Machine Learning, 2023 Spring
Assignment 1

Notice

Plagiarizer will get 0 points.
LaTeXis highly recommended. Otherwise you should write as legibly as possible.

Problem 1 If $\mu = 0.9$, what is the probability that a sample of 10 marbles will have $\nu \leq 0.1$?

[Hints: 1. Use binomial distribution. 2. The answer is a very small number]

By using binomial distribution, we can assume the situation:
There exists 10 marbles and among marbles 90% red marble and 10% green marbles,
What's the probability that the number of red marbles smaller than or equal to 1?

$P(X=k) = C_{10}^{k} \mu^{k}(1-\mu)^{10-k}$ with $\mu = 0.9$

$\Rightarrow$ The number of red marbles is 0 or 1, meaning $X = 0.1$

Therefore,

$$P(V \leq 0.1) = P(X=0) + P(X=1) = C_{10}^{0} \times 0.9^{0} \times 0.1^{10} + C_{10}^{1} \times 0.9^{1} \times 0.1^{9}$$
$$= 0.1^{10} + 9 \times 0.1^{9}$$
$$= 9.1 \times 10^{-9}$$

Problem 2 If $\mu = 0.9$, use the Hoeffding Inequality to bound the probability that a sample of 10 marbles will have $\nu \leq 0.1$ and compare the answer to the previous exercise.

The Hoeffding inequality:

$$P[|\mu - \nu| > \varepsilon] \leq 2e^{-2\varepsilon^2 N}$$

$$\therefore P[\nu \leq 0.1] = P[0.9 - \nu \geq 0.8]$$

$$= P[\mu - \nu \geq 0.8] \leq P[|\mu - \nu| \geq 0.8]$$

$$\leq 2e^{-2\times 0.8^2 \times 10} \approx 5.52 \times 10^{-6}$$

The result in Problem 1 shows $P[\nu \leq 0.1] = 9.1 \times 10^{-9}$ surely satisfy $9.1 \times 10^{-9} < 5.52 \times 10^{-6}$

We can find that the restriction of Hoeffding inequality is flexible

Problem 3 We are given a data set $\mathcal{D}$ and of 25 training examples from an unknown target function $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{-1, +1\}$. To learn f, we use a simple hypothesis set $\mathcal{H} = \{h_1, h_2\}$ and, where h_1 is the constant $+1$ function and h_2 is the constant -1.

We consider two learning algorithms, S (smart) and C (crazy). S chooses the hypothesis that agrees the most with $\mathcal{D}$ and C chooses the other hypothesis deliberately. Let us see how these algorithms perform out of sample from the deterministic and probabilistic points of view. Assume in the probabilistic view that there is a probability distribution on $\mathcal{X}$, and let $\mathbb{P}[f(x) = +1] = p$.

(a) Can S produce a hypothesis that is guaranteed to perform better than random on any point outside $\mathcal{D}$?

(b) Assume for the rest of the exercise that all the examples in $\mathcal{D}$ have $y_n = +1$. Is it possible that the C hypothesis that produces turns out to be better than the hypothesis that S produces?

(c) If $p = 0.9$, what is the probability that S will produce a better hypothesis than C?

(d) Is there any value of p for which it is more likely than not that C will produce a better hypothesis than S ?

(a) No, it can't.
We suppose that f has 25+1 on D but -1 on all other points of X. In this situation, S will choose h_1 and can't match any outside D. Now there exists a hypothesis that choose +1 and -1 equally, which performs better than S.

(b) We can use the example I have proposed in part(a).
In that situation C performs better than S. satisfying the restrictions in part(b).
So it is possible that C performs better than S.

(c) We suppose the result of S is f_S, the result of C is f_C
what we need to calculate is $P(P[f_S=f] > P[f_C=f])$
with $f_S(x)=1$. $P[f(x)=+1] = p = 0.9$. so:
$P[f_S=f] = P[f(x)=+1] = p = 0.9$ $\quad P[f_C=f] = 1 - P[f_S=f] = 0.1$
so $P(P[f_S=f] > P[f_C=f]) = P(0.9 > 0.1) = 1$

(d) We can do this problem after part(c)
What we need to find is the value of p. making $P(P[f_S=f] > P[f_C=f]) < 0.5$
Additionally. $P[f(x)=+1] = p$.
$\Rightarrow P[f_S=f] = p$. $P[f_C=f] = 1-p$. $\Rightarrow P(P[f_S=f] > P[f_C=f]) = P(p > 1-p)$
Then $P(p > 1-p) = 0$, meaning $p < 1-p$ (when $p = 1-p$, it not satisfy)
$\Rightarrow 2p < 1$, $p < 0.5$
So the answer is: $p < 0.5$

Problem 4 A friend comes to you with a learning problem. She says the target function f is completely unknown, but she has 4,000 data points. She is willing to pay you to solve her problem and produce for her a g which approximate f. What is the best that you can promise her among the following:

(a) After learning you will provide her with a g that you will guarantee approximates f well out of sample.

(b) After learning you will provide her with a g, and with high probability the g which you produce will approximate f well out of sample.

(c) One of two things will happen.

(i) You will produce a hypothesis g;

(ii) You will declare that you failed.

If you do return a hypothesis g, then with high probability the g which you produce will approximate f well out of sample.

The best is (c), the reasons are as follows:
The first possibility is that the unknown f is too complex to learn.
In the other hand, the nature of f changes for points out of sample,
then the hypothesis g fails. According to Hoeffding inequality, the probability
that g matches f is highly relative, and the error on g might be small when the
dataset is large.
So (c) choice is the best.

Problem 5

Given target function $f(x) = ax$ (a unknown), now we have the dataset $\{x_1, y_1\}$ ... $\{x_n, y_n\}$. Which hypothesis class will you chose, $H = \{ax + b\}$ or $H = \{ax\}$, explain your reason.

I will choose $H = \{ax+b\}$. Reasons are as follows:

Instead of $y = f(x)$, we usually take the output y to be a random variable, that is affected by, not determined by, the input x. Formally, the target distribution $P(y|x)$ is better than the target function $y = f(x)$. When we consider the target we should think of a noisy target as a deterministic target plus added noise. In this problem, the deterministic one is ax and the noisy one is b. Additionally, A deterministic target $H = \{ax\}$ is completely a special case of the noisy one: $H = \{ax+b\}$ where $b=0$. Therefore, there is no loss of generality if we consider the target involved a noisy b. So $H = \{ax+b\}$ is my choice.