Qualifier: CS 6375
Machine Learning
Spring 2015

The exam is closed book. You are allowed to use two double-sided cheat sheets and a calculator. If you run out of room for an answer, use an additional sheet (available from the instructor) and staple it to your exam.

- **Time:** 2 hours 30 minutes.

Question	Points	Score
Support Vector Machines	15	
Bayesian networks: Representation	10	
Bayesian networks: Inference	16	
Naive Bayes	15	
Learning Theory	12	
AdaBoost	12	
Short Questions	20	
Total:	100	

Question 1: Support Vector Machines (15 points)

(a) (5 points) Consider a 2-D dataset that is separable by an axis-aligned ellipse (namely there exists an axis-aligned ellipse that has 100% accuracy on the dataset). Show that this dataset is linearly separable in the 6-D feature space $(1, x_1, x_2, x_1^2, x_2^2, x_1 x_2)$ where x_1, x_2 denote the attributes in the 2-D space.

Hint: Recall that the equation of a axis-aligned ellipse is $c(x_1 - a)^2 + d(x_2 - b)^2 = 1$ where a, b, c, d are real-numbers.

Consider the dataset given below (x_1, x_2, x_3 are the attributes and y is the class variable):

x_1	x_2	x_3	y
1	0	0	$+1$
1	$-\sqrt{2}$	1	-1
1	$\sqrt{2}$	1	-1

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

Question 2: Bayesian networks: Representation (10 points)

(a) (10 points) Consider the formula $F = (X_1 \vee X_2) \wedge (\neg X_3 \vee X_2)$. Construct a Bayesian network over the Boolean variables X_1, X_2, X_3 that represents the following function.

- If the assignment $(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ evaluates the formula F to True, then $P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = 1/n$ where n is the number of solutions of F. Notice that $n = 5$.
- If the assignment $(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ evaluates F to False, then $P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = 0$.

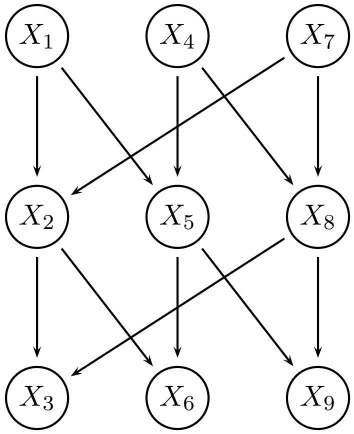
Your Bayesian network should have the minimal number of edges to get full credit. Moreover, to get full credit, you should provide the precise structure of the network (namely the directed acyclic graph) and the parameters (the CPTs).

Question 3: Bayesian networks: Inference (16 points)

(a) (3 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.

(b) (3 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.

Consider the Bayesian network given below.



For the following two questions, assume that all variables in the network are Binary. Namely, they take values from the set $\{True, False\}$. Further assume that all conditional probabilities tables are such that $P(X_i = True|parents(X_i)) = 0.8$ where $1 \leq i \leq 9$, and $parents(X_i)$ is any truth-assignment to all parents of X_i (note that for the root nodes, the parent set is empty and therefore $P(X_1 = True) = P(X_4 = True) = P(X_7 = True) = 0.8$).

(c) (5 points) What is the probability that X_1 is True given that X_3 and X_6 are True? Explain your answer.

(d) (5 points) What is the probability that X_3 is True given that X_4 and X_6 are True? Explain your answer.

Question 4: Naive Bayes (15 points)

In this problem, we will train a probabilistic model to classify a document written in a simplified language. However, instead of training the naive Bayes model with multivariate Bernoulli distributions, we will train it with a model that uses multinominal distributions.

Assume that all the documents are written in a language which has only three words a, b and c. All the documents have exactly n words (each word can be either a, b or c). We are given a labeled document collection $\{D_1, D_2, \ldots, D_m\}$. The label y_i of document D_i is 1 or 0, indicating whether D_i is "good" or "bad." This model uses the multinominal distributions in the following way. Given the i-th document D_i, we denote by a_i (b_i, c_i, respectively) the number of times that word a (b, c, respectively) appears in D_i. Therefore, $a_i + b_i + c_i = |D_i| = n$. In this model, we define

$$P(y_i = 1) = \eta$$

$$P(D_i|y_i = 1) = \frac{n!}{a_i!b_i!c_i!}\alpha_1^{a_i}\beta_1^{b_i}\gamma_1^{c_i}$$

where α_1 (respectively β_1, γ_1) is the probability that word a (respectively b, c) appears in a "good" document and $\alpha_1 + \beta_1 + \gamma_1 = 1$. Similarly,

$$P(D_i|y_i = 0) = \frac{n!}{a_i!b_i!c_i!}\alpha_0^{a_i}\beta_0^{b_i}\gamma_0^{c_i}$$

where α_0 (respectively β_0, γ_0) is the probability that word a (respectively b, c) appears in a "bad" document and $\alpha_0 + \beta_0 + \gamma_0 = 1$.

