{j0}}$$

$new\ \beta_1 =$

Solution:

$$\frac{\sum_{j=1}^{m} f_{j1}(1 - x^{(j)})}{\sum_{j=1}^{m} f_{j1}}$$

$new\ \lambda_0 =$

Solution:

$$\frac{\sum_{j=1}^{m} f_{j0}(1 - y^{(j)})}{\sum_{j=1}^{m} f_{j0}}$$

$new\ \lambda_1 =$

Solution:

$$\frac{\sum_{j=1}^{m} f_{j1}(1 - y^{(j)})}{\sum_{j=1}^{m} f_{j1}}$$

Question 2: Learning Theory (20 points)

Recall that for hypothesis class H, its VC dimension is greater than or equal to k if there exists a set S of k data points, namely $|S| = k$, which is shattered by H.

(a) (5 points) Consider the class of concepts that can be expressed as an axis parallel rectangle. Assume further that the rectangles can be defined in terms of 4 integers: a, b, c, and d that lie between 1 and 100 (both included). Namely, $a, b, c, d \in \{1, \ldots, 100\}$. We assume that a consistent algorithm is available.

i. What is the size of the hypothesis space.

Solution: The hypothesis space is 100^4. If the students make a combinatorial argument (e.g., a, b, c and d are not the same points), please give them full points as long as the function increases as the same rate as 100^4.

ii. How many randomly obtained examples are needed for PAC learning with accuracy parameter $\epsilon = 0.1$ and confidence parameter $\delta = 0.05$?

Solution: Since a consistent learner is available, we use the formula:

$$m \geq \frac{1}{\epsilon}\left(\ln(1/\delta) + \ln(|H|)\right)$$

$$m \geq \frac{1}{0.1}\left(\ln(1/0.05) + \ln(100^4)\right)$$

$$m \geq 214.16$$

iii. What is the accuracy level (what is ϵ) if it is known that 1000 randomly chosen examples were used, and the desired confidence level is $\delta = 0.05$?

Solution: Rearraging the formula given above,

$$\epsilon \geq \frac{1}{m}\left(\ln(1/\delta) + \ln(|H|)\right)$$

$$\epsilon \geq \frac{1}{1000}\left(\ln(1/0.05) + \ln(100^4)\right)$$

$$\epsilon \geq 0.0214$$

(b) (5 points) Let $x \in \mathbb{R}^2$ (2-dimensional reals), H is the set of axis-parallel squares (equal height and width). Find the largest k such that the VC dimension of H is greater than or equal to k.

Solution: Here $k = 3$. Student need to prove this using 3 points.

(c) (5 points) Let $x \in \mathbb{R}$ (1-dimensional reals), H is the set of functions defined as follows: $H(x) = sign(ax^2 + bx + c)$ where a, b and c are real numbers and are parameters in H, $sign(x) = +1$ if $x \geq 0$ and $sign(x) = -1$ otherwise. Find the largest k such that the VC dimension of H is greater than or equal to k.

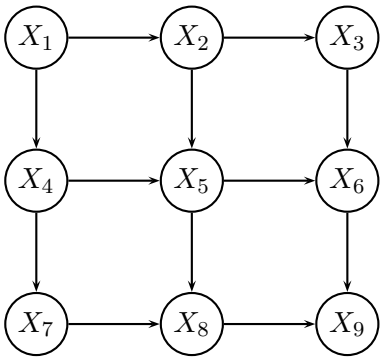
Solution: Here $k = 3$. Student need to prove this using 3 points on a line.

(d) (5 points) H is a finite hypothesis class such that $|H| < \infty$ where $|H|$ is the size of H. Show that the VC dimension of H is upper bounded by $\log_2 |H|$.

Solution: Consider a set of size $\log_2 |H| + 1$. Since each point can belong to one of two classes, the set can be labeled in $2^{(\log_2 |H|+1)} = 2|H|$ ways. For a given set of points, a particular hypothesis provides a single labeling. Therefore $|H|$ can at most only provide $|H|$ labelings for the chosen set. So it cannot shatter a set of size $\log_2 |H| + 1$. Therefore its VC dimension is bounded above by $\log_2 |H| + 1$.

