MAC0460/5832 – Introdução ao Aprendizado de Máquina

DCC / IME-USP

Lista 1 (Data para entrega: até 25/04/2022)

Entregar em papel, manuscrito ou impresso.

Explique as suas respostas.

1. Problema 1.12 do livro-texto.

Problem 1.12 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 estimate a 'representative' value.

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

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

(b) If your algorithm is to find the hypothesis h that minimizes the 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 are at least 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 two estimators h_{mean} and h_{med}?

2. Em regressão logística, vimos que a função $h(\mathbf{x}) = \theta(\mathbf{w}\,\tilde{\mathbf{x}})$ é usada para aproximar $P(y = +1 \,|\, \mathbf{x})$. Desta forma, podemos por exemplo considerar que uma dada instância $\mathbf{x}$ é da classe +1 se $h(\mathbf{x}) > T$ e é da classe -1 se $h(\mathbf{x}) < T$, para um certo limiar $T \in [0, 1]$. Caso $h(\mathbf{x}) = T$ então $\mathbf{x}$ encontra-se na fronteira de decisão. Mostre que, qualquer que seja o limiar T escolhido, a fronteira de decisão é um hiperplano.

3. Problema 3.16 do livro-texto

Problem 3.16 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

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

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

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

(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 thresholds you get.