Final: CS 7301
Spring 2017

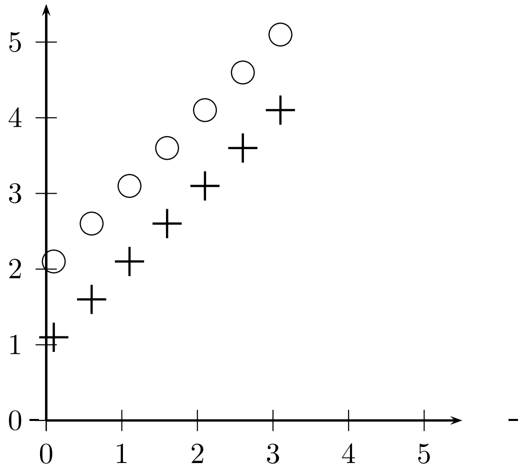
The exam is closed book (2 cheat sheets allowed). Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, use an additional sheet (available from the instructor) and staple it to your exam.

- **NAME** ____
- **UTD-ID if known** ___________________________________

Question	Points	Score
Decision Trees	10	
Neural Networks	10	
Support Vector Machines	10	
Short Questions	20	
Bayesian networks	15	
VC Dimensions	10	
AdaBoost	10	
Hidden Markov Models	10	
Regression	5	
Total:	100	

Question 1: Decision Trees (10 points)

Consider a large dataset D having n examples in which the positive (denoted by $+$) and negative examples (denoted by $\circ$) follow the pattern given below. (Notice that the data is clearly linearly separable).



(a) (5 points) Which among the following is the "best upper bound" (namely the smallest one that is a valid upper bound) on the number of leaves in an optimal decision tree for D (n is the number of examples in D)? By optimal, I mean a decision tree having the smallest number of nodes. Circle the answer and explain why it is the best upper bound. No credit without a correct explanation.

1. $O(n)$

2. $O(\log n)$

3. $O(\log \log n)$

4. $O((\log n)^2)$

Consider the dataset given below. X_1, X_2, X_3 and X_4 are the attributes (or features) and Y is the class variable.

X_1	X_2	X_3	X_4	Y
3	0	0	1	$+$
1	1	0	0	$-$
2	0	1	1	$-$
5	1	1	0	$+$
4	1	0	1	$+$
6	0	1	0	$-$

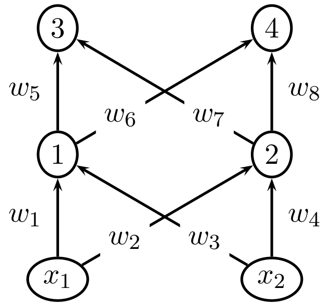
(b) (2 points) Which attribute (among X_1, X_2, X_3 and X_4) has the highest information gain?

(c) (3 points) In the above dataset, is the attribute having the highest information gain useful (namely will it help improve generalization)? Answer YES/NO and then

- Explain why the attribute is useful if your answer is “YES.”
- If your answer is “NO”, explain how will you change the information gain criteria so that such useless attributes are not selected.

Question 2: Neural Networks (10 points)

Consider the Neural network given below.



Assume that all internal nodes and output nodes compute the sigmoid $\sigma(t)$ function. In this question, we will derive an explicit expression that shows how back propagation (applied to minimize the least squares error function) changes the values of w_1, w_2, w_3, w_4, w_5, w_6, w_7 and w_8 when the algorithm is given the example (x_1, x_2, y_1, y_2) with y_1 and y_2 being outputs at 3 and 4 respectively (there are no bias terms). Assume that the learning rate is η. Let o_1 and o_2 be the output of the hidden units 1 and 2 respectively. Let o_3 and o_4 be the output of the output units 3 and 4 respectively.

Hint: Derivative: $\frac{d}{dt}\sigma(t) = \sigma(t)(1 - \sigma(t))$.

(a) (2 points) Forward propagation. Write equations for o_1, o_2, o_3 and o_4.

(b) (4 points) Backward propagation. Write equations for δ_1, δ_2, δ_3 and δ_4 where δ_1, δ_2 , δ_3 and δ_4 are the values propagated backwards by the units denoted by 1, 2, 3 and 4 respectively in the neural network.

(c) (4 points) Give an explicit expression for the new (updated) weights w_1, w_2, w_3, w_4, w_5, w_6, w_7 and w_8 after backward propagation.

Question 3: Support Vector Machines (10 points)

Consider the following 2-D dataset (x_1 and x_2 are the attributes and y is the class variable).

