Ejercicios - *Learning from Data*

Universidad Nacional de Colombia, Sede Bogotá
Matemáticas para el Aprendizaje de Máquina
2022-I

Sebastian Leonardo Molina Diaz
smolinad@unal.edu.co

1.2. Suppose that we use a perceptron to detect spam messages. Let's say that each email message is represented by the frequency of occurrence of keywords, and the output is +1 if the message is considered spam.

(a) Can you think of some keywords that will end up with a large positive weight in the perceptron?

Solution.

Words like "OFF", "money", "win", "selected", "claim", and "prize" have a higher chance to appear in spam messages.

□

(b) How about keywords that will get a negative weight?

Solution.

Words like "email", "unsubscribe", "`user@email`", "privacy", "copyright", and "never" have a higher chance to appear in your inbox and being safe to open.

□

(c) What parameter in the perceptron directly affects how many borderline messages end up being classified as spam?

Solution.

The threshold or bias will affect how borderline messages end up being classified, as it defines where the decision bound will be located.

□

1.3. The weight update rule in (1.3) has the nice interpretation that it moves in the direction of classifying $\mathbf{x}(t)$ correctly.

(a) Show that $y(t)\mathbf{w}^{\mathsf{T}}(t)\mathbf{x}(t) < 0$. [Hint: $\mathbf{x}(t)$ is misclassified by $\mathbf{w}(t)$.]

Solution.

As $\mathbf{x}(t)$ is misclassified by $\mathbf{w}(t)$, $\text{sign}(\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)) \neq y(t)$. Therefore, $y(t) = -\text{sign}\left(\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)\right)$, and

$$\begin{aligned} y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t) &= -\text{sign}\left(\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)\right)\left(\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)\right) \\ &< 0. \end{aligned}$$

□

(b) Show that $y(t)\mathbf{w}^{\mathrm{T}}(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)$. [Hint: Use (1.3).]

Solution.

By (1.3) we have that

$$\mathbf{w}(t+1) = \mathbf{w}(t) + y(t)\mathbf{x}(t).$$

Therefore,

$$\begin{aligned} y(t)\mathbf{w}^{\mathrm{T}}(t+1)\mathbf{x}(t) &= y(t)\big(\mathbf{w}(t) + y(t)\mathbf{x}(t)\big)^{\mathrm{T}}\mathbf{x}(t) \\ &= y(t)\big(\mathbf{w}^{\mathrm{T}}(t) + y(t)\mathbf{x}^{\mathrm{T}}(t)\big)\mathbf{x}(t) \\ &= y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t) + \big(y(t)\big)^2\mathbf{x}^{\mathrm{T}}(t)\mathbf{x}(t) \\ &= y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t) + \big(y(t)\big)^2\|\mathrm{x}(t)\|^2 \\ &> y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t). \end{aligned}$$

□

(c) As far as classifying $\mathbf{x}(t)$ is concerned, argue that the move from $\mathbf{w}(t)$ to $\mathbf{w}(t+1)$ is a move 'in the right direction'.

Solution.

Observe that there are two cases:

- If $y(t) = -1$, then

$$y(t)\mathbf{w}^{\mathrm{T}}(t+1)\mathbf{x}(t) = y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t) + \|\mathrm{x}(t)\|^2,$$

  and as $\mathbf{w}^{\mathrm{T}}(t+1)\mathbf{x}(t) \geq 0$ because $\mathbf{x}(t)$ is misclassified by $\mathbf{w}^{\mathrm{T}}(t)$, then in the iteration $t+1$, the weights vector $\mathbf{w}^{\mathrm{T}}(t)$ is moving in the negative direction, *i.e.* in the -1 direction, by a magnitude of $\mathbf{w}^{\mathrm{T}}(t+1)\mathbf{x}(t)$.
- If $y(t) = +1$, by similar reason of the previous case, the weights vector $\mathbf{w}^{\mathrm{T}}(t)$ is moving in the positive direction, *i.e.* in the $+1$ direction, by a magnitude of $\mathbf{w}^{\mathrm{T}}(t+1)\mathbf{x}(t)$.

In any case, the move from $\mathbf{w}(t)$ to $\mathbf{w}(t+1)$ is a move 'in the right direction', *i.e.* in the direction of the correct label $y(t)$.

□

1.10. 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 ν_{min} be the fraction of heads you obtain for the respective three coins.

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

Solution.

From the experiment, we have that $\nu_1 = 0.522$, $\nu_{\text{rand}} = 0.495$, and $\nu_{\text{min}} = 0.438$. As each of the coins is a fair coin, $\mu = 0.5$.

□

(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 ν_{min} and plot the histograms of the distributions of ν_1, ν_{rand} and ν_{min}. Notice that which coins end up being c_{rand} and c_{min} may differ from one run to another.

Solution.

The experiment was run 100,000 times. Figure 1 corresponds to the required histogram.

□

(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).

Solution.

Figure 2 corresponds to the plot of the estimations for Hoeffding bound.

□

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

Solution.

