Final: CS 6375
Fall 2019

Duration: 2 hours 30 minutes

The exam is closed book (2 cheat sheets allowed). Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, use an additional sheet and staple it to your exam.

- **NAME** ______________________________

- **UTD-ID** ______________________________

Question	Points	Score
Decision Trees	10	
Poisson Naive Bayes	20	
Support Vector Machines	13	
Bayesian networks	10	
Bayesian networks: Learning	10	
VC Dimension	12	
Hidden Markov Models	10	
Short Questions	20	
Total:	105	

Question 1: Decision Trees (10 points)

(a) (4 points) Which of the following four are reasonable strategies for handling missing values at test time in a decision tree (namely after a decision tree is learned and you are using it to predict the class of a given example)? Clearly Mark the reasonable strategies (you do not have to explain your choice):

Note: Zero, one or more strategies described below may be reasonable.

1. (Majority class) Let p and n denote the number of leaf nodes in the decision tree that are labeled with positive and negative class respectively. Output positive if $p > n$ else negative.

2. Ignore the features/attributes that are missing and choose any leaf node l such that the path from the root to l is consistent with the assignment of values to the attributes that are not missing. Output the class label associated with l.

3. Let $L = \{l_1, \ldots, l_k\}$ denote the set of leaves of the decision tree such that the path from the root to each leaf $l_i \in L$ is consistent with the assignment of values to the attributes that are not missing. Output the majority class label in L.

4. Use a separate generative model to fill in the most likely values of the missing attributes and then use the decision tree to predict the class.

Solution: (3) and (4) are reasonable.

(b) (3 points) Draw a decision tree which represents the concept:

If the following CNF formula defined over 5 attributes $\{X_1, \ldots, X_5\}$ evaluates to True then the class is positive, otherwise the class is negative.

$$(X_1 \vee X_2 \vee X_3) \wedge (\neg X_1 \vee X_2 \vee X_3) \wedge (X_4 \vee X_5) \wedge (\neg X_4 \vee X_5)$$

(c) (3 points) Consider the following approach to learning decision trees. Use a learning algorithm A to learn a CNF representation from data (assume that an efficient algorithm A exists) and then you generalize the method you used in your answer to the previous question (question 1(b)) to convert the CNF to a decision tree. Explain why the above is not a good approach for learning decision trees even if an efficient algorithm A exists.

Solution: The DTree generated from a CNF can be exponential in the number of variables/clauses. CNF is a compact representation; its useless to convert it to a DTree.

Question 2: Poisson Naive Bayes (20 points)

In this question, we consider the problem of classifying a variable Y into two categories: good (A), and bad (B). We have two attributes X_1, and X_2 and assume that each attribute (X_i, $i = 1, 2$) is related to each value (A/B) of Y via a Poisson distribution with a particular mean ($\lambda_{A,i}$/$\lambda_{B,i}$). That is

$$Pr[X_i = x|Y = A] = \frac{e^{-\lambda_{A,i}}(\lambda_{A,i})^x}{x!} \quad \text{and} \quad Pr[X_i = x|Y = B] = \frac{e^{-\lambda_{B,i}}(\lambda_{B,i})^x}{x!} \text{ for } i = 1, 2$$

(a) (10 points) Derive a general expression for estimating $\lambda_{A,i}$ and $\lambda_{B,i}$ where $i \in \{1, 2\}$ from data using the maximum likelihood estimation (MLE) approach. Assume that you are given m examples: $\{(x_1^{(1)}, x_2^{(1)}, y^{(1)}), \ldots, (x_1^{(m)}, x_2^{(m)}, y^{(m)})\}$.

X_1	X_2	Y
0	3	A
4	8	A
2	4	A
6	2	B
3	5	B
2	1	B
5	4	B

Table 1: Dataset for Poisson Naive Bayes

Assume that the data given in Table 1 is generated by a Poisson Naive Bayes model.

$\Pr(Y=A) =$	$\Pr(Y=B) =$
$\lambda_{A,1} =$	$\lambda_{B,1} =$
$\lambda_{A,2} =$	$\lambda_{B,2} =$

Table 2: Parameters for Poisson Naive Bayes. Fill in the estimated values (from data) of the six parameters.

(b) (10 points) Using the expression for λs you derived in part (a) and the dataset given in Table 1, compute λs as well as the prior probabilities, namely fill in Table 2.

Question 3: Support Vector Machines (13 points)

Recall that $K(\mathbf{x}, \mathbf{y})$ is a valid kernel where $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$ if there exists a transformation $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^k$ where typically $k >= n$ such that:

$$K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y})$$

In layman's terms, a Kernel K is valid if there exists a transformation which converts two data points $\mathbf{x}$ and $\mathbf{y}$ from n dimensional space to k dimensional space such that K is equal to the dot product of the transformed data points.

