SHANGHAITECH UNIVERSITY

Machine Learning, 2024 Spring
Solution for Assignment 1

March 30, 2024

Exercise 1.8

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.][10pt]
By the problem, $\nu \leq 0.1 \Leftrightarrow$ a sample of 10 marbles will have at most 1 red marble, then we have

$$\begin{aligned}
P(\nu \leq 0.1) &= P(1 \text{ red marble}) + P(0 \text{ red marble}) \\
&= (1-0.9)^{10} + C_{10}^{1} \times 0.9 \times (1-0.9)^{9} \\
&= 9.1 \times 10^{-9}
\end{aligned}$$

Exercise 1.9

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.[10pt]
In this problem,

$$P[|\mu - \nu| \geq 0.8] \leq 2e^{-2*0.8^2*10} \approx 5.52 * 10^{-6}$$

$9.1 \times 10^{-9} < 5.52 * 10^{-6}$, which satisfies Hoeffding Inequality.

Problem 1.10

Assume that $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}, \ldots, \mathbf{x}_{N+M}\}$ and $\mathcal{Y} = \{-1, +1\}$ with an unknown target function $f : \mathcal{X} \to \mathcal{Y}$. The training data set $\mathcal{D}$ is $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$. Define the off-training-set error of a hypothesis h with respect to f by

$$E_{\text{off}}(h, f) = \frac{1}{M} \sum_{m=1}^{M} [\![h(\mathbf{x}_{N+m}) \neq f(\mathbf{x}_{N+m})]\!].$$

(a) Say $f(\mathbf{x}) = +1$ for all $\mathbf{x}$ and

$$h(\mathbf{x}) = \begin{cases} +1, & \text{for } \mathbf{x} = \mathbf{x}_k \text{ and } k \text{ is odd and } 1 \leq k \leq M+N \\ -1, & \text{otherwise} \end{cases}$$

What is $E_{\text{off}}(h, f)$? [10pt]

(b) We say that a target function f can 'generate' $\mathcal{D}$ in a noiseless setting if $y_n = f(\mathbf{x}_n)$ for all $(\mathbf{x}_n, y_n) \in \mathcal{D}$. For a fixed $\mathcal{D}$ of size N, how many possible $f : \mathcal{X} \rightarrow \mathcal{Y}$ can generate $\mathcal{D}$ in a noiseless setting? [10pt]

(c) For a given hypothesis h and an integer k between 0 and M, how many of those f in (b) satisfy $E_{\text{off}}(h, f) = \frac{k}{M}$? [10pt]

(d) For a given hypothesis h, if all those f that generate $\mathcal{D}$ in a noiseless setting are equally likely in probability, what is the expected off-trainingset error $\mathbb{E}_f\left[E_{\text{off}}(h, f)\right]$? [10pt]

(e) A deterministic algorithm A is defined as a procedure that takes $\mathcal{D}$ as an input, and outputs a hypothesis $h = A(\mathcal{D})$. Argue that for any two deterministic algorithms A_1 and A_2, [10pt]

$$\mathbb{E}_f\left[E_{\text{off}}\left(A_1(\mathcal{D}), f\right)\right] = \mathbb{E}_f\left[E_{\text{off}}\left(A_2(\mathcal{D}), f\right)\right]$$

Solution

(a) From the problem, we can get

$$\begin{aligned} \boldsymbol{E}_{\text{off}}(h, f) &= \frac{1}{M} \sum_{m=1}^{M} [\![h\left(\mathbf{x}_{N+m}\right) = -1]\!] \\ &= \frac{1}{M}\left(\left\lceil \frac{N+M}{2} \right\rceil - \left\lceil \frac{N}{2} \right\rceil\right) \end{aligned}$$

(b) In a noiseless setting, there is no error on the training set $\mathcal{D}$, and the values on $\{x_{N+1}, \ldots, x_{N+M}\}$ are arbitrary, with two values at each point, so there are a total of 2^M species that can be fitted to f

(c) There are M points in total, and there are k points that can be different from the objective function, so there are C_M^k kinds of f

