Final: CS 6375
Spring 2018

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 (available from the instructor) and staple it to your exam.

- **NAME** _____
- **UTD-ID if known** _____

Question	Points	Score
Decision Trees	10	
Neural Networks	13	
Support Vector Machines	10	
Short Questions	22	
Bayesian networks	15	
VC Dimensions	10	
AdaBoost	10	
Hidden Markov Models	10	
Total:	100	

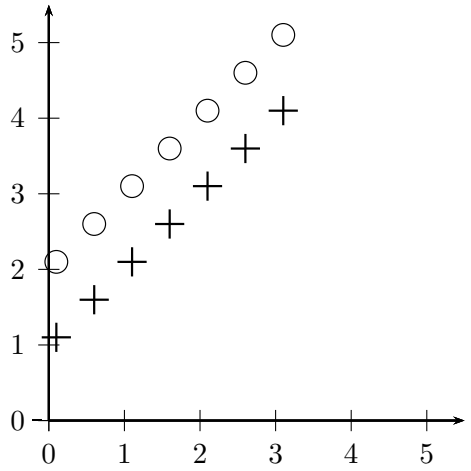
Question 1: Decision Trees (10 points)

(a) (5 points) Does the decision tree algorithm we discussed in class (the one using the information gain heuristic) guarantee a globally optimal decision tree? By optimality, we mean a decision tree that perfectly fits the training data and also has a minimal depth. Justify your answer.

Solution: False. It is a greedy algorithm and will only find a sub-optimal tree. In general inducing an optimal decision tree is a NP-complete problem.

(b) (5 points) True/False. Justify your answer. No credit if the justification/explanation is incorrect. You are given a d-dimensional linearly separable training data having n examples. The size of the decision tree for this data is guaranteed to be polynomial in d.

Solution: False. The size of the decision tree will be polynomial in n but not d. For example, consider the dataset given below.



$O(n)$ nodes will be required here. Since all splits are either horizontal or vertical, each of them will classify at most one point correctly.

Question 2: Neural Networks (13 points)

(a) (6 points) Draw a neural network having minimum number of nodes that represents the following function. Please provide a precise structure as well as a setting of weights. You can only use simple threshold units (namely, $o = +1$ if $\sum_i w_i x_i > 0$ and $o = -1$ otherwise) as hidden units and output units. X_1, X_2 and X_3 are attributes and Y is the class variable.

X_1	X_2	X_3	Y
0	0	0	+1
0	0	1	−1
0	1	0	+1
0	1	1	+1
1	0	0	+1
1	0	1	−1
1	1	0	+1
1	1	1	+1

Solution: This can be represented using a Perceptron. The concept is $X_2 \vee \neg X_3$.

(b) (4 points) True/False. Explain your answer in 1-2 sentences. Given a neural network having at least one hidden layer, the back-propagation algorithm is susceptible to initialization. Namely, the parameters returned by the algorithm will be different for different initialization strategies.

Solution: True. Backpropagation on such networks reaches a local minima and typically there are a large number of them. Therefore, different initialization strategies will likely yield different parameters.

(c) (3 points) Describe one approach to prevent over-fitting in neural networks.

Solution: Early Stopping.

Question 3: Support Vector Machines (10 points)

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	$+1$

(a) (10 points) Find the linear SVM classifier for the dataset given above. Do your optimization using the primal problem. Provide a precise setting of the weights **w** and the bias term b. What is the size of the margin?

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: Short Questions (22 points)

(a) (5 points) True/False. The SVM learning algorithm (using the dual formulation described in class) will find the globally optimal model with respect to its objective function. Explain your answer.

Solution: True. It is a quadratic optimization problem and a quadratic programming solver will solve it exactly.

(b) (5 points) Assume that you are given a Naive Bayes model defined over n binary features and one class variable having two values. Assume that at test time, k out of the n features have missing values. Then, is the following statement True or False. *The Naive Bayes model will be impractical because determining the posterior marginal probability distribution over the class variable given the observed features will require time that scales exponentially in k in the worst case*. Explain your answer.

Solution: False. As we discussed in class, when we sum out the unknown variables, we get a 1 and thus the complexity of inference (computing the posterior marginal distribution over the class) is actually linear in the number of observed features.

(c) (4 points) The bias of a k nearest neighbor classifier increases with k. True/False. Explain your answer.

Solution: True. Increasing k will decrease variance and increase bias because the functions become smoother and less susceptible to change if we add or delete data points. While decreasing k will increase variance and decrease bias

(d) (8 points) Consider a scheme to generate a series of numbers as follows: For each element in the series, first a dice is rolled. If it comes up as one of the numbers from 1 to 5, it is shown to the user. On the other hand, if the dice roll comes up as 6, then the dice is rolled for the second time, and the outcome of this roll is shown to the user.

Assume that the probability of a dice roll coming up as 6 is p. Also assume if a dice roll doesn't come up as 6, then the remaining numbers are equally likely. Suppose you see a sequence `3463661622` generated based on the scheme given above. What is the most likely value of p for this given sequence?