(a) (7 points) Prove that $(100 + \mathbf{x}^T \mathbf{y})^2$ is a valid kernel where $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ are two dimensional, namely each data point is composed of two attributes. (Hint: Find a transformation that maps the two dimensions to six dimensions).

Consider the dataset given below (x_1, x_2 are the attributes and y is the class variable):

x_1	x_2	y
0	0	-1
$-1/3$	1	$+1$
$1/3$	1	$+1$

(b) (6 points) Give the precise primal and dual formulation for linear SVM without slack penalty for the dataset given above.

Primal Formulation (3 points):

Dual Formulation (3 points):

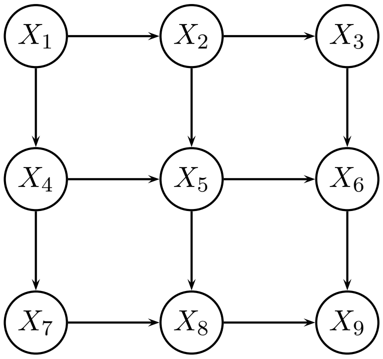
Solution: Primal problem: $L(w, \lambda) = \frac{1}{2}||w||^2 + \sum_{i=1}^{3} \lambda_i(y_i(w^T.x_i + b) - 1)$ Using data points and differentiating, we get the following equations:

$w_1 + 1/3\lambda_2 - 1/3\lambda_3 = 0$
$w_2 - \lambda_2 - \lambda_3 = 0$
$\lambda_1 - \lambda_2 - \lambda_3 = 0$

Since all three points are support vectors, we have: $y_i(w^T x_i + b) - 1 = 0$ for all i. Therefore,
$b = 1$
$-1/3w_1 + w_2 + b = -1$
$1/3w_1 + w_2 + b = -1$
From this, we see that $w_2 = 0$ and $w_i = 1/3$

Question 4: Bayesian networks (10 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.

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 $O(8 \exp(4))$ and the space complexity is $O(8 \exp(3))$.

(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.

Question 5: Bayesian networks: Learning (10 points)

Consider a Bayesian network with edges $A \to B$ and $B \to C$, and parameters which are given below:

- $P(A = 1) = 0.3$
- $P(B = 1|A = 1) = 0.6$, $P(B = 1|A = 0) = 0.2$
- $P(C = 1|B = 1) = 0.1$, $P(C = 1|B = 0) = 0.9$

Consider the dataset given below:

A	B	C
0	1	?
0	1	1
?	0	1
1	1	?
0	0	0
1	1	1

Assume that the CPTs given above are the CPTs at some iteration of EM. What you are going to do is derive the new set of parameters after running one iteration of EM.

(a) (5 points) Show the calculations involved in the E-step. Recall that in the E-step, you make the dataset bigger (by considering all possible completions) and weigh each new data point appropriately.

Solution: The possible completions are:

A	B	C	weight
0	1	0	w_1
0	1	1	w_2
0	1	1	1.0
0	0	1	w_3
1	0	1	w_4
1	1	0	w_5
1	1	1	w_6
1	0	0	w_7
1	0	1	w_8
0	0	0	1.0
1	1	1	1.0

where:

$$w_1 = (1 - 0.9)(0.6)(1 - 0.3); w_2 = (1 - 0.9)(0.6)(0.3); \text{ such that } w_1 + w_2 = 1$$

$$w_3 = (1 - 0.9)(1 - 0.6)(0.3); w_4 = (0.9)(1 - 0.1)(0.7); \text{ such that } w_3 + w_4 = 1$$

$$w_5 = (0.9)(0.1)(1 - 0.7); w_6 = (0.9)(0.1)(0.7); \text{ such that } w_5 + w_6 = 1$$

$$w_7 = (0.9)(1 - 0.1)(1 - 0.7); w_8 = (0.9)(1 - 0.1)(0.7); \text{ such that } w_7 + w_8 = 1$$

(b) (5 points) Show the calculations involved in the M-step. What are the new parameters? (Continue your answer on the next page if required; the next page is blank)

Solution:

$$P(A=1) = \frac{w_4 + w_5 + w_6 + w_7 + w_8 + 1}{7}$$

$$P(B=1|A=1) = \frac{w_5 + w_6}{w_4 + w_5 + w_6 + w_7 + w_8 + 1}$$

$$P(B=1|A=0) = \frac{w_1 + w_2 + 1}{w_1 + w_2 + 1 + w_3 + 1}$$

$$P(C=1|A=1) = \frac{w_4 + w_6 + w_8 + 1}{w_4 + w_5 + w_6 + w_7 + w_8 + 1}$$

$$P(C=1|A=0) = \frac{w_2 + 1 + w_3}{w_4 + w_5 + w_6 + w_7 + w_8 + 1}$$

(Blank page:)

