Learning Theory

Machine Learning II (2023-2024)
UMONS

1 Exercise 1

Consider a sample of 10 marbles drawn independently from a bin that holds red and green marbles. The probability of a red marble is μ. For $\mu = 0.05$, $\mu = 0.5$, and $\mu = 0.8$, compute the probability of getting no red marbles ($\nu = 0$) in the following cases.

(a) We draw only one such sample. Compute the probability that $\nu = 0$.

(b) We draw 1,000 independent samples. Compute the probability that (at least) one of the samples has $\nu = 0$.

(c) Repeat (b) for 1,000,000 independent samples.

2 Exercise 2

Here is an experiment that illustrates the difference between a single bin and multiple bins. Run a computer simulation for flipping 1,000 fair coins. Flip each coin independently 10 times. Let's focus on 3 coins as follows: c_1 is the first coin flipped; c_{rand} is a coin you choose at random; $c_{\min}$ is the coin that had the minimum frequency of heads (pick the earlier one in case of a tie). Let ν_1, ν_{rand} and $\nu_{\min}$ be the fraction of heads you obtain for the respective three coins.

(a) What is μ for the three coins selected?

(b) Repeat this entire experiment a large number of times (e.g., 100,000 runs of the entire experiment) to get several instances of ν_1, ν_{rand} and $\nu_{\min}$ and plot the histograms of the distributions of ν_1, ν_{rand} and $\nu_{\min}$. Notice that which coins end up being c_{rand} and $c_{\min}$ may differ from one run to another.

(c) Using (b), plot estimates for $\mathbb{P}[|\nu - \mu| > \epsilon]$ as a function of ϵ, together with the Hoeffding bound $2e^{-2\epsilon^2 N}$ (on the same graph) .

(d) Which coins obey the Hoeffding bound, and which ones do not? Explain why.

(e) Relate part (d) to the multiple bins in Figure 1.

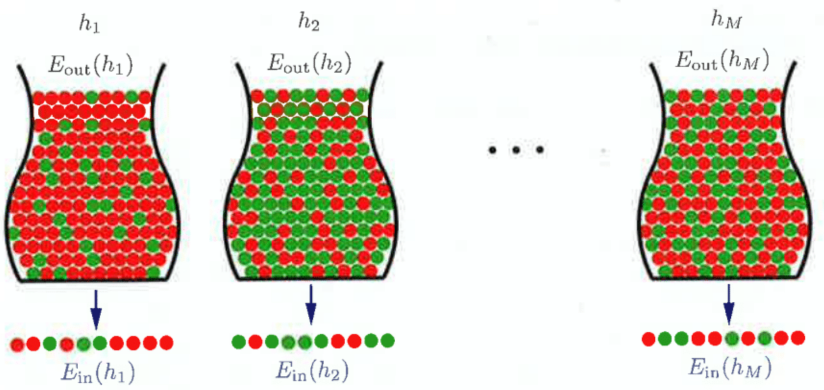

Figure 1.10: Multiple bins depict the learning problem with M hypotheses

Figure 1: Source: Abu-Mostafa et al. Learning from data. AMLbook.

See this Python notebook

https://colab.research.google.com/drive/1CB3s2RpkfdU9y7tx7RLT5hS_bGNZoM_s?usp=sharing

3 Exercise 3

The Hoeffding Inequality is one form of the law of large numbers. One of the simplest forms of that law is the Chebyshev Inequality, which you will prove here.

(a) If t is a non-negative random variable, prove that for any $\alpha > 0$, $\mathbb{P}[t \geq \alpha] \leq \mathbb{E}(t)/\alpha$.

(b) If u is any random variable with mean μ and variance σ^2, prove that for any $\alpha > 0$, $\mathbb{P}[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{\alpha}$. [**Hint:** Use (a)]

(c) If $u_1, \cdots, u_N$ are iid random variables, each with mean μ and variance σ^2, and $u = \frac{1}{N}\sum_{n=1}^{N} u_n$, prove that for any $\alpha > 0$,

$$\mathbb{P}[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{N\alpha}.$$

Notice that the RHS of this Chebyshev Inequality goes down linearly in N, while the counterpart in Hoeffding's Inequality goes down exponentially. In Exercise 5, we develop an exponential bound using a similar approach.

4 Background

The moment generating function (MGF) of a random variable X is given by:

$$M_X(s) = \mathbb{E}[e^{Xs}].$$

We called it the moment generating function because its derivatives evaluated at 0 provides the moments of X. In fact,

$$M'_X(0) = \left[\frac{d}{ds}\mathbb{E}[e^{Xs}]\right]_{s=0} = \mathbb{E}\left[\frac{d}{ds}e^{Xs}\right]_{s=0} = \mathbb{E}\left[Xe^{Xs}\right]_{s=0} = \mathbb{E}[X].$$

More generally, we have

$$M_X^{(k)}(0) = \mathbb{E}[X^k],$$

for $k = 1, 2, \ldots$.

There are two important properties of MGFs:

- *Sums of independnet random variables*: If we have random variables $X_1, X_2, \ldots, X_N$, which are independent, and $Y = \sum_{n=1}^{N} X_n$, then

  $$M_Y(s) = \prod_{n=1}^{N} M_{X_n}(s).$$

  Basically, this allows us to calculate effectively every moment of a sum of independent random variables.

- *Equality of MGFs*: If the MGF of X and Y exist, and are equal, then X and Y have the same distribution.

5 Exercise 4

In this problem, we derive a form of the law of large numbers that has an exponential bound, called the Chernoff bound. We focus on the simple case of flipping a fair coin, and use an approach similar to Exercise 3.