Question 3: Bayesian networks: Variable Elimination (20 points)

Consider the Bayesian network given below:



(a) (7 points) Show the steps in the variable elimination algorithm for computing $P(X_9 = x_9)$ along the ordering $(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$.

Solution: The treewidth along the ordering is 3. The steps involved in VE are given below. I appears in only one CPT $P(I|F, H)$. Let the evidence instantiated CPT for it be $\psi(F, H)$.

- Elim A: $\sum_A P(A)P(B|A)P(D|A)$. This yields a function $\phi(B, D)$
- Elim B: $\sum_B P(E|B, D)P(C|B)\phi(B, D)$. This yields a function $\phi(C, D, E)$.
- Elim C: $\sum_C P(F|C, E)\phi(C, D, E)$. This yields a function $\phi(D, E, F)$
- Elim D: $\sum_D P(G|D)\phi(D, E, F)$. This yields a function $\phi(E, F, G)$
- Elim E: $\sum_E P(H|E, G)\phi(E, F, G)$. This yields a function $\phi(F, G, H)$
- Elim F: $\sum_F \psi(F, H)\phi(F, G, H)$. This yields a function $\phi(G, H)$
- Elim G: $\sum_G \phi(G, H)$. This yields a function $\phi(H)$
- Elim H: $\sum_H \phi(H)$. This equals probability of evidence.

(b) (3 points) What is the time and space complexity of variable elimination along the ordering given above? Assume that each variable has d values in its domain.

Solution: The maximum scope size of any new functions created is 3 and therefore the width of the ordering is 3. Therefore, the treewidth is at least 3. The time complexity of VE is $O(8d^4)$ and the space complexity is $O(8d^3)$. Acceptable answers include $O(d^4)$ and $O(d^3)$ for both; if the student gets the width right, give them full points.

(c) (5 points) Consider a Bayesian network defined over a set of Boolean variables $\{X_1, \ldots, X_n\}$. Assume that the number of parents of each node is bounded by a constant. Then, the probability $P(X_i = True)$ for any arbitrary variable $X_i \in \{X_1, \ldots, X_n\}$ can be computed in polynomial time (namely, in $O(n^k)$ time, where k is a constant). True or False. Explain your answer. No credit without correct explanation.

Solution: No, the problem is NP-hard. Unless P=NP, this is not possible.

(d) (5 points) Let $\{X_1, \ldots, X_r\}$ be a subset of root nodes of a Bayesian network (the nodes having no parents). Then, $P(X_1 = x_1, \ldots, X_r = x_r) = \prod_{i=1}^{r} P(X_i = x_i)$ where the notation $X_j = x_j$ denotes an assignment of value x_j to the variable X_j, $1 \leq j \leq r$. True or False. Explain your answer. No credit without correct explanation.

