Lista 1 - Introdução ao Aprendizado de Máquina (MAC5832)

Gustavo Quintero - NUSP: 11350395

May 17, 2022

1. 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 estimated a "representative" value.

   (a) If your algorithm is to find the hypothesis h that minimize 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,

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

Proof

Differentiating the in sample sum of squared deviations with respect h, we obtain

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

Now, if $E'_{\text{in}}(h) = 0$, then,

$$\begin{array}{rcl} \sum_{n=1}^{N} (h - y_n) & = & 0 \\ \Rightarrow \sum_{n=1}^{N} h & = & \sum_{n=1}^{N} y_n \\ \Rightarrow h \sum_{n=1}^{N} & = & \sum_{n=1}^{N} y_n \\ \Rightarrow h\, N & = & \sum_{n=1}^{N} y_n \end{array}$$

Therefore, $h_{\text{mean}} = \frac{1}{N} \sum_{n=1}^{N} y_n$ is the only stationary point of E_{in}. Note that,

$$\begin{array}{rcl} E''_{\text{in}}(h) & = & 2 \sum_{n=1}^{N} \\ & = & 2N. \end{array}$$

Then, $E''_{\text{in}}(h) > 0$ for all h and we deduced that h_{mean} minimize the in sample sum of squared deviations.

(b) If your algorithm is to find the hypothesis h that minimize in sample sum of absolute deviations,

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

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 at least h_{med}.

Proof

First of all, note that

$$\begin{aligned} E'_{\text{in}}(h) &= \sum_{n=1}^{N} \frac{h - y_n}{|h - y_n|} \\ &= \sum_{n=1}^{N} \operatorname{sign}(h - y_n), \end{aligned} \tag{1}$$

for all $h \neq y_n$, $n = 1, \ldots, N$. Then, for all $h < y_1$, we have that $h < y_n$ for all $n = 1 \ldots, N$. So,

$$h - y_n < 0 \Rightarrow \operatorname{sign}(h - y_n) = -1, \quad \forall n = 1 \ldots, N.$$

Hence, by (1), $E'_{\text{in}}(h) = \sum_{n=1}^{N} \operatorname{sign}(h - y_n) = -N < 0$ and this implies that $E_{\text{in}}(h)$ is decreasing for all $h < y_1$. Analogously, we can show that $E_{\text{in}}(h)$ is increasing for all $h > y_N$.
Now, for all $h \in [y_1, y_N]$, we have that

$$\begin{aligned} E_{\text{in}}(h) &= \sum_{n=1}^{N} |h - y_n| \\ &= \left(\sum_{n=2}^{N-1} |h - y_n| \right) + |\underbrace{h - y_1}_{\geq 0}| + |\underbrace{h - y_N}_{\leq 0}| \\ &= \left(\sum_{n=2}^{N-1} |h - y_n| \right) + (y_N + y_1). \end{aligned} \tag{2}$$

Therefore, if N is odd, applying the above identity $(N-1)/2$ times, we obtain

$$E_{\text{in}}(h) = |h - y_{(N+1)/2}| + C,$$

where $C = (y_N - y_1) + (y_{N-1} - y_2) + \cdots + \left(y_{(N+3)/2} - y_{(N-1)/2}\right)$ is a constant value. So, if we define $h_{\text{med}} = y_{(N+1)/2}$ (that is, h_{med} is the in sample median), we have that $E_{\text{in}}(h_{\text{med}}) \leq E_{\text{in}}(h)$ for all $h \in [y_1, y_N]$. Therefore, we conclude that $E_{\text{in}}(h_{\text{med}}) \leq E_{\text{in}}(h)$ for all $h \in \mathbb{R}$, because $E_{\text{in}}(h)$ is decreasing and increasing for all $h < y_1$ and $h > y_N$, respectively.