Dataset:

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

(a) (5 points) Precisely write the expression for the dual problem. Let $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ be the lagrangian multipliers associated with the four data points.

(b) (5 points) Identify the support vectors and compute the value of α_1, α_2, α_3 and α_4.

Question 4: Short Questions (20 points)

(a) (5 points) Circle YES or NO. You don't have to explain your answer.
Which of the following models can learn a non-linear concept? x denotes the attributes while a, v and w are the weights. σ denotes the sigmoid activation function, namely $\sigma(t) = \frac{1}{1+e^{-t}}$.

- $\sigma(\sum_{i=1}^{n} w_i x_i + \sum_{j=1}^{n} v_j x_j)$ YES NO
- $\sigma(\sum_{i=1}^{n} w_i x_i)$ YES NO
- $a_1\sigma(\sum_{i=1}^{n} w_i x_i) + a_2\sigma(\sum_{j=1}^{n} v_j x_j)$ YES NO
- $\sigma(a_1\sigma(\sum_{i=1}^{n} w_i x_i) + a_2\sigma(\sum_{j=1}^{n} v_j x_j))$ YES NO
- $\sigma(\sum_{i=1}^{n} w_i x_i + \sum_{j=1}^{n}\sum_{k=1}^{n} v_{i,j} x_i x_j)$ YES NO

(b) (5 points) When do you expect the learning algorithm for the logistic regression classifier to produce the same parameters as the ones produced by the learning algorithm for the Gaussian Naive Bayes model with class independent variances.

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

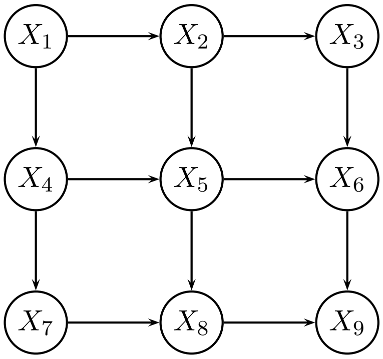
(d) (6 points) You are given a coin and a thumbtack and you put Beta priors $Beta(5, 5)$ and $Beta(20, 20)$ on the coin and thumbtack respectively. You perform the following experiment: toss both the thumbtack and the coin 100 times. To your surprise, you get 20 heads and 80 tails for both the coin and the thumbtack. Are each of the following two statements true or false.

- The MLE estimate of both the coin and the thumbtack is the same.
- The MAP estimate of the parameter θ (probability of landing heads) for the coin is greater than the MAP estimate of θ for the thumbtack.

Explain your answer mathematically.

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.

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.

(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.

(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.

(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 $k > 2$?

Question 9: Regression (5 points)

Consider a linear regression problem $y = w_1x + w_2z + w_0$, with a training set having m examples $(x_1, z_1, y_1), \ldots, (x_m, z_m, y_m)$. Suppose that we wish to minimize squared error (loss function) given by:

$$Loss = \sum_{i=1}^{m} (y_i - w_1x_i - w_2z_i - w_0)^2$$

under the assumption $w_1 = w_0$.

(a) (5 points) Derive a batch gradient descent algorithm that minimizes the loss function. (Note the assumption $w_0 = w_1$; also known as parameter tying).

AdaBoost

1. Initialize the data weighting coefficients $\{w_n\}$ by setting $w_n^{(1)} = 1/N$ for $n = 1, \ldots, N$.
2. For $m = 1, \ldots, M$:
   (a) Fit a classifier $y_m(\mathbf{x})$ to the training data by minimizing the weighted error function

$$J_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n) \qquad (14.15)$$

   where $I(y_m(\mathbf{x}_n) \neq t_n)$ is the indicator function and equals 1 when $y_m(\mathbf{x}_n) \neq t_n$ and 0 otherwise.

   (b) Evaluate the quantities

$$\epsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n)}{\sum_{n=1}^{N} w_n^{(m)}} \qquad (14.16)$$

   and then use these to evaluate

$$\alpha_m = \ln\left\{\frac{1-\epsilon_m}{\epsilon_m}\right\}. \qquad (14.17)$$

   (c) Update the data weighting coefficients

$$w_n^{(m+1)} = w_n^{(m)} \exp\left\{\alpha_m I(y_m(\mathbf{x}_n) \neq t_n)\right\} \qquad (14.18)$$

3. Make predictions using the final model, which is given by

$$Y_M(\mathbf{x}) = \operatorname{sign}\left(\sum_{m=1}^{M} \alpha_m y_m(\mathbf{x})\right). \qquad (14.19)$$