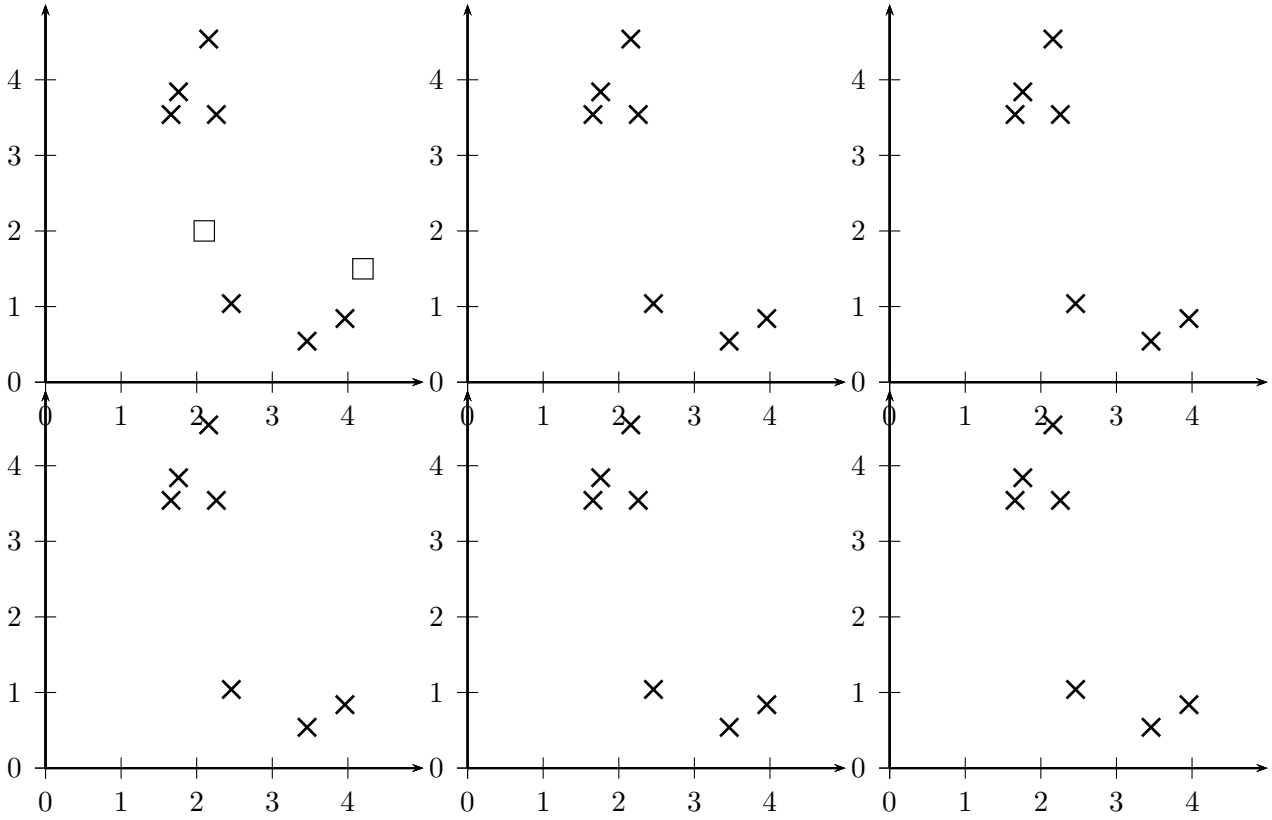
Solution: True. All roots are independent of each other. Therefore, the joint probability is just a product of individual probabilities.

Question 4: Hidden Markov Models (10 points)

(a) (10 points) Recall that a HMM makes a 1-Markov assumption, i.e., the state variable X_t is conditionally independent of $X_{1:t-2}$ given X_{t-1}. Consider a HMM that makes a 2-Markov assumption instead, i.e., the state variable X_t is conditionally independent of $X_{1:t-3}$ given $\{X_{t-1}, X_{t-2}\}$. Describe in your own words how the filtering algorithm for this 2-Markov HMM will be different from the filtering algorithm for HMMs (you don't have to provide a pseudo code). Also, compare the time and space complexity of the 2-Markov HMM filtering algorithm with 1-Markov HMM filtering algorithm. What will be the computational complexity of filtering if we make a k-Markov assumption instead, where $k > 2$?

Question 5: Clustering (10 points)

(a) (5 points) Starting with two cluster centers indicated by squares, perform k-means clustering on the following data points (denoted by $\times$). In each panel, indicate the data assignment and in the next panel show the new cluster means. Stop when converged or after 6 steps whichever comes first. Use Euclidean distance as the distance measure.



Solution: k-means will converge in two steps. In the first step, the points at the top will be assigned to the cluster center at the top and the points at the bottom will be assigned to the cluster center at the bottom. In the second step, the new cluster centers will be the mean of the points at the top and the points at the bottom respectively.

(b) (5 points) Do single link clustering and complete link clustering have the same computational complexity. Explain your answer. No credit will be given without a correct explanation.

Solution: Complete link clustering $O(n^2 \log(n))$ has higher complexity than single link clustering $O(n^2)$. The reason for this difference between single-link and complete-link is that distance defined as the distance of the two closest members is a local property that is not affected by merging; distance defined as the diameter of a cluster is a non-local property that can change during merging.