Universidad Nacional de Colombia
Matemáticas para el Aprendizaje de Máquinas 2022-I
Exercise List 1

1 de junio de 2022

Plazas Muñeton, Juan Felipe.

Problem 1 (Exercise 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 as spam.

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

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

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

Solution.

a) Free, giveaway, gift, specially, earn, money, Visa, MasterCard, ...

b) The, not, of, to, be, I, and, for, have, in, with, ...

c) In the perceptron, the parameter **b** is the threshold used to classify the points (emails) into positives (spam) and negatives (non-spam), so **b** directly affects how many borderline messages end up being classified as spam.

■

Problem 2 (Exercise 1.3) The weight update rule in (1.3), $\mathbf{w}(t+1) = \mathbf{w}(t) + y(t)\mathbf{x}(t)$, has the nice interpretation that it moves in the direction of classifying $\mathbf{x}(t)$ correctly.

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

b) Show that $y(t)\mathbf{w}^T(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^T(t)\mathbf{x}(t)$.

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.

a) If $\mathbf{x}(t)$ is misclassified by $\mathbf{w}(t)$, then $\mathbf{y}(t) \neq \text{Sign}(\mathbf{w}^T(t)\mathbf{x}(t))$, so we have two cases

- $y(t) = +1$ and $\text{Sign}(\mathbf{w}^T(t)\mathbf{x}(t)) = -1$.
- $y(t) = -1$ and $\text{Sign}(\mathbf{w}^T(t)\mathbf{x}(t)) = +1$.

In both cases we have that $y(t)\mathbf{w}^T(t)\mathbf{x}(t) < 0$.

b) Note that

$$\begin{aligned} y(t)\mathbf{w}^T(t+1)\mathbf{x}(t) &= y(t)(\mathbf{w}(t) + y(t)\mathbf{x}(t))^T\mathbf{x}(t) \\ &= y(t)(\mathbf{w}(t)^T + y(t)\mathbf{x}(t)^T)\mathbf{x}(t) \\ &= y(t)\mathbf{w}(t)^T\mathbf{x}(t) + y(t)^2\mathbf{x}(t)^T\mathbf{x}(t) \\ &= y(t)\mathbf{w}(t)^T\mathbf{x}(t) + y(t)^2||\mathbf{x}(t)||_2^2 \end{aligned}$$

and as the term $y(t)^2||\mathbf{x}(t)||_2^2 \geq 0$, then $y(t)\mathbf{w}^T(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^T(t)\mathbf{x}(t)$.

c) Due to the previous point, as the value of t increases $y(t)\mathbf{w}^T(t)\mathbf{x}(t)$ also increases. So we have two cases

- If $y(t) = -1$ and $\text{Sign}(\mathbf{w}^T(t)\mathbf{x}(t)) = +1$, the increase of $y(t)\mathbf{w}^T(t)\mathbf{x}(t)$ means that $\mathbf{w}^T(t)\mathbf{x}(t)$ decreases, i.e, moving toward negative region.
- If $y(t) = +1$ and $\text{Sign}(\mathbf{w}^T(t)\mathbf{x}(t)) = -1$, the increase of $y(t)\mathbf{w}^T(t)\mathbf{x}(t)$ means that $\mathbf{w}^T(t)\mathbf{x}(t)$ also increases, i.e, moving toward positive region.

From the above cases, we can conclude that the move from $\mathbf{w}(t)$ to $\mathbf{w}(t+1)$ is a move "in the right direction", as far as classifying $\mathbf{x}(t)$ is concerned.

■

Problem 3 (Exercise 1.10) Here is an experiment that illustrates the difference between a single bin and multiple bins. Run a computer simulation for flipping 1000 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 v_1, v_{rand} and v_{min} be the fraction of heads you obtain for the respective three coins.

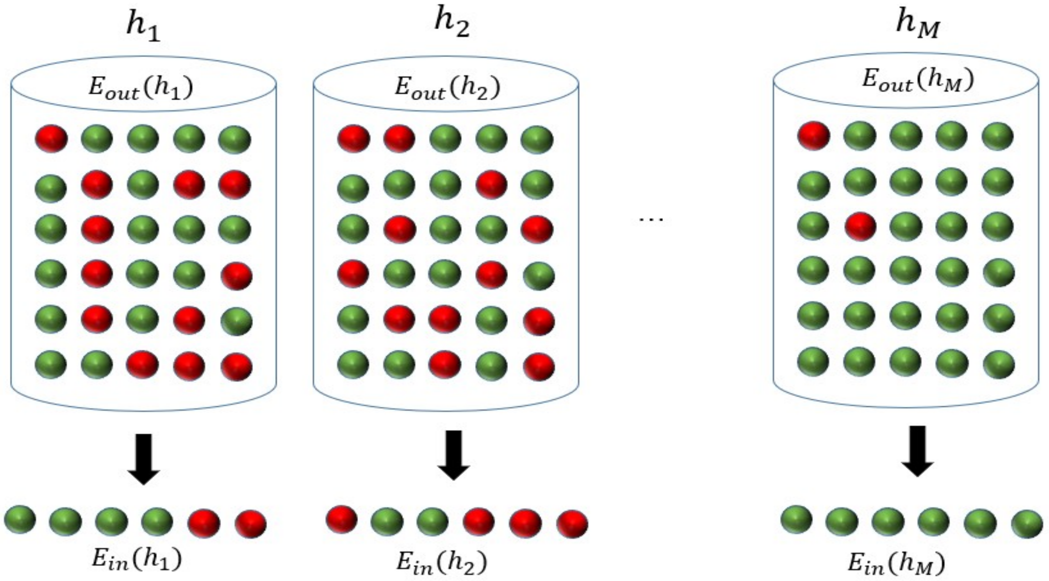
a) What is μ for the three coins selected?