Solution:

$$\begin{aligned} \Pr(S) &= (p^2)^k(1-p^2)^{n-k} \\ L = \log \Pr(S) &= 2k \log p + (n-k)\log(1-p^2) \\ \therefore \frac{\partial L}{\partial p} &= \frac{2k}{p} + \frac{(n-k)(-2p)}{1-p^2} = \frac{2k - 2kp^2 - 2np^2 + 2kp^2}{p(1-p^2)} = 0 \\ \therefore 2k = 2np^2 &\Rightarrow p^2 = \frac{k}{n}, \text{ or } p = \sqrt{\frac{k}{n}}. \end{aligned}$$

Note that if we use $\Pr(S) = (p^2)^k(\frac{1-p^2}{5})^{n-k}$ instead, we have

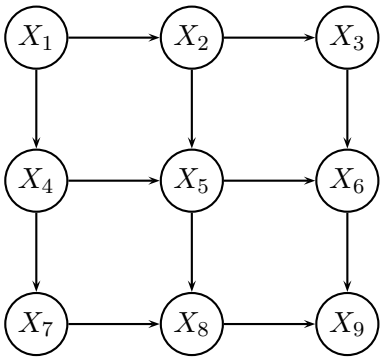
$$L = \log \Pr(S) = 2k \log p + (n-k)\log(1-p^2) - (n-k)\log 5,$$

where the last term is independent of p and does not affect our final solution. Both solutions are accepted. (What do these two different $\Pr(S)$'s represent, respectively?)

For the sequence `3463661622`, $k = 4$ and $n = 10$. Hence, the maximum likelihood estimate of $p = \sqrt{\frac{4}{10}} = 0.632$.

Question 5: Bayesian networks (15 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 $(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$. What is the time and space complexity of variable elimination along this ordering. Assume that each variable has d values in its domain.

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 $O(8\exp(3))$.

Suppose that you will learn the parameters of the grid Bayesian network given in the previous question using the Expectation-Maximization (EM) algorithm.

(b) (4 points) Assume that you are given a partially observed dataset in which variable "X_1" is always missing while the remaining variables are observed. Let n be the number of examples in the dataset and d be the number of values that each variable can take. What is the time and space complexity of the EM algorithm for this case.

Solution: As discussed in class, you can either update the sufficient statistics using the variable elimination algorithm for each example or complete each example. The complexity is the minimum of the two.

The complexity of completing the dataset is $O(2n \times 9)$ because there are two possible completions for each instance and there are nine functions which we have to multiply to compute the probability or weight for each example. Since the maximum CPT size is $O(\exp(3))$, we will need $O(9\exp(3))$ to normalize the CPTs at the end. Thus the overall time complexity is $O(18n + 9\exp(3))$.

(c) (4 points) Assume that you are given a partially observed dataset in which variables $\{X_1, X_2, X_3, X_4\}$ are always missing while the remaining variables are observed (i.e., you have complete data on X_5, X_6, X_7, X_8 and X_9). Let n be the number of examples in the dataset and d be the number of values that each variable can take. What is the time and space complexity of the EM algorithm for this case.

Question 6: VC Dimensions (10 points)

(a) (5 points) Given a hypothesis space H defined over an instance space X, prove that if there exists a subset of X of size d (for some integer d) that is shattered by H, then for any $1 \leq k < d$ there also exists a subset of size k of X that is shattered by H.

Solution: Suppose the set of points $S = \{x_1, \ldots, x_d\}$ are shattered by H. Let $J = \{x_1, \ldots, x_k\}$ be a subset of S. Consider the following labeling. All points in the set J are labeled with any label from the set $\{0, 1\}$ and all points in the set $S \setminus J$ (S setminus J) are labeled with 0. Since S is shattered by H, there exists a $h \in H$ such that above labeling is also shattered by h. In other words, J is also shattered by H.

(b) (5 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 $VC(H) \geq 2k$.

Question 7: AdaBoost (10 points)

Consider the following dataset. $(X_1, X_2) \in \mathbb{R}^2$ and Y is the class variable.

X_1	X_2	Y
0	8	$-$
1	4	$-$
3	7	$+$
-2	1	$-$
-1	13	$-$
9	11	$-$
12	7	$+$
-7	-1	$-$
-3	12	$+$
5	9	$+$

We will use two rounds of AdaBoost to learn a hypothesis for this data set. Consult the AdaBoost algorithm given on the last page of this exam. In each round m, AdaBoost chooses a weak learner that minimizes the error ϵ_m. As weak learners, use hypotheses of the form (a) $h_1 \equiv [X_1 > \theta_1]$ or (b) $h_2 \equiv [X_2 > \theta_2]$, for some integers θ_1, θ_2 (either one of the two forms, not a disjunction of the two). There should be no need to try many values of θ_1, θ_2; appropriate values should be clear from the data.

(a) (3 points) Which weak learner will AdaBoost choose in the first iteration ($m = 1$)? Be sure to provide a precise value for θ_1 or θ_2 for this learner.

Solution: For h_1, the error rates are as follows (format is θ: err): $-7.5 : 0.6; -3.5 : 0.5; -2.5 : 0.6; -1.5 : 0.5; -0.5 : 0.4; 0.5 : 0.3; 1.5 : 0.2; 3.5 : 0.3; 5.5 : 0.4; 9.5 : 0.3; 12.5 : 0.4$

For h_2, the error rates are as follows (format is θ: err): $-1.5 : 0.6; 0.5 : 0.5; 1.5 : 0.4; 4.5 : 0.3; 7.5 : 0.5; 8.5 : 0.4; 9.5 : 0.5; 11.5 : 0.4; 12.5 : 0.5; 13.5 : 0.4$

Since $h_1 \equiv [X_1 > 1.5]$ has the smallest error rate, Adaboost will choose h_1.

(I have copied the data from the previous page to this page for your convenience.)

X_1	X_2	Y
0	8	$-$
1	4	$-$
3	7	$+$
-2	1	$-$
-1	13	$-$
9	11	$-$
12	7	$+$
-7	-1	$-$
-3	12	$+$
5	9	$+$

(b) (7 points) Which weak learner will AdaBoost choose in the second iteration ($m = 2$)? Again, be sure to provide a precise value for θ_1 or θ_2 for this learner.

Question 8: 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 $k > 2$?