
CS 760 Homework 5:

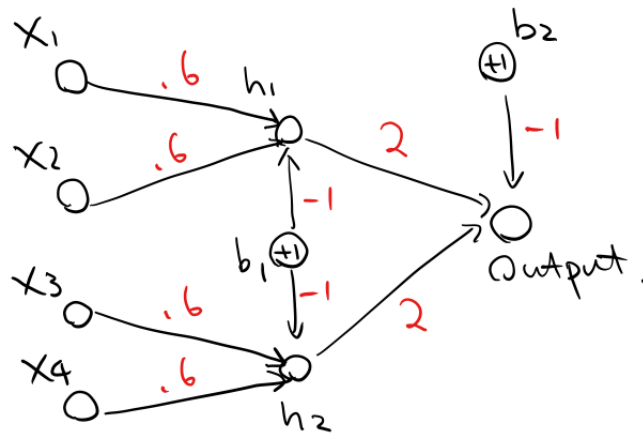
Qihong Lu

December 6, 2016

QUESTION 1

Show a neural network that represents the logical function $y = (x_1 \wedge x_2) \vee (x_3 \wedge x_4)$. Specifically, show the network topology, weights and biases. You should assume that hidden and output units use sigmoid output functions, and an output-unit activation of 0.5 or greater represents a true prediction for y .

The network architecture is shown below. The weights are denoted in black. X_i are the input units. h_i are the hidden units. b_i are the bias units. Finally, there is a single output unit.



The unit h_1 is representing an AND function over x_1 and x_2 . Similarly, the unit h_2 is representing an AND function over x_3 and x_4 . Finally, the output unit represents an OR function over h_1 and h_2 .

QUESTION2

Consider the concept class C in which each concept is an interval on the line of real numbers. Each training instance is represented by a single real-valued feature x , and a binary class label $y \in \{0, 1\}$. A learned concept is represented by an interval $[a, a + b]$ where a is a real value and b is a positive real value, and the concept predicts $y=1$ for values of x in the interval, and $y=0$ otherwise. Show that C is PAC learnable.

QUESTION3

Consider the concept class that consists of disjunctions of exactly two literals where each literal is a feature or its negation, and at most one literal can be negated. Suppose that the number of features $n = 3$. Show what the Halving algorithm would do with the following two training instances in an on-line setting. Specifically, show the initial version space, the prediction made by the Halving algorithm for each instance, and the resulting version space after receiving the label of each instance.

x1	x2	x3	y
T	F	F	pos
F	T	T	neg

QUESTION4

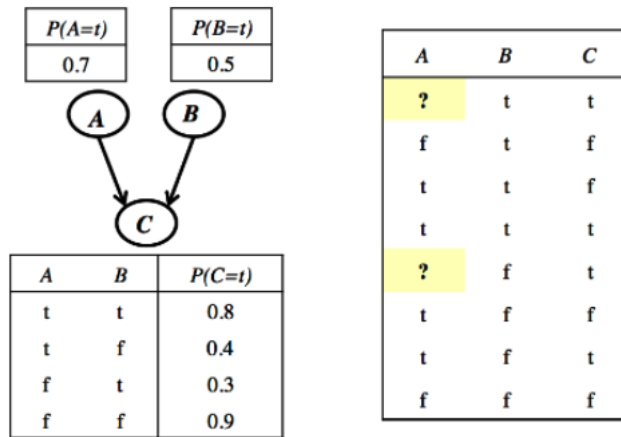
In this same setting, suppose the learner can pick the next training instance it will be given. That is, the learner can pick the feature vector part of the instance; the class label will be provided by the teacher. Which instance should it ask for next? Justify your answer.

QUESTION5

How many mistakes will the Halving algorithm make for this concept class in the worst case? Justify your answer.

QUESTION6

Given the initial Bayes net parameters and training set depicted below, show how the network parameters would be updated after one step of the EM procedure. The '?' symbol indicates that the value for the variable A is missing in a given training instance.