On the other hand, if N is even, applying $(N-2)/2$ times the identity (2), for all $h \in [y_1, y_N]$, we obtain

$$E_{\text{in}}(h) = |h - y_{N/2}| + |h - y_{(N+2)/2}| + C,$$

where $C = (y_N - y_1) + (y_{N-1} - y_2) + \cdots + \left(y_{(N-2)/2} - y_{(N+4)/2}\right)$ is a constant value. Then,

$$\begin{aligned} E'_{\text{in}}(h) &= \frac{h - y_{N/2}}{|h - y_{N/2}|} + \frac{h - y_{(N+2)/2}}{|h - y_{(N+2)/2}|} \\ &= \operatorname{sign}\left(h - y_{(N+2)/2}\right) + \operatorname{sign}\left(h - y_{(N+2)/2}\right). \end{aligned}$$

So, $E'_{\text{in}}(h) = 0$ whenever $h \in \left(y_{N/2}, y_{(N+2)/2}\right)$. Indeed, if $h \in \left(y_{N/2}, y_{(N+2)/2}\right)$, then $h - y_{(N+2)/2} > 0$ and $h - y_{(N+2)/2} < 0$ implies that $\operatorname{sign}\left(h - y_{(N+2)/2}\right) = 1$ and $\operatorname{sign}\left(h - y_{(N+2)/2}\right) = -1$, respectively. Since $E_{\text{in}}(h)$ is decreasing for all $h < y_1$ and increasing for all $h > y_N$, for any

$h_{\text{med}} \in \left(y_{N/2}, y_{(N+2)/2}\right)$ (that is, h_{med} is the in sample median), we have that $E_{\text{in}}(h_{\text{med}}) \leq E_{\text{in}}(h)$ for all $h \in \mathbb{R}$.

Therefore, for all $N \in \mathbb{N}$, the hypothesis h that minimize E_{in} is the in sample median 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 estimators h_{mean} and h_{med}?

Solution

For h_{mean} we have that

$$h_{\text{mean}} = \frac{1}{N}\left(\sum_{n=1}^{N-1} y_n + (y_N + \epsilon)\right).$$

Then, $h_{\text{mean}} \to \infty$, because $\epsilon \to \infty$. On the other hand,

$$h_{\text{med}} = \text{median}\{y_1, \ldots, y_{N-1}, y_N + \epsilon\},$$

where $y_1 \leq \cdots \leq y_{N-1} \leq y_N + \epsilon$. Therefore, by definition of median, h_{med} remains unchanged, because $h_{\text{med}} < y_N < y_N + \epsilon$, for any $\epsilon \to \infty$.

2. In logistic regression, we saw that the function $h(\mathbf{x}) = \theta(\mathbf{w}^T\tilde{\mathbf{x}})$ is used to approximate $P(y = +1|\mathbf{x})$. In this way, we can, for example, consider that a given instance $\mathbf{x}$ belongs to the class $+1$ if $h(\mathbf{x}) > T$ and belongs to the class -1 if $h(\mathbf{x}) < T$, for a certain threshold $T \in [0, 1]$. If $h(\mathbf{x}) = T$ then $\mathbf{x}$ lies on the decision boundary. Show that, whatever threshold T is chosen, the decision boundary is a hyperplane.

Proof

Let $T \in [0, 1]$ as the hypothesis. Then, $\mathbf{x}$ lies on the decision boundary whenever that $\theta(\mathbf{w}^T\tilde{\mathbf{x}}) = T$, that is,

$$\frac{1}{1 + e^{-\mathbf{w}^T\tilde{\mathbf{x}}}} = T \Rightarrow e^{-\mathbf{w}^T\tilde{\mathbf{x}}} = \frac{1}{T} - 1 \Rightarrow -\mathbf{w}^T\tilde{\mathbf{x}} = \ln\left(\frac{1-T}{T}\right).$$

Note that, $\dfrac{1-T}{T} > 0$, then $\ln\left(\dfrac{1-T}{T}\right) \in \mathbb{R}$, and we deduced that the decision boundary consists of hyperplane

$$\tilde{\mathbf{w}}^T\tilde{\mathbf{x}} = 0,$$

where, $\tilde{\mathbf{w}} = \left(w_0 + \ln\left(\dfrac{1-T}{T}\right), w_1, \ldots, w_d\right)$.

