Computer Science & Engineering Department, IIT Kharagpur

CS60050 Machine Learning
Midterm Examination, Spring 2013

Time: 2 hours Full Marks: 50

1. Consider the following concept learning scenario. There are n boolean valued features: $x_1, x_2, \ldots x_n$, and the targer concept is Y. The hypothesis space is conjunction of positive and negative literals. [6]

   (a) What is the size of the hypothesis space?

   (b) Suppose the training data has the following positive instances:

x_1	x_2	x_3	x_4	x_5	Y
T	T	T	F	T	T
T	F	T	T	T	T
T	F	T	F	T	T
T	F	F	T	T	T

   i. What is the size of the version space?

   ii. Specify the G-set (the set of most general hypotheses) and the S-set (the set of the most specific hypothesis) of the version space

2. Suppose that you have a 2-dimensional real-valued feature space and two classes, and you apply nearest neighbour classifier (NN). [6]

   (a) Is it possible for a 2-class 1-NN classifier to always classify all new examples as positive even though there are negative examples in the training data? If yes, show an example. If no, briefly explain.

   (b) Is it possible for a 2-class 3-NN classifier to always classify all new examples as positive even though there are negative examples in the training data? If yes, show an example. If no, briefly explain.

   (c) Draw a set of 6 instances (3 positive and 3 negative) so that the "eave one out cross-validation" error when using a 1-NN classifier is always 1, that is, every point is misclassified.

3. Consider the problem of diagnosing avian influenza in a patient, given the following binary-valued (i.e., true or false) attributes: [10]

   - Fever = patient has fever
   - Cough = patient has cough
   - Breathing = patient has difficulty breathing
   - HumanContact = patient had recent contact with human infected with avian influenza
   - AvianContact = patient had recent contact with bird infected with avian influenza

Consider the training set S given below. In each of the six training examples, true or false values for each of the five attributes are given, as well as the correct classification for each example.

Patient	Fever	Cough	Breathing	HumContact	AvContact	AvInfluenza
P1	T	T	F	F	F	F
P2	T	T	T	T	F	T
P3	F	F	T	F	T	F
P4	F	T	T	F	T	T
P5	T	T	F	F	T	T
P6	T	F	F	T	T	F

(a) Give the information gain, $Gain(S, A)$, for each attribute A with respect to the training set S. Show your work clearly

(b) Find a decision tree that ID3 would return, by tracing the steps of ID3 by hand, using information gain as the splitting criterion. (If there is a tie for highest information gain, choose one of the highest-gain attributes at random.) Include your work that shows how you traced the steps of ID3. Verify that your resulting tree is consistent with the training data.

(c) Does pruning a decision tree such as that produced by the basic ID3 algorithm increase or decrease performance on the training set? sometimes or always? Does it increase or decrease performance on the test set? sometimes or always?

Some approximate values of $\log_2$: $\log_2(3) = 1.585$, $\log_2(5) = 2.322$, $\log(2(7) = 2.807$.

4. (a) Design an n-input perceptron that implements the function: k or more of the inputs are true. [6]

(b) If you add more hidden layers to a feedforward neural network, can you always improve performance on both your training and test sets? Explain.

(c) For a fixed network architecture, is the backpropagation algorithm guaranteed to find the best possible set of weights, given sufficient training data? Explain.

5. Assume we have a set of data from patients who have visited BCR hospital during the year 2011. A set of features have been also extracted for each patient. Our goal is to decide whether a new visiting patient has any of diabetes, heart disease, or kidney disease, or Alzheimers (a patient can have one or more of these diseases). [6]

(a) You decide to use a neural network to solve this problem. You have two choices: either to train a separate neural network for each of the diseases or to train a single neural network with one output neuron for each disease, but with a shared hidden layer. Under what assumption will a shared hidden layer network be more effective, and when will independent network be more effective?

(b) Some patient features are expensive to collect (e.g., brain scans) whereas others are not (e.g., temperature). Therefore, we have decided to first ask our classification algorithm to predict whether a patient has a disease, and if the classifier is 80% confident that the patient has a disease, then we will do additional examinations to collect additional patient features In this case, which classification methods do you recommend: neural networks, decision tree, or naive Bayes? Briefly outline a suitable method using one of the above methods and justify your method in one or two sentences.

6. Suppose you are given hypothesis $h1$ and sample $S1$ of size 100, drawn from unknown distribution D. You find that h1 misclassifies 20 of the examples in $S1$. [4]

   (a) Give the approximate standard deviation of $errorS(h1)$ over samples S of size 100.

   (b) Use the results from part (a) to give the upper bound U of a one-sided confidence interval such that $errorD(h1) \leq U$ with 99% confidence.

   The following table gives the multiples z_N of σ such that X will lie in the range $\mu \pm z_N\sigma$ with a specified probability p.

Confidence level N%	80%	90%	95%	98%	99%	99.5%
Constant z_N	1.28	1.64	1.96	2.33	2.58	2.81

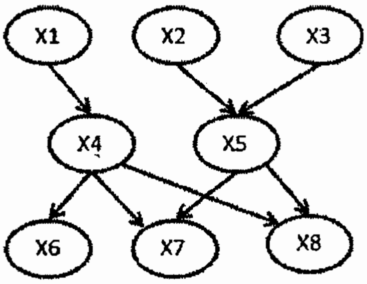
7. We have a training set consisting of samples and their labels. All samples come from one of two classes, 0 and 1. Samples are two dimensional vectors. The input data is the form $\{X1, X2, Y\}$ where$X1$ and $X2$ are the two values for the input vector and Y is the label for this sample. After learning the parameters of a Naive Bayes classifier we arrived at the following table: [4]

	$Y = 0$	$Y = 1$
$X1$	$P(X1 = 1\|Y = 0) = 1/5$	$P(X1 = 1\|Y = 1) = 3/8$
$X1$	$P(X2 = 1\|Y = 0) = 1/3$	$P(X2 = 1\|Y = 1) = 3/4$

   Denote by w_1 the probability of class 1 (that is $w_1 = P(Y = 1)$). If we know that the likelihood of the following two samples: $\{1, 0, 1\}$, $\{0, 1, 0\}$ given our Naive Bayes model is $1/180$, what is the value of w_1? You do not need to derive an explicit value for w_1. You may write an equation that has w_1 as the only unknown and that when solved would provide the value of w_1. You may then derive an explicit value for w_1.

8. (a) Consider a classification problem with two classes and eight binary attributes, $X_1, X_2, \ldots X_8$. How many parameters would you need to learn with a Naive Bayes classifier? How many parameters would you need to learn with a Bayes optimal classifier? What are they? [8]

   (b) Consider the Bayesian network for these features.



   Write the joint probability $P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$ factored according to the Bayes net. How many parameters are necessary to define the conditional probability distributions for this Bayesian network?