b) Repeat this entire experiment a large number of times (e.g., 100000 runs of the entire experiment) to get several instances of v_1, v_{rand} and v_{min} and plot the histograms of the distributions of v_1, v_{rand} and v_{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 1.10. Multiple bins depict the learning problem with M hypotheses.



Solution.

a) The μ for the three coins are all 0.5 since the coins are fair.

b) *Solution code link*

c) *Solution code link*

d) From the graph above, we can see that the first coin and the random coin obey the Hoeffding bound while the coin with minimum frequency does not. This is because the first two coins were chosen before the experiment, while the third was chosen ***after*** the experiment and this violates the Hoeffding inequality condition which says the hypothesis has been fixed before samples were drawn.

e) When we choose the coin having the minimum frequency of heads. We are like choosing the bin from 1000 bins (our hypothesis space). But we choose bin after we finish sampling the data. This is akin to learning algorithm for the final hypothesis. The other two coins were chosen before the sampling, which is choosing bin beforehand.

■

Problem 4 (Exercise 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 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 $\mathcal{D}$?

b) Assume for the rest of the exercise that all the examples in $\mathcal{D}$ have $y_n = +1$. Is it possible that the hypothesis that C 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?

Solution.

a) No. If f classifies all the 25 training examples as $+1$ on D and as -1 all the elements of $\mathcal{X} \setminus \mathcal{D}$, i.e., f is of the next form

$$f(x) = \begin{cases} +1, & \text{if } x \in \mathcal{D} \\ -1, & \text{if } x \notin \mathcal{D} \end{cases}$$

then the hypothesis that agrees the most with $\mathcal{D}$ is h_1, so the learning algorithm S will produce the hypothesis h_1. Nevertheless, $f(x) = h_1(x)$ iff $x \in \mathcal{D}$, then the function f doesn't match with h_1 on $\mathcal{X} \setminus \mathcal{D}$ at no point. On the other hand, a random function g will have $+1$ in almost 50 % of the points and -1 in almost 50 % of the points, so g will match with h_1 in almost one point of on $\mathcal{X} \setminus \mathcal{D}$, then g is better than the function f produced by the learning algorithm $\mathcal{S}$.

b) Yes. Is the same example of a).

c) We have that $f(x) = +1$ for all $x \in \mathcal{D}$, then the hypothesis that agrees the most with $\mathcal{D}$ is h_1, so the learning algorithm S will produce the hypothesis h_1 and the learning algorithm C will produce the hypothesis h_2. So if $\mathbb{P}[f(\mathbf{x}) = +1] = 0.9$ then the hypothesis h_1 have a probability of 90 % to match with f on $\mathcal{X} \setminus \mathcal{D}$ and the hypothesis h_2 have a probability of 10 % to match with f on $\mathcal{X} \setminus \mathcal{D}$, so the learning algorithm S will always produce a better hypothesis than C.

d) Yes. From the exercise c), with the same reasoning we can conclude that for $p < 0.5$, C will always produce a better hypothesis than S.

■

Problem 5 (Exercise 1.12) A friend comes to you with a learning problem. She says the target function f is *completely* unknown, but she has 4000 data points. She is willing to pay to 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 that I can promise her is c) for two principal reasons, I have a considerable amount of data (4000 data points) and I **could** produce a hypothesis g. The Hoeffding inequality $\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$ allows us to assure that $E_{in}(g) \approx E_{out}(g)$; $E_{out}(g)$ is an unknown quantity since f is unknown, but $E_{in}(g)$ is a quantity that we can ascertain, as we can assure that g will approximate f well when $E_{out}(g) \approx 0$ then we only need to determinate if $E_{in}(g) \approx 0$ an this could be possible since we have a large dataset, so we have a high probability that g will approximate f well out of example. In other case, for example, in that the target fuction f is too complex and we don't have enough data to learn it, I will declare that I failed. ■