(d) Each noiseless f on the training set has the same probability, so these f should have the same probability of being wrong at each point on the test set, so

$$\begin{aligned} \mathbb{E}_f\left[E_{\text{off}}(h, f)\right] &= \sum_{k=0}^{M} \frac{k}{M} \frac{C_M^k}{2^M} \\ &= \frac{\sum_{k=0}^{M} k C_M^k}{M 2^M} \\ &= \frac{\sum_{k=1}^{M} M C_{M-1}^{k-1}}{M 2^M} \\ &= \frac{2^{M-1}}{2^M} \\ &= \frac{1}{2} \end{aligned}$$

(e) From (d), we can get that the expected off-training-set error only depends M, not on the specific form of the algorithm. So, for any two deterministic algorithms A_1 and A_2, $\mathbb{E}_f\left[E_{\text{off}}\left(A_1(\mathcal{D}), f\right)\right] = \mathbb{E}_f\left[E_{\text{off}}\left(A_2(\mathcal{D}), f\right)\right]$

0.1 Problem 1.12

This problem investigates how changing the error measure can change the result of the learning process. You have N data points $y_1 \leq \cdots \leq y_N$ and wish to estimate a 'representative' value.

(a) If your algorithm is to find the hypothesis h that minimizes the in-sample sum of squared deviations,

$$E_{\text{in}}(h) = \sum_{n=1}^{N} (h - y_n)^2$$

then show that your estimate will be the in-sample mean, [10pt]

$$h_{\text{mean}} = \frac{1}{N} \sum_{n=1}^{N} y_n$$

(b) If your algorithm is to find the hypothesis h that minimizes the in-sample sum of absolute deviations, [10pt]

$$E_{\text{in}}(h) = \sum_{n=1}^{N} |h - y_n|$$

then show that your estimate will be the in-sample median h_{med}, which is any value for which half the data points are at most h_{med} and half the data points are at least h_{med}.

(c) Suppose y_N is perturbed to $y_N + \epsilon$, where $\epsilon \to \infty$. So, the single data point y_N becomes an outlier. What happens to your two estimators h_{mean} and h_{med}? [10pt]

Solution

(a)The derivation of E_{in} with respect to h yields

$$E'_{\text{in}}(h) = 2 \sum_{n=1}^{N} (h - y_n)$$
$$E''_{\text{in}}(h) = 2N > 0$$

So E_{in} takes a minimum value at $E'_{\text{in}}(h) = 0$, and we can get

$$h = h_{\text{mean}} = \frac{1}{N} \sum_{n=1}^{N} y_n$$

(b)Before we give the solution, we prove a general conclusion:

Suppose that the probability density function of the random variable x is $f(x)$. When $F(a) = \int_{-\infty}^{+\infty} |x - a| f(x)dx$ takes a minimum value, a needs to satisfy $\int_{-\infty}^{a} f(x)dx = \frac{1}{2}$.

Prove: The derivation of $F(a) = \int_{-\infty}^{a} (a - x) f(x)dx + \int_{a}^{\infty} (x - a) f(x)dx$ with respect to a yields

$$F'(a) = \int_{-\infty}^{a} f(x)dx - \int_{a}^{\infty} f(x)dx$$
$$F''(a) = 2f(a) \geq 0$$

$F(a)$ takes a minimum value when $F'(a) = 0$, which means that $\int_{-\infty}^{a} f(x)dx = \int_{a}^{\infty} f(x)dx = \frac{1}{2}$. Therefore, $F(a)$ takes a minimum value at $a = x_{\text{med}}$.

Construct a distribution $P\left(y=y_i\right)=\frac{1}{N}(i=1,2\dots N)$, and we can get

$$F(h)=\frac{1}{N}E_{\text{in}}\left(h\right)=\frac{1}{N}\sum_{n=1}^{N}\left|h-y_n\right|$$

From the general conclusion, we can know that $F(h)$ takes a minimum value at $h=y_{\text{med}}$

(c)When y_N is perturbed to $y_N+\epsilon$, where $\epsilon\rightarrow\infty$, $h_{\text{mean}}=\frac{1}{N}\sum_{n=1}^{N}y_n\rightarrow\infty$, while h_{med} doesn't change because only the relative order of y_N changes.