DATA/STAT/BIOSTAT 558
SPRING QUARTER 2025

Homework # 1
Online Submission to Canvas: due Wednesday April 16th, 5pm PST

Instructions: You may discuss the homework problems in small groups, but you must write up the final solutions and code yourself. Please turn in your code used for the problems that involve coding. However, code without written answers will receive no credit. To receive credit, you must explain your answers and show your work. All plots should be appropriately labeled and legible, with axis labels, legends, etc., as needed.

Problem 1 This problem has to do with the bias-variance trade-off and related ideas. It's okay to submit hand-sketched plots: you are not supposed to actually compute the quantities referred to below on data.

(a) Make a plot, like the one we saw in class, with "flexibility" (sometimes called complexity) of the function $\hat{f}$ on the x-axis. Sketch the following curves on the y-axis: squared bias, variance, irreducible error, reducible error, and expected prediction error. Be sure to label each curve. Indicate which level of flexibility is "best". Justify your answer (i.e., explain why this level of flexibility is best).

(b) For each of the following pairs of models, which do you expect to have higher bias and lower variance versus lower bias and higher variance?

  (a) Linear regression with just an intercept versus linear regression with an intercept and one predictor?

  (b) A model that has a training MSE of 100 versus a model with a training MSE of 200?

  (c) A model that interpolates the training data versus a model that does not interpolate the training data?

  Explain your answers.

(c) Consider two models: one has an expected prediction error of 2398, and the other has an expected prediction error of 1212. Based on this information, can you determine which model has higher squared bias and which model has higher variance? Explain your answer.

Problem 2 In this problem, we will consider constructing a k-nearest-neighbors classifier in a setting with $p = 2$ features.

Let the first feature $X_1 \sim \text{Unif}[-1, 1]$ and second feature $X_2 \sim \text{Unif}[-1, 1]$, i.e. the observations for each feature are independent and identically distributed (i.i.d.) from a uniform distribution.

Each observation belongs to one of two classes: the red class or the blue class. Given some value of $\tau \in (0, 1)$ to be specified below, the value of Y for a given observation is as follows:

- If $|X_1| < 0.2$ and $|X_2| < 0.2$, then $Y = \text{red}$.
- If $|X_1| \geq 0.2$ and/or $|X_2| \geq 0.2$, $Y = \text{red}$ with probability τ, and $Y = \text{blue}$ with probability $1 - \tau$.

(a) Letting $\tau = 0.4$, generate a training set of $n = 200$ observations as described above. Plot the training data. Make sure that the axes are properly labeled, and that each observation is colored according to its class label.

(b) The Bayes classifier assigns each observation to the most probable class, based on its (X_1, X_2) values. Write out a mathematical expression for the Bayes classifier for part (a).

(Hint: this should not require algebra. It should require geometry and logic.)

(c) On the plot from (a), color or shade each possible value of (X_1, X_2) either red or blue, depending on the prediction for Y arising from the Bayes classifier for that (X_1, X_2) pair.

(See Figure 2.13 of the textbook for guidance on what this might look like. You are being asked to color/shade EVERY possible value of (X_1, X_2), not just the pairs that happen to be in your training set.)

(d) For each possible value of (X_1, X_2), provide a mathematical expression for the classification error (i.e., the probability of misclassification) of the Bayes classifier.

(e) The expected classification error of the Bayes classifier is called the Bayes error. Compute the Bayes error.

(Hint: The Bayes error is the *expected* classification error, where the expectation is computed over the distribution of the X's. Thus, to compute the Bayes error, average the errors you computed in (d) over the distribution of (X_1, X_2). To do this, you can take advantage of the fact that X_1 and X_2 are uniformly distributed on $[-1, 1]$.)

(f) Now generate a test set consisting of another 200 observations. Fit a k-nearest neighbors model on the training set, for a range of values of k from 1 to $n/2$. Make a plot that displays the value of $1/k$ on the x-axis, and classification error (both training error and test error) on the y-axis. Display the Bayes error rate (computed in (e)) as a horizontal line. Make sure all axes and curves are properly labeled. Explain your results.

(g) Repeat the above steps (a)-(f), but with $\tau = 0.1$. How do your results differ? Explain your answer.

Problem 3 Consider patient data consisting of the measured amount of an antibody (covariate X) and whether the patient became infected with the flu within a year after this measurement (outcome Y).

(a) Given an example of a prediction task involving this data. What quantities might be reported in an analysis involving this prediction task?

(b) Given an example of an inference task involving this data. What quantities might be reported in an analysis involving this inference task?

Problem 4 Consider two functions, $f_1(X) := 1+X$ and $f_2(X) := 1+X+X^2+X^3$, where $X \in \mathbb{R}$. Suppose that for $k = 1, 2$, we observe $Y^{(k)} = f_k(X) + \varepsilon$, where $\varepsilon \sim \text{Norm}(0, 1)$. Our goal is to estimate f_1 and f_2 using $Y^{(1)}$ and $Y^{(2)}$, respectively.

In what follows, we will generate the observations of X according to $X \sim \text{Norm}(0, 1)$.

(a) During Lecture #1, we talked about the "irreducible error" associated with the expected prediction error, in the context of our discussion of the bias-variance trade-off. Write down the mathematical expression for the "irreducible error" that we saw in lecture.

(b) What is the irreducible error associated with estimating $f_1(\cdot)$ from $Y^{(1)}$? What is the irreducible error associated with estimating $f_2(\cdot)$ from $Y^{(2)}$?

(c) For each $k = 1, 2$:

- Simulate a test dataset consisting of 10,000 independent $(X_i, Y_i^{(k)})$ pairs.
- Simulate $1,000$ training datasets, each of which consists of 5,000 independent $(X_i, Y_i^{(k)})$ pairs.

  (Note: in a real data analysis, we would of course only have access to one training dataset. However, in this homework problem we will work in an imagined scenario where we have access to a HUGE number of training datasets, in order to strengthen our understanding of the bias-variance trade-off.)
- On each training dataset, fit a linear model $\hat{f}^{(k)}$ to predict $Y^{(k)}$ (you can use, e.g., the `lm` function in **R** to fit this model), with an intercept and the single feature X.
- For each value x_0 in the test dataset, estimate the squared bias $(\mathrm{E}[\hat{f}^{(k)}(x_0)] - f_k(x_0))^2$, using the 1,000 training sets.
- For each value x_0 in the test dataset, estimate the variance $\mathrm{Var}[\hat{f}^{(k)}(x_0)]$, using the 1,000 training sets.

- Recall that the expected prediction error at a given point x_0 is the sum of the (i) irreducible error, (ii) squared bias of $\hat{f}^{(k)}(x_0)$, and (iii) variance of $\hat{f}^{(k)}(x_0)$. Combine your answers to the previous subproblems in order to estimate the expected prediction error at each value x_0 in the test dataset. Plot the expected prediction error as a function of x_0.

(d) Repeat (c), but this time fit a linear model containing an intercept, and features X, X^2, and X^3.

(e) Repeat (c), but this time fit a linear model containing just an intercept.

(f) Compare and contrast your results in (c)–(e). Interpret your results in terms of the bias-variance trade-off.