STAT 535 Homework 3
Out October 26, 2022
Due November 3, 2022

mmp@stat.washington.edu

Problem 1 – Bias and Variance again The questions in this problem refer to the setup of Problem 1 from Homework 2 ((Pb 1, Hw 2)).

a. Consider all the quantities you are asked to calculate or plot in (Pb 1, Hw 2), e.g. $\hat{l}_b, L_b, \hat{l}, V, \ldots$. List 3 of these which are *approximations*; of them, at least 2 should be *statistical approximations*. For example, computing a mean from samples is a statistical approximation, computing an integral by discretization is a numerical approximation. We assume computing is done with infinite precision, hence computing an integral by calling a function such as `erf, sin` is considered exact.

Explain (1 line or less) in each case what is (are) the approximation(s) made.

b. For one of your answers above, explain how you could increase the approximation accuracy.

c. Consider all the quantities you are asked to calculate or plot in (Pb 1, Hw 2). Is any of them exact? (No explanation here)

d. Assume that in (Pb 1, Hw 2), $n \to \infty$. Will the Bayes error $L^* \to 0$? Explain (1 line).

e. Assume that in (Pb 1, Hw 2), $n \to \infty$. Will the error bars on $L \to 0$? Explain (1 line).

f. Assume that in (Pb 1, Hw 2), $n \to \infty$. Will the error bars on $\hat{l} \to 0$? Explain (1 line).

g. Now consider that instead of K-NN, you have another classifier and another classification problem. Answer **d, e, f** again in this case.

h. For some prediction problem, not necessarily (Pb 1, Hw 2), $L(\hat{f}) = 0$; $\hat{f}$ is a predictor trained on a data set of size n. Does this imply $L^*(\hat{f}) = 0$? Explain (1 line).

i. For some prediction problem, not necessarily (Pb 1, Hw 2), $\hat{l}(\hat{f}) = 0$ (all training set predicted correctly). Does this imply $L(\hat{f}) = 0$? Does this imply $Var(\hat{f}) = 0$? Explain (1 line) in both cases.

[Problem 2 – The rate of decrease of MISE for Nadaraya-Watson regression – Extra credit] In Lecture II.1 it was shown that in $\mathbb{R}^d$, the kernel width h depends on n by $h \propto n^{-\frac{1}{d+4}}$ and that this is the optimal *rate* of decrease of h. The MISE is given by (note that in the lecture notes $d = 1$.)

$$MISE(h) = C_1 h^4 + C_2 \frac{1}{nh^d}. \tag{1}$$

a. What is the rate of decrease of MISE if h has the optimal rate? In other words, replace h in (1) with $h = C_3 n^{-\frac{1}{d+4}}$, then find an exponent a so that for $n \to \infty$

$$\frac{MISE}{n^a} \to C_4. \tag{2}$$

b. Now assume we make the choice $h = C_3 n^{-\frac{1}{d+5}}$. Repeat the previous question for this choice of h, and find the new exponent a' that represents the rate of decrease of MISE. Which is larger, a or a'? If our goal is a faster rate of decrease of MISE with respect to n, which choice of h is preferable?

In this problem, constants $C_1, C_2, C_3, \ldots$ are assumed to be > 0 and $< \infty$ (otherwise the answers become trivial). For example: $f(n) = 3n^4 + n + \ln(n)$. In this case, $f(n)/n^4 \to 3$ a finite, nonzero value. For any other exponent of n, the limit $f(n)/n^a$ is either 0 or infinity. Hence, we say the rate (of increase) of f is n^4. Similarly $f(n) = 5n^{-3} + 2n^{-1}$ has rate (of decrease) n^{-1}.

Problem 3 – Logit loss and backpropagation - NOT GRADED

The *logit loss*

$$L_{logit}(w) \;=\; \ln(1 + e^{-yw^T x}),\; x, w \in \mathbb{R}^n,\; y = \pm 1 \tag{3}$$

is the negative log-likelihood of observation (x, y) under the logistic regression model $P(y = 1|x, w) = \phi(w^T x)$ where ϕ is the logistic function.

a. Show that the partial derivatives $\frac{\partial L_{logit}}{\partial w_i}$, $\frac{\partial L_{logit}}{\partial x_i}$ for L_{logit} in (3) can be rewritten as

$$\frac{\partial L_{logit}}{\partial w_i} \;=\; -(1 - P(y|x, w)) y x_i \tag{4}$$

$$\frac{\partial L_{logit}}{\partial x_i} \;=\; -(1 - P(y|x, w)) y w_i. \tag{5}$$

The elegant formulas above hold for a larger class of statistical models, called Generalized Linear Models.

Problem 4 – Decision regions for the neural network

In this problem, the inputs are of the form $[x_1\; x_2]^T \in \mathbb{R}^2$ and if necessary we introduce the dummy variable $x_0 \equiv 1$.

a. Consider the following two-layer neural network

$$f(x) \;=\; \beta_0 + \sum_k \beta_k z_k \tag{6}$$

$$z_k \;=\; \phi\Big(\sum_{j=0}^{2} w_{jk} x_j\Big),\ \text{for}\, k = 1:K \tag{7}$$

$$x \;=\; \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} \tag{8}$$

$$W \;=\; [w_{jk}] \;=\; \begin{bmatrix} 1 & 0 & 2 & 2 & 2 \\ 1 & 1 & 0 & -1 & -0.5 \\ -1 & 1 & -1 & 0 & 1 \end{bmatrix} \times 20 \tag{9}$$

$$\phi(u) \;=\; \frac{1}{1 + e^{-u}}\ \text{the sigmoid function} \tag{10}$$

$$\beta_0 \;=\; -4.9,\; \beta_{1:5} \;=\; 1 \tag{11}$$

Plot the decision regions of this neural network, i.e the regions $D_\pm = \{x \,|\, f(x) \lessgtr 0\}$ and the decision boundary $\{x \,|\, f(x) = 0\}$.

b. Repeat the plots for $\beta_0 = -3.9$.