

# Introducción al Aprendizaje de Máquina

David Ricardo Pedraza Silva

April 9, 2021

## Exercise 1.1

Express each of the following tasks in the framework of learning from data by specifying the input space  $\chi$ , output space  $Y$ , target function  $f : \chi \rightarrow Y$ , and the specifics of the data set that we will learn from.

### Solution:

(a) Medical diagnosis: A patient walks in with a medical history and some symptoms, and you want to identify the problem.

$\chi$  is the set of all patients codified through this relevant information (medical history and symptoms) while  $Y$  is the set of all illnesses that could coincide with that info. Our target function  $f$  associates  $\chi$  with  $Y$  and, by definition, there is no better function which could fit  $\chi$  to  $Y$ .

(b) The handwritten digits must be represented in a way it is understandable to the machine. Let it be a matrix of 0's and 1's which represent blank and black space, respectively.  $f$  map the set of all matrices of this kind to  $Y = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ .

(c)  $f$  maps  $\chi$ , the set of lists of words; each of these lists the words in an email, to  $Y = \{False, True\}$ . "True"; it's spam. "False": it's not spam.

(d)  $\chi$  is the set of triples of the form  $(price, temperature, \text{day of the week})$  and  $f$  maps  $\chi$  to  $Y$ , where  $Y$  is a subset of the real numbers which, in turn, represents the electric load.

(e) Predicting if a videogame is going to go viral.  $\chi = \{x : x = (x_1, \dots, x_n)\}$  where  $x_i$  is a significative characteristic (Is it a psequel?, an RTS?, multiplayer?, etc...).  $Y = \{True, False\}$ , although this assumes "viral" is well defined.

## 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 output is +1 if the message is considered spam.

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

Everything having to do with prices "Winner", "Congratulations". Too many exclamation signs after a word (we could treat this as a word with some regex). The phrase "Visitor number x".

(b) What about words that will get a negative weight?

If the remitent is an official account, it might be legit. Using words in context and a measured use of punctuation are also signs of legitimate activity.

(c) Our bias, I suppose. How many of these words are we willing to tolerate in a single message.

## Exercise 1.3

The wight update rule in (1.3) has the nice interpretaton that it moves in the direction of classifying  $x(t)$  correctly.

(a) Show that  $y(t)W^T(t)x(t) < 0$ .

**solution:** Note that if  $x(t)$  is missclassified  $y(t) \neq \text{sig}(W^T(t)x(t))$ , then  $y(t)W^T(t)x(t) < 0$  by mere definition.

(b) Show that  $y(t)W^T(t+1)x(t) > y(t)W^T(t)x(t)$ .

**Solution:** Let's use that  $W(t+1) = W(t) + y(t)x(t)$ , then

$$y(t)W^T(t+1)x(t) = y(t)(W(t) + y(t)x(t))^T x(t)$$

That is

$$y(t)W^T(t+1)x(t) = y(t)W^T(t)x(t) + y(t)(x^T(t)y^T(t))x(t)$$

Note that  $y(t)(x^T(t)y^T(t))x(t) > 0$ . We conclude that

$$y(t)W^T(t+1)x(t) > y(t)W^T(t)x(t)$$

(c) Argue that the move from  $W(t)$  to  $W(t+1)$  is a move in the "right direction".

**Solution:** suppose  $y(t)$  is negative and  $W^T(t)x(t)$  is positive. If we take  $W(t+1) = W(t) + y(t)x(t)$  we see that  $W^T(t+1)x(t) = W^T(t)x(t) + y(t)|x(t)|^2$ . In this case  $W^T(t+1)x(t) < W^T(t)x(t)$ , which is good. If we have the opposite case; that in which  $y(t)$  is positive and  $W^T(t)x(t)$  negative, we get that  $W^T(t+1)x(t) > W^T(t)x(t)$ , which is also good.

## Exercise 1.11

From now on I won't write the formulation of the problem.

(a) No. As good as  $S$  might be at adjusting  $D$  the possibility of it being completely useless outside of  $D$  is non zero.

(b) Yes: Consider  $f(x) = 1$  if  $x \in D$  and  $f(x) = -1$  otherwise. In this case  $C$  is better than  $S$ .

(c) We know that for  $x \in \mathbb{R}$ ,  $P[f(x) = 1] = p$  with  $p = 0.9$ .

Let  $D = \{(x_1, y_1), \dots, (x_{25}, y_{25})\}$ . If  $D' = \{d \in D : d = (x_i, y_i) \text{ and } y_i = 1\}$ , for  $S$  to predict worst than  $C$  we need  $|D'| < 13$ . The probability of this is easy to calculate:

$$P(|D'| < 13) = \sum_{k=0}^{12} \binom{25}{k} p^k (1-p)^{25-k}$$

The result of this calculation is  $1.62083 \times 10^{-7}$ , extremely unlikely.

(d) If  $p > 0.5$  we need  $|D'| < 13$  for  $C$  to predict better than  $S$ . The probability of this happening is always less than 0.5.

If  $p < 0.5$  we need  $|D'| \geq 13$ . Again, the probability of this happening is less than 0.5. We conclude that taking  $S$  as our model is always smarter.