The first and random coins are below the Hoeffding bound, as shown in Figure 2. However the coins with minimum fraction of heads is above

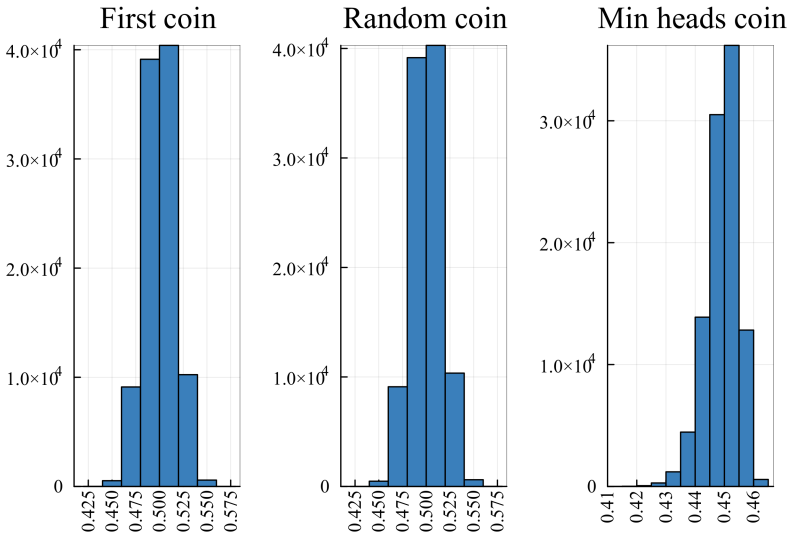

Figure 1: Histogram for the fraction of heads.

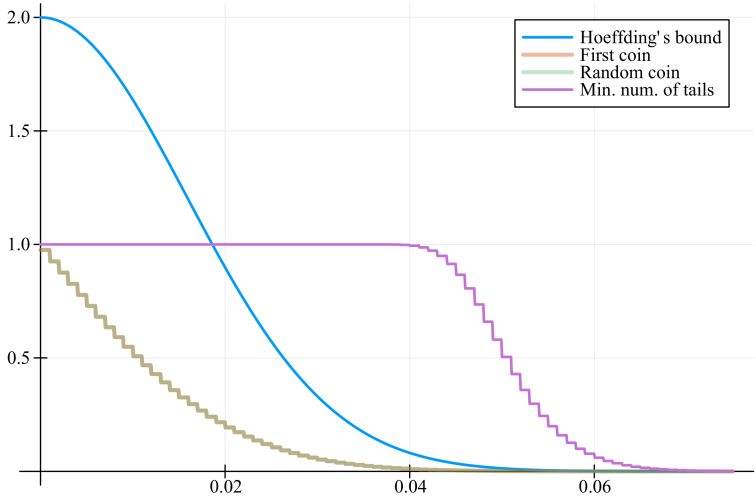

Figure 2: Estimation of $\mathbb{P}[|\nu - \mu| > \epsilon]$ for the experiment.

the bound. This happens because c_{first} and c_{rand} are in fact random samples, but c_{min} was selected with a set of rules from the observed data,

hence not random, and Hoeffding bound only applies for random samples.

□

(e) Relate part (d) to the multiple bins in Figure 1.10.

Solution.

Choosing c_{first} and c_{rand} is equivalent to choosing a hypothesis before generating the data, so it is equivalent to applying the Hoeffding bound to two different bins, which behave as the entire data. However, choosing c_{min} corresponds to consider all bins at the same time, as the rule of choosing the coin with minimum number of heads can be done only after generating all the data.

□

1.11. We are given a data set $\mathcal{D}$ 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 $\mathcal{H} = \{h_1, h_2\}$ 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 $\mathbb{P}[f(\mathbf{x}) = +1] = p$.

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

Solution.

No. S will choose a hyphotesis that will "memorize" the data in $\mathcal{D}$, but it may not be able to predict data outside the data set. Suppose that all data in $\mathcal{D}$ is labelled as $+1$, but somehow, all data outside $\mathcal{D}$ is labelled -1. In the other hand, C may choose a hypothesis which labels data 50% of the time as -1, and 50% of the time as $+1$. In this case, C will have a better approximation to f.

□

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

Solution.

Yes, by the same reasoning of the previous exercise. □

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

Solution.

Assuming S chooses h_1, as all the examples in $\mathcal{D}$ have $y_n = +1$, then $\mathbb{P}[h_1 = f] = \mathbb{P}[(\mathbf{x}) = +1] = 0.9$. In the other hand, as C chooses h_2, $\mathbb{P}[h_2 = f] = 0.1$. In this case,

$$\mathbb{P}\big[\mathbb{P}[h_1 = f] > \mathbb{P}[h_2 = f]\big] = \mathbb{P}[0.9 > 0.1] = 1,$$

so S will always 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?

Solution.

From the previous exercise, we can generalize to answer this question. We want to know when does C produce a better hypothesis than S. Therefore, we are looking when does

$$\mathbb{P}\big[\mathbb{P}[h_1 = f] > \mathbb{P}[h_2 = f]\big] < 0.5.$$

As $\mathbb{P}[h_1 = f] = p$ and $\mathbb{P}[h_2 = f] = 1 - p$, the inequality is true for $p < 0.5$.

□

1.12. 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 approximates 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.

Solution.

The best one can promise of the options above is (c). One cannot promise (a), as the target function can be intractable, and by the same reasoning, one

cannot promise (b), with high probability of g approximating f. If one can indeed learn from the data, by Hoeffding's Inequality we can ensure that $g \approx f$, because we have a large quantity of data points. Specifically, we have

$$\mathbb{P}\left[\|\mathrm{E}_{\text{in}}(g) - \mathrm{E}_{\text{out}}(g)\| > \varepsilon\right] < 2e^{-8000\varepsilon^2}.$$

With $\varepsilon > 0.06$, the previous probability is almost 0%, meaning that in-sample error and out-sample error will be separated between a range of 0.06 with almost 100% probability.

□