Homework-5

Question 1

As discussed in class the concept of a medium built man can be expressed as an axis parallel rectangle. Therefore, it can be defined in terms of 4 real numbers: a, b, c, d such that:

$$a \leq \text{height} \leq b, \qquad c \leq \text{weight} \leq d$$

We assume that a consistent algorithm is available.

Part I

In this part we take as hypotheses 4 floating point numbers, each represented in terms of 4 bytes (32 bits).

a How many hypotheses are in the hypotheses class?

b How many rendomly obtained examples are needed for PAC learning with accuracy parameter $\epsilon = 0.1$ and confidence parameter $\delta = 0.05$?

c What is the confidence level (what is δ) if it is known that 1000 randomly chosen examples were used, and the desired accuracy is level is $\epsilon = 0.1$?

d What is the accuracy level (what is ϵ) if it is known that 1000 randomly chosen examples were used, and the desired confidence level is $\delta = 0.05$?

Part II

In this part we assume that we are able to represent arbitrarily accurate axis parallel rectangles.

a Prove that the VC dimension of the concept class of all axis parallel rectangles is $d = 4$.

b How many rendomly obtained examples are needed for PAC learning with accuracy parameter $\epsilon = 0.1$ and confidence parameter $\delta = 0.05$?

c What is the confidence level (what is δ) if it is known that 1000 randomly chosen examples were used, and the desired accuracy is level is $\epsilon = 0.1$?

d What is the accuracy level (what is ϵ) if it is known that 1000 randomly chosen examples were used, and the desired confidence level is $\delta = 0.05$?

Part III

Do you think that your answer to Part I can be improved based on your answer to Part II? Explain.

Question 2

Consider the problem of learning the concept of “ *possible temperature for this month*” which gives minimum and maximum temperatures for each one of the twelve months. The concept has the following form:

> The temperature in January is at least **min1** and at most **max1**.
> The temperature in February is at least **min2** and at most **max2**.
> The temperature in March is at least **min3** and at most **max3**.
> The temperature in April is at least **min4** and at most **max4**.
> The temperature in May is at least **min5** and at most **max5**.
> The temperature in June is at least **min6** and at most **max6**.

The temperature in July is at least **min7** and at most **max7**.
The temperature in August is at least **min8** and at most **max8**.
The temperature in September is at least **min9** and at most **max9**.
The temperature in October is at least **min10** and at most **max10**.
The temperature in November is at least **min11** and at most **max11**.
The temperature in December is at least **min12** and at most **max12**.

The values of **min1, max1, ..., min12, max12** are computed from randomly chosen training examples. A training example has the form of (date and time , temperature), e.g., (4/26/92 at 2 PM , 78).
A test example is a question such as:

“ Is 53 a possible temperature in July?”

It is assumed that the test questions and the training examples come from the same probability distribution. In the following two cases compute how many randomly chosen training examples are needed to guarantee with confidence of at least 95% that at least 90% of the test examples are answered correctly. Specify which formula you use for the computation, and what is the value of each of the variables in the formula.

a. It is known that a solution can be found in which each value of **min1, max1, ..., min12, max12** is represented by 4 bits.

Answer: The number of training examples should be at least _______. The formula used: _______. The variables in the formula have the values:

b. It is known that a solution can be found in which each value of **min1, max1, ..., min12, max12** is kept as a real number.

Answer: The number of training examples should be at least _______. The formula used: _______. The variables in the formula have the values:

Question 3

Three learning algorithms in a binary classification problem are applied to a set of 1000 training examples, selected at random. (All three are applied to the same training examples.)

Algorithm A produces **Classifier A** that correctly classifies 800 examples and incorrectly classifies 200 examples.

Algorithm B produces **Classifier B** that correctly classifies 800 examples and incorrectly classifies 200 examples. All the mistakes of Classifier B are on examples that were correctly classified by Classifier A.

Algorithm C produces **Classifier C** that correctly classifies 900 examples and incorrectly classifies 100 examples. All the mistakes of Classifier C are on examples that were correctly classified by both Classifier A and Classifier B.

Combine the three classifiers using the AdaBoosting technique. Use Classifier A as the first weak classifier, Classifier B as the second weak classifier, and Classifier C as the third weak classifier.

1 What would be the weight of Classifier A?

Answer: $\alpha_1 =$

2 What would be the weight of Classifier B?

Answer: $\alpha_2 =$

3 What would be the weight of Classifier C?

Answer: $\alpha_3 =$

4 The test example x_1 is classified as POSITIVE by Classifier A, POSITIVE by Classifier B, and NEGATIVE by Classifier C. What would be the classification of x_1 by the combined (boosted) classifier?

Answer: POSITIVE / NEGATIVE / Impossible-to-tell

5 The test example x_2 is classified as POSITIVE by Classifier A, NEGATIVE by Classifier B, and POSITIVE by Classifier C. What would be the classification of x_2 by the combined (boosted) classifier?

Answer: POSITIVE / NEGATIVE / Impossible-to-tell

6 The test example x_3 is classified as NEGATIVE by Classifier A, POSITIVE by Classifier B, and POSITIVE by Classifier C. What would be the classification of x_3 by the combined (boosted) classifier?

Answer: POSITIVE / NEGATIVE / Impossible-to-tell