STAT 535 Homework 3
Out October 19, 2023
Due October 26, 2023
©Marina Meilă
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 each case.

[Problem 2 – The rate of decrease of MISE for Nadaraya-Watson regression – Extra credit] *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}.*

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 $n \to \infty$

$$\frac{MISE}{n^{-a_*}} \to C_4. \tag{2}$$

Denote $\alpha_* = \frac{1}{d+4}$. Note that with this notation h decreases at rate $n^{-\alpha_*}$ and $MISE(n)$ at rate n^{-a_*}.

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_b that represents the rate of decrease of MISE. Denote $\alpha_b = \frac{1}{d+5}$.

c. Which is larger, a_* or a_b? If our goal is a faster rate of decrease of MISE with respect to n, which choice of h is preferable? Which is larger, α_b or α_*? For which case is h decreasing faster with n? *Make a plot if you aren't sure*

d. Now assume we make the choice $h = C_3 n^{-\frac{1}{d+1}}$. Repeat the previous question for this choice of h, and find the new exponent a_d that represents the rate of decrease of MISE. Denote $\alpha_d = \frac{1}{d+1}$. For which case, **a.** or this case, is h decreasing faster with n?

e. Denote by $MISE_*(n), MISE_b(n), MISE_d(n)$ the functions obtained in questions **a, b d**. Make a well labeled plot of $\log_{10} MISE_{a,b,d}$ vs. $\log_{10} n$ for $d = 1$, from $n = 1 : 10^k$, where $k \geq 10$. Define the *dominant term* in the $MISE$ expression as the term which decreases to 0 slowest (hence it is reponsible for the rate a.) Choose the constants so that in all 3 functions the dominant term has the same value 1 at $n = 1$. Which of the three functions decreases faster? Plot also $\log_{1-} n^{-1}$ the rate of decrease of the variance for parametric estimation.

f. For this question, use only the dominant term in the $MISE$ expression (that is, neglect the non-dominant term). Suppose that, for the constants chosen above, we need $MISE_*(n) = \epsilon = 10^{-1}$. What is $n_{\epsilon,*}$ for which this error is achieved? Same question for $MISE_b = \epsilon$: find $n_{\epsilon,b}$ for which this error is achieved. What is the ratio $\frac{n_{\epsilon,b}}{n_{\epsilon,*}}$?

g. Give an expressions for $\text{excess}(\epsilon, d) = \frac{n_{\epsilon,b}}{n_{\epsilon,*}}$ as a function of ϵ and data dimension d. Does the excess grow or decrease with d?

h. Repeat questions **f, g** for $MISE_d$.

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

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

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

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