Question 6: VC Dimension (12 points)

(a) (6 points) Consider the space of concentric circles in $\mathbb{R}^2$ centered at the origin. They are defined by two constants r_1 and r_2 where r_1, $r_2 \in \mathbb{R}$ and $r_1^2 \leq r_2^2$. Let H be the set of all classifiers h that classify a point (x, y) as $h(x, y) = 1$ if $r_1^2 \leq x^2 + y^2 \leq r_2^2$ and $h(x, y) = 0$ otherwise. Find the largest $d \geq 1$ such that $VC(H) \geq d$. Be sure to provide the proof. No credit if the proof is incorrect.

(b) (6 points) Consider the space of **non-intersecting** k intervals in 1-dimension. More formally, we consider **non-intersecting** intervals $[a_i, b_i]$ for $i = 1$ to k where a_i and b_i are real numbers and $a_i < b_i$. Let H be the set of all classifiers h that classify a point x as $h(x) = 1$ if x lies in any of the k intervals (namely $h(x) = 1$ if there exists an integer i such that $1 \leq i \leq k$, $x \geq a_i$ and $x \leq b_i$) and $h(x) = 0$ otherwise. Prove that $VC(H) \geq 2k$. No credit if the proof is incorrect.

Solution: It is not hard to see that any $2k$ distinct points on the real line can be shattered using k intervals: it suffices to shatter each of the k pairs of consecutive points with an interval. Assume now that $2k + 1$ distinct points $x_1 < \ldots < x_{2k+1}$ are given. For any $i \in [1, 2k + 1]$, label x_i with $(1)^{i+1}$, that is alternatively label points with 1 or 1. This leads to $k + 1$ points labeled positively and requires $2k + 1$ intervals to shatter the set since no interval can contain two consecutive points. Thus, no set of $2k + 1$ points can be shattered by k intervals and the VC dimension of the union of k intervals is $2k$.

Question 7: 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 8: Short Questions (20 points)

(a) (5 points) Construct a one dimensional classification dataset for which the Leave-one out cross validation error of the One Nearest Neighbors algorithm is always 1. Stated another way, the One Nearest Neighbor algorithm never correctly predicts the held out point. Explain your answer. No credit if the explanation is incorrect.

Solution: For this question you simply need an alternating configuration of the points $+, , +, , ...$ along the real line. In leave-one-out cross validation we compute the predicted class for each point given all the remaining points. Because the neighbors of every point are in the opposite class, the leave-one-out cross validation predictions will always be wrong.

(b) (5 points) Would we expect that running AdaBoost using the decision tree learning algorithm (without pruning) as the weak learning algorithm would have a better true error rate than running the decision tree algorithm alone (i.e., without boosting and without pruning)? Explain your answer. No credit if the explanation is incorrect.

Solution: No. Unless two differently labeled examples have the same feature vectors, ID3 will find a consistent classifier every time. In particular, after the first iteration of AdaBoost, $\epsilon_1 = 0$, so the first decision tree learned gets an infinite weight $\alpha_1 = \infty$, and the example weights $D_{t+1}(i)$ would either all become 0, all become ∞, or would remain uniform (depending on the implementation). In any case, we either halt, overflow, or make no progress, none of which helps the true error rate.

(c) (6 points) For the following classifiers, write down what happens to the bias and variance (You do not have to explain your answer)

- As k is increased in the k nearest neighbor classifier
  - The bias (Circle one) (a) Increases; (b) Decreases; (c) Remains the Same.
  - The Variance (Circle one) (a) Increases; (b) Decreases; (c) Remains the Same.
- As the number of hidden nodes in a neural network is increased while the number of layers remains the same
  - The bias (Circle one) (a) Increases; (b) Decreases; (c) Remains the Same.
  - The Variance (Circle one) (a) Increases; (b) Decreases; (c) Remains the Same.
- As the regularization constant is increased in the Logistic Regression Classifier with ℓ_1 penalty
  - The bias (Circle one) (a) Increases; (b) Decreases; (c) Remains the Same.
  - The Variance (Circle one) (a) Increases; (b) Decreases; (c) Remains the Same.

(d) (4 points) In class, we saw that k-means is a special case of hard EM. Precisely describe the modeling assumptions under which this is true. (namely, describe the probabilistic model used as well as what modification will you make to the EM algorithm).