(a) (2 points) Given a document D_i, we want to classify it using $P(y_i|D_i)$. Write down an expression for $P(y_i|D_i)$ and explain what parameters are to be estimated from data in order to calculate $P(y_i|D_i)$.

(b) (13 points) Derive the most likely value of the parameters. Namely, write the expression for likelihood of the data and derive a closed form expression for the parameters.

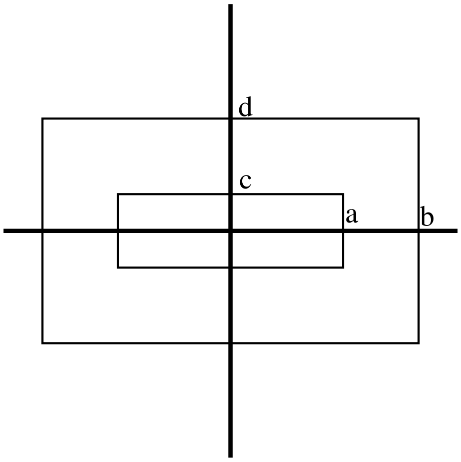
[Hint: You may have to use Lagrange multipliers because you have to maximize the log-likelihood under the constraint that α_1, β_1 and γ_1 (similarly α_0, β_0 and γ_0) sum to one.]

Derive your answer on the next page.

(Write your answer here.)

Question 5: Learning Theory (12 points)

(a) (6 points) Consider a concept space H of axis-parallel origin-centered embedded rectangles (see the figure below). Formally, a concept $h \in H$ is defined by four non-negative real parameters $a, b, c, d \in R^+$, such that $a < b$ and $c < d$. An example (x, y) is labeled positive if and only if (x, y) is in the rectangle $-b < x < b$ and $-d < y < d$, but not in the rectangle $-a < x < a$ and $-c < y < c$. Is the VC-dimension of H greater than or equal to 4? Explain your answer. No credit without correct explanation.



(b) (6 points) What is the VC-dimension of the following concept defined over 2-dimensions x_1, x_2.

- If $(\alpha x_1^2 + \alpha x_2^2 + \beta) > 0$ then class is positive. Otherwise, the class is negative. Here, $\alpha, \beta \in \mathbb{R}$ are real numbers.

Explain your answer. No credit without correct explanation.

Question 6: AdaBoost (12 points)

Consult the AdaBoost algorithm given on the last page for this question. Suppose you have two weak learners, h_0 and h_1, and a set of 17 points.

(a) (2 points) You find that h_1 makes one mistake and h_2 makes four mistakes on the dataset. Which learner will AdaBoost choose in the first iteration (namely $m = 1$)? Justify your answer.

(b) (2 points) What is α_1?

(c) (2 points) Calculate the data weighting co-efficients w_2 for the following two cases: (1) the points on which the chosen learner made a mistake and (2) the points on which the chosen learner did not make a mistake.

(d) (6 points) Consider a simple modification to the AdaBoost algorithm in which we normalize the data weighting co-efficients. Namely, we replace $w_n^{(m+1)}$ by $w_n^{(m+1)}/Z^{(m+1)}$ where $Z^{(m+1)} = \sum_{n=1}^{N} w_n^{(m+1)}$. Prove that $Z^{(m+1)} = 2(1-\epsilon_m)$.

Hint: Notice that if the weights are normalized, then $\epsilon_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n)$.

Question 7: Short Questions (20 points)

(a) (4 points) Consider a 2-D dataset that is linearly separable. Your Boss tells you to use ID3 (a decision tree classifier) instead of a linear classifier (e.g. SVMs). Make an argument based on the representation size that this is not a good idea.

(b) (5 points) Let p be the probability of a coin landing heads up when tossed. You flip the coin 3 times and observe 2 tails and 1 head. Suppose p can only take two values: 0.3 or 0.6. Find the Maximum likelihood estimate of p over the set of possible values $\{0.3, 0.6\}$.

(c) (5 points) Suppose that you have the following prior on the parameter p: $P(p = 0.3) = 0.3$ and $P(p = 0.6) = 0.7$. Given that you flipped the coin 3 times with the observations described above, find the MAP estimate of p over the set $\{0.3, 0.6\}$, using the prior.

(d) (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	$+1$

AdaBoost

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

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

   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

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

   and then use these to evaluate

$$\alpha_m = \ln\left\{\frac{1-\epsilon_m}{\epsilon_m}\right\}. \tag{14.17}$$

   (c) Update the data weighting coefficients

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

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

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