3. In Example 3.4, it is mentioned that the output of the final hypothesis $g(\mathbf{x})$ learned using logistic regression can be thresholded to get a "hard" (± 1) classification. This problem shows how to use the risk matrix introduced in Example 1.1 to obtain such a threshold.

Consider fingerprint verification, as in Example 1.1. After learning from the data using logistic regression, you produce the final hypothesis

$$g(\mathbf{x}) = \mathbb{P}[y = +1 \mid \mathbf{x}],$$

which is your estimate of the probability that $y = +1$. Suppose that the cost matrix is given by

		True classification	
		+1 (correct person)	-1 (intruder)
you say	+1	0	c_a
	-1	c_r	0

For a new person with fingerprint $\mathbf{x}$, you compute $g(\mathbf{x})$ and you now need to decide whether to accept or reject the person (i.e., you need a hard classification). So, you will accept if $g(\mathbf{x}) \geq \kappa$, where κ is the threshold.

(a) Define the cost(accept) as your expected cost if you accept the person. Similarly define cost(reject). Show that

$$\begin{aligned} \text{cost(accept)} &= (1 - g(\mathbf{x}))c_a \\ \text{cost(reject)} &= g(\mathbf{x})c_r. \end{aligned}$$

Proof

Using the weights established in the cost matrix of hypothesis, we have that

$$\begin{aligned} \text{cost(accept)} &= \mathbb{P}[y = +1 \,|\, \mathbf{x}](0) + \mathbb{P}[y = -1 \,|\, \mathbf{x}](c_a) \\ &= \mathbb{P}[y = -1 \,|\, \mathbf{x}]c_a \\ &= (1 - g(\mathbf{x}))c_a \end{aligned}$$

and

$$\begin{aligned} \text{cost(reject)} &= \mathbb{P}[y = +1 \,|\, \mathbf{x}](c_r) + \mathbb{P}[y = -1 \,|\, \mathbf{x}](0) \\ &= \mathbb{P}[y = +1 \,|\, \mathbf{x}]c_r \\ &= g(\mathbf{x})c_r \end{aligned}$$

(b) Use part (a) to derive a condition on $g(\mathbf{x})$ for accepting the person and hence show that

$$\kappa = \frac{c_a}{c_a + c_r}.$$

Proof

A condition for the fulfillment of $g(\mathbf{x}) \geq k$ can be cost(accept) = cost(reject). Indeed, by part (a),

$$\begin{aligned} c_a(1 - g(\mathbf{x})) &= c_r\, g(\mathbf{x}) \\ \Rightarrow c_a - c_a\, g(\mathbf{x}) &= c_r\, g(\mathbf{x}) \\ \Rightarrow (c_a + c_r)g(\mathbf{x}) &= c_a \\ \Rightarrow g(\mathbf{x}) &= \frac{c_a}{c_a + c_r}. \end{aligned}$$

Therefore, if cost(accept) = cost(reject) we have that $g(\mathbf{x}) = k$, where

$$\kappa = \frac{c_a}{c_a + c_r}.$$

(c) Use the cost matrices for the Supermarket and CIA applications in Example 1.1 to compute the threshold κ for each of these two cases. Give some intuition for the threshold you get.

Proof

For the Supermarket we have that $c_a = 1$ and $c_r = 10$, and then $\kappa = 1/11$. This mean that the chances that the supermarket reject a consumer is very low, that is, the supermarket is not so rigorous when it comes to accepting the fingerprint of a person, even if he or she is not a frequent customer.

On the other hand, for the CIA we have that $c_a = 1000$ and $c_r = 1$. So,

$$\kappa = \frac{1000}{1001} = 1 - \frac{1}{1001}.$$

This means that the CIA will accept the fingerprint of a person only when the final hypothesis satisfies $g(\mathbf{x}) \geq \kappa \approx 1$, that is, when the probability that the fingerprint of a person is correct is very high. Equivalently, this means that the chance of having a false accept is very low.