

DS 3010 HOMEWORK 4

Instructions: Homework is to be submitted on Canvas by the deadline stated above. Please clearly print your name and student ID number on your HW.

Show your work (including calculations) to receive full credit. Please work hard to make your submission as readable as you possibly can - **this means no raw R output or code in your solutions.** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework *as a separate file* (for example, a .R file, text file, Word file, or .Rmd are all acceptable). You should mark sections of the code that correspond to different homework problems using comments (e.g. ##### Problem 1 #####).

Problem 1: Concept Review

- (a) You fit a multiple linear regression model with 3 predictors. The output is seen below:

```
> summary(lm(y~x1+x2+x3))
```

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:

	Min	1Q	Median	3Q	Max
	-1.82903	-0.55227	0.07452	0.46413	2.39613

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.87750	0.08782	21.380	<2e-16 ***
x1	2.95472	0.02991	98.774	<2e-16 ***
x2	1.04783	0.02907	36.042	<2e-16 ***
x3	-1.4306	1.47332	-0.971	0.334

To gain insight into the model, you propose testing whether the regression coefficient related to X_3 is statistically significantly different from 0. Your collaborator (who knows nothing about data science) says that is a waste of time. They state that it's obvious that $\hat{\beta}_3$ is different from 0 because the estimate is not 0! So why bother doing any hypothesis testing? Provide a clear and compelling explanation for why hypothesis testing is valuable here and why data scientists rely on hypothesis test to draw meaningful conclusions about the data.

- (b) This same colleague claims that if you knew the true $f(X)$, then we could obtain perfect prediction. Do you agree or disagree? Explain clearly.

For the following statement, evaluate whether they are True or False and **justify your answer**.

- (c) The expected test MSE is defined as: $E(y_0 - \hat{f}(x_0))^2$. Here y_0 is from our training set and $\hat{f}()$ is the model we built from our training set. We evaluate $\hat{f}(x_0)$ on the x_0 values from our test set.
- (d) The bias-variance decomposition tells us that sometimes reducing the complexity of our model (for example, removing a predictor), can actually improve our expected test MSE.
- (e) **(practice)** The expected test MSE can be smaller than the irreducible error.
- (f) **(practice)** The training MSE can be smaller than the irreducible error.

Problem 2: Bias-variance decomposition

- (a) On a single plot, provide a sketch of typical curves for (squared) bias, variance, expected test MSE, training MSE, and the irreducible error as we go from less flexible statistical learning methods towards more flexible methods. The x -axis should represent the amount of flexibility in the method, and the y -axis should represent the values for each curve. There should be 5 curves. Make sure to label each one.
- (b) Define in plain language (so that a non-data scientist can understand) what the quantities expected test MSE, training MSE, bias, variance and irreducible error mean.
- (c) **(Practice)** Explain why each of the five curves has the shape displayed in part (a).
- (d) Suppose we collect a data set with $n = 100$ observations, consisting of a single predictor X and a quantitative response Y . We fit two models to the data:
 - a linear regression model, and
 - a cubic regression model: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$.

Assume that the true relationship between X and Y is **linear**.

Compare the **training MSE** of the linear regression model and the cubic regression model. Would you expect one to be lower than the other, or is there not enough information to tell? Explain your reasoning.

- (e) Using the same setup as in part (d), now compare the **test MSE** of the linear regression model and the cubic regression model. Would you expect one to be lower than the other, or is there not enough information to tell? Explain your reasoning.

Problem 3: (Practice) Optimal degree for Boston dataset

We'll use the **Boston** dataset. Our response is **medv** and our predictor is **lstat**. The setup is the same as our in-class activity (**bv_example.R**): we want to determine the optimal level of flexibility for our model. But now instead of just doing a one time split into a training/test set, use 5-fold CV to determine the optimal degree d for a polynomial regression model. **Write your own code**

to do so - that means you should not be using other another package's functions to implement this (other than `caret` to create your folds). You can set $d = 1$ through 9. Report your 5-fold CV error for each model you considered and show a plot here. What degree did you choose?

Problem 4: Cross-validation

- (a) Explain how k -fold cross-validation is implemented.
- (b) What are the advantages and disadvantages of k -fold cross-validation relative to:
 - i. The validation set approach?
 - ii. LOOCV?
- (c) For the following questions, we will perform cross-validation on a simulated data set. Generate a simulated data set such that $Y = X - 2X^2 + \epsilon$, with $\epsilon \sim N(0, 1^2)$. Fill in the following code:

```
set.seed(1)
x = rnorm(100)
error = ??
y = ??
```

- (d) Set a random seed, and then compute the LOOCV errors that result from fitting the following 4 models using `lm` and `poly`:

$M1$: a linear model with X

$M2$: a polynomial regression model with degree 2

$M3$: a polynomial regression model with degree 3

$M4$: a polynomial regression model with degree 4

You may find it helpful to use the `data.frame()` function to create a single data set containing both X and Y .

- (e) **(Practice)** Repeat the above step using another random seed, and report your results. Are your results the same as what you got in (d). Why?
- (f) Which of the models in (d) had the smallest LOOCV error? Is this what you expected? Explain your answer.
- (g) **(practice)** LOOCV will provide approximately unbiased estimates of the test error. Provide some intuition as to why.
- (h) There is bias-variance trade-off associated with the choice of k in k -fold cross-validation. As k increases in k -fold cross-validation, does the bias of the test error estimate increase, decrease, or stay the same? Explain why.
- (i) As k increases in k -fold cross-validation, does the variance of the test error estimate increase, decrease, or stay the same? Explain why.

End of assignment.