QUESTION7

Consider a learning task in which you are given n features and you want to use a feature selection method along with your learning algorithm. Specifically, suppose you are using forward selection along with k -fold cross validation to evaluate each feature set during the search process. Assume that there are r relevant features, and forward selection stops after selecting r features. Given a single training set, how many models are learned in the process of finding a feature subset of size r ?

For forward selection, we begin with the null model (the model with no feature) and add one informative feature at each iteration, where the informativeness of a given feature is determined based on the cross-validated performance.

At the 1st iteration, there are n candidate features, I need to evaluate n models.

At the 2nd iteration, there are $n-1$ candidate features, I need to evaluate $n-1$ models.

.....

At the r -th iteration, there are $n-r$ candidate features, I need to evaluate $n-r$ models. ($r \leq n$)

Therefore, to chose r features, we need to evaluate $n(n-1) \cdots (n-r)$ models.

When evaluating each model, a k -folds cross validation is required. So in total, we need to fit $kn(n-1) \cdots (n-r)$ models.

QUESTION8

Now suppose instead you are using backward elimination for the same task. Again, assume that the search process stops after selecting $>r$ features and does not consider feature subsets smaller than this. Given a single training set, how many models are learned in the process of finding a feature subset of size r ?

For backward elimination, we begin with the full model (the model with all features) and remove the least informative feature at each iteration, where the informativeness of a given feature is determined based on the cross-validated performance.

1st iteration: there are n possible features can be removed, need to evaluate n models.

2nd iteration: there are $n-1$ possible features can be removed, need to evaluate $n-1$ models.

.....

r -th iteration: there are $n-r$ possible features can be removed, need to evaluate $n-r$ models.
($r \leq n$)

Therefore, to chose r features, we need to evaluate $n(n-1) \cdots (n-r)$ models.

When evaluating each model, a k -folds cross validation is required. So in total, we need to fit $kn(n-1) \cdots (n-r)$ models.

QUESTION9

Consider the relational learning task defined below. List all of the literals that would be considered by FOIL algorithm on the first step of learning a rule for the aunt(X, Y) relation.

constants y, o, a, b, m, s, t, d

target relation

$$\text{aunt}(X, Y) \quad \oplus = \left\{ \langle o, t \rangle, \langle y, t \rangle, \right. \\ \left. \langle a, s \rangle, \langle b, s \rangle \right\}$$

$$\ominus = \left\{ \langle o, d \rangle, \langle y, d \rangle, \right. \\ \left. \langle b, m \rangle \right\}$$

aunt(X, Y) relation.

background relations

$$\text{sibling}(X, Y) = \left\{ \begin{array}{l} \langle s, t \rangle, \langle t, s \rangle, \\ \langle d, s \rangle, \langle s, d \rangle, \\ \langle t, d \rangle, \langle d, t \rangle, \\ \langle o, y \rangle, \langle y, o \rangle, \\ \langle a, b \rangle, \langle b, a \rangle \end{array} \right\} \quad \text{parent}(X, Y) = \left\{ \begin{array}{l} \langle o, m \rangle, \langle o, s \rangle, \\ \langle y, m \rangle, \langle y, s \rangle, \\ \langle a, t \rangle, \langle b, t \rangle \end{array} \right\}$$

$$\text{female}(X) = \left\{ \begin{array}{l} \langle y \rangle, \langle b \rangle, \\ \langle s \rangle, \langle t \rangle \end{array} \right\} \quad \text{male}(X) = \left\{ \begin{array}{l} \langle o \rangle, \langle a \rangle, \\ \langle m \rangle, \langle d \rangle \end{array} \right\}$$

QUESTION10

Show the FOIL_gain calculation when $\text{sibling}(Y, Z)$ is considered as the first literal to be added to the first rule learning by FOIL. Also show the tuples that are involved in this calculation.