(a) Let t be a (finite) random variable, a be a positive constant, and s be a positive parameter. If $T(s) = \mathbb{E}_t(e^{st})$, prove that

$$\mathbb{P}[t \geq \alpha] \leq e^{-s\alpha} T(s).$$

[**Hint:** e^{st} is monotonically increasing in t]

(b) Let $u_1, \cdots, u_N$ be iid random variables, and let $u = \frac{1}{N} \sum_{n=1}^{N} u_n$. If $U(s) = \mathbb{E}_{u_n}(e^{su_n}$ (for any n), prove that

$$\mathbb{P}[u \geq \alpha] \leq (e^{-s\alpha} U(s))^N.$$

(c) Suppose $\mathbb{P}[u_n = 0] = \mathbb{P}[u_n = 1] = \frac{1}{2}$ (fair coin). Evaluate $U(s)$ a a function of s, and minimize $e^{s\alpha} U(s)$ with respect to s for fixed α, $0 < \alpha < 1$.

(d) Conclude in (c) that, for $0 < \epsilon < \frac{1}{2}$,

$$\mathbb{P}[u \geq \mathbb{E}(u) + \epsilon] \leq 2^{-\beta N},$$

where $\beta = 1 + (\frac{1}{2} + \epsilon) \log_2(\frac{1}{2} + \epsilon) + (\frac{1}{2} - \epsilon) \log_2(\frac{1}{2} - \epsilon)$ and $\mathbb{E}(u) = \frac{1}{2}$. Notice that this bound is exponentially decreasing in N.

6 Exercise 5

Lemma 1 (Chernoff's method). Let X be a random variable. Then, for any $\varepsilon > 0$, we have

$$P(X > \varepsilon) \leq \inf_{s>0}\ e^{-s\varepsilon}\mathbb{E}[e^{Xs}] \text{ and } P(X < -\varepsilon) \leq \inf_{s>0}\ e^{-s\varepsilon}\mathbb{E}[e^{-Xs}].$$

Lemma 2 (Hoeffding's lemma). Suppose that $a \leq X \leq b$ and $\mu = \mathbb{E}[X]$. Then,

$$\mathbb{E}[e^{Xs}] \leq e^{s\mu}e^{\frac{s^2(b-a)^2}{8}}.$$

Hoeffding's inequality. Let $X_1, X_2, \ldots, X_N$ be i.i.d. observations such that $\mathbb{E}[X_n] = \mu$, $a \leq X_n \leq b$ and $\bar{X} = \frac{1}{N}\sum_{n=1}^{N} X_n$. Then, for any $\varepsilon > 0$,

$$P(|\bar{X} - \mu| > \varepsilon) \leq 2e^{-2N\varepsilon^2/(b-a)^2}.$$

Prove Hoeffding's inequality using the Chernoff's method and Hoeffding's Lemma, and without loss of generality, you can assume that $\mu = 0$.

7 Exercise 6

Which of the following are possible growth functions $m_{\mathcal{H}}(N)$ for some hypothesis set:

$$1 + N; 1 + N + \frac{N(N-1)}{2}; 2^N; 2^{\lfloor\sqrt{N}\rfloor}; 2^{\lfloor N/2\rfloor}; 1 + N + \frac{N(N-1)(N-2)}{6}.$$

8 Exercise 7

Compute the maximum number of dichotomies, $m_{\mathcal{H}}(N)$, for these learning models, and consequently compute d_{VC}, the VC dimension.

(a) Positive or negative ray: $\mathcal{H}$ contains the functions which are $+1$ on $[a, \infty)$ (for some a) together with those that are $+1$ on $(-\infty, a]$ (for some a).

(b) Positive or negative interval: $\mathcal{H}$ contains the functions which are $+1$ on $[a, b]$ and -1 elsewhere or -1 on an interval $[a, b]$ (for some a) together and $+1$ elsewhere.

(c) Two concentric spheres in $\mathbb{R}^d$: $\mathcal{H}$ contains the functions which are $+1$ for $a \leq \sqrt{x_1^2 + x_2^2 + \cdots + x_d^2} \leq b$

9 Exercise 8

$B(N,k)$ is the maximum number of dichotomies on N points such that no subset of size k of the N points can be shattered by these dichotomies. Now, Show that $B(N,k) = \sum_{i=0}^{k-1} \binom{N}{i}$ by showing the other direction to

$$B(N,k) \geq \sum_{i=0}^{k-1} \binom{N}{i}.$$

To do so, construct a specific set of $\sum_{i=0}^{k-1} \binom{N}{i}$ dichotomies that does not shatter any subset of k variables. [**Hint:** Try limiting the number of -1's in each dichotomy.]

10 Exercise 9

Prove by induction that $\sum_{i=0}^{D} \binom{N}{i} \leq N^D + 1$, hence

$$m_{\mathcal{H}}(N) \leq N^{d_{\mathrm{VC}}} + 1$$

11 Exercise 10

1. Let $\mathcal{H} = \{h_1, h_2, \ldots, h_M\}$ with some finite M. Prove that $d_{\mathrm{VC}}(\mathcal{H}) \leq \log_2 M$.
2. For hypothesis sets $\mathcal{H}_1, \mathcal{H}_2, \cdots, \mathcal{H}_K$ with finite VC dimensions $d_{\mathrm{VC}}(\mathcal{H}_k)$, derive and prove the highest upper and lower bound that you can get on $d_{\mathrm{VC}}(\cap_{k=1}^{K} \mathcal{H}_k)$.
3. For hypothesis sets $\mathcal{H}_1, \mathcal{H}_2, \cdots, \mathcal{H}_K$ with finite VC dimensions $d_{\mathrm{VC}}(\mathcal{H}_k)$, derive and prove the highest upper and lower bound that you can get on $d_{\mathrm{VC}}(\cup_{k=1}^{K} \mathcal{H}_k)$.