Homework 3 - Written Answer Key

Question 1:

(Question)

If your training data set D consists of $N = 100$ points, $(\mathbf{x}^{(i)}, y^{(i)})$ for $i = 1, \ldots, N$ where $\mathbf{x}^{(i)}$ consists of a single feature (i.e. $\mathbf{x}^{(i)} = [x_1^{(i)}]$), and we fit two linear regression models

model 1: $w_0 + w_1 x_1 + w_2 x_1^2$.

model 2: $w_0 + w_1 x_1$

- When using model 1, what transformation function[1] $\Phi(\mathbf{x})$ would we use? (Include x_0 as part of your transformation.)
- For model 1, express RSS (residual sum of squares) in terms of $\Phi(\mathbf{x})$.
- Suppose the true relationship between $\mathbf{x}$ and y is $y = w_0 + w_1 x_1 + \epsilon$. Which model would we expect to have a smaller training error, E_{in}? Which model would we expect to have a smaller generalization error, E_{out}? Explain.
- Suppose the true relationship between $\mathbf{x}$ and y is $y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \epsilon$. Which model would we expect to have a smaller training error, E_{in}? Which model would we expect to have a smaller generalization error, E_{out}? Explain.

(Answer(s))

1.

For model 1, we'd use the transformation function:

$$\Phi(x) = \begin{bmatrix} 1 \\ \Phi_1(x) \\ \Phi_2(x) \end{bmatrix} = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}$$

2.

Originally, $RSS = \sum_{i=1}^{N} (\hat{y} - y^{(i)})^2$
Now, we just translate this summation given the transformed x

$\tilde{w} = w$ vector for the data transformed by $\Phi(x)$
$\Phi(x)$ = transformation for model 1 (from part 1)

$RSS = \sum_{i=1}^{N} (\tilde{w}^T \Phi(x^{(i)}) - y^{(i)})^2$

3.

We would expect model 1 to have smaller E_{in} because it is raised to a higher order (has more parameters) than model 2, allowing it to better model the noise in the training data. This makes model 1 more capable of reducing error in the training set (Model 1 overfits the true model).

However, we would expect model 2 to have smaller E_{out} because it is of the same class as the true model. Model 1, on the other hand, is more likely to model noise from the training data, making its accuracy on new data less than model 2 's.

4.

Both model 1 and model 2 would suffer from underfitting the true model because they do not consider parameters x_2 and x_3. Thus, the difference in E_{in} and E_{out} between models 1 and 2 would result from the same reasons in the above part; model 1 would have smaller E_{in} while model 2 would have smaller E_{out}.

Question 2:

(Question)

A medical researcher wishes to evaluate a new diagnostic test for cancer. A clinical trial is conducted where the diagnostic measurement y of each patient is recorded along with attributes of a sample of cancerous tissue from the patient. Three possible models are considered for the diagnostic measurement:

- Model 1: The diagnostic measurement y depends linearly only on the cancer volume.
- Model 2: The diagnostic measurement y depends linearly on the cancer volume and the patient's age.
- Model 3: The diagnostic measurement y depends linearly on the cancer volume and the patient's age, but the dependence (slope) on the cancer volume is different for two types of cancer – Type I and II. (Hint: Use a variable x_3 which is assigned the value 1 if the cancer is Type I, and x_3 has the value 0 if the cancer is of Type II.)

(a) Define variables for the cancer volume, age and cancer type and write a linear model for the predicted value $\hat{y}$ in terms of these variables for models 1 & 2 above.
(Do not turn in this question): Do the same for model 3. For Model 3, you will want to use one-hot coding as mentioned above.

(b) What are the number of parameters in model 1 & 2? Which model is the most complex?

(c) Since the models in part (a) are linear, given training data, we should have $\hat{\mathbf{y}} = X\mathbf{w}$ where $\hat{\mathbf{y}}$ is the vector of predicted values on the training data, X is a design matrix (feature matrix) and $\mathbf{w}$ is the vector of parameters. To test the different models, data is collected from 100 patients. The records of the first three patients are shown below:

Patient ID	Measurement y	Cancer type	Cancer volume	Patient age
12	5	I	0.7	55
34	10	II	1.3	65
23	15	II	1.6	70
⋮	⋮	⋮	⋮	⋮

For model 1 in part (a), based on this data, what are the first three rows of the matrix X?
For model 2 in part (a), based on this data, what are the first three rows of the matrix X?
(Do not turn in this question): For model 3 in part (a), based on this data, what are the first three rows of the matrix X?

(d) To evaluate the models, 10-fold cross validation is used with the following results.

Model	training MSE	test MSE
1	2.0	2.01
2	0.7	0.72
3	0.65	0.74

Which model should be selected?

(Answer(s))

1.

We define x_1 to be the variable for cancer volume; x_2 to be the variable for the patient's age; and x_3 to be the variable for the cancer type.

For Model 1:

$$\hat{y}^{(i)} = w_0 + w_1 x_1^{(i)}$$

For Model 2:

$$\hat{y}^{(i)} = w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)}$$

For Model 3:

$$\hat{y}^{(i)} = w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)} + w_3 x_3^{(i)}$$

2.

Model 1 uses two parameters and Model 2 uses three. Model 2 is more complex out of these two models because it has more parameters.

3.

In Model 1, the first 3 rows of X are:

$$\begin{bmatrix} 1 & 0.7 \\ 1 & 1.3 \\ 1 & 1.6 \end{bmatrix}$$

In Model 2, the first 3 rows of X are:

$$\begin{bmatrix} 1 & 0.7 & 55 \\ 1 & 1.3 & 65 \\ 1 & 1.6 & 70 \end{bmatrix}$$

In Model 3, the first 3 rows of X are:

$$\begin{bmatrix} 1 & 0.7 & 55 & 1 \\ 1 & 1.3 & 65 & 0 \\ 1 & 1.6 & 70 & 0 \end{bmatrix}$$

4.

Generally, the model with the lowest validation MSE should be selected. This means we select Model 2.

Question 3:

(Question)

3. Suppose you trained your data[2] on three different models and then plotted how well the different fitted models performed with varying amounts of data:

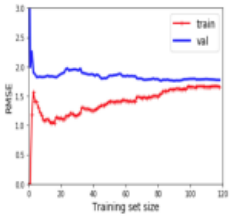


Figure 1: *A*

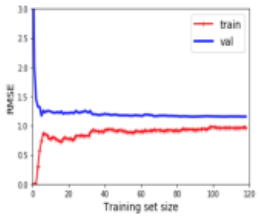


Figure 2: *B*

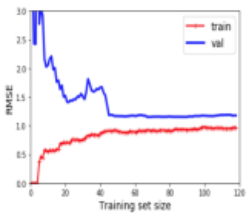


Figure 3: *C*

What can you say about overfitting and underfitting? What can you say about the number of examples and the fit of the model?

(Answer(s))

Figure A likely displays a model that is underfitting the data because the errors converge at a relatively high value.

Figure B is probably neither underfitting nor overfitting the data because the errors converge both at a relatively low error and with a relatively low number of examples.

Figure C likely overfits the data because although the errors converge at a relatively low value, a large number of examples is needed to do so.

The number of examples needed to converge the errors is highest in Figure C, lower in Figure B, and lowest in Figure A. However, the best trade-off between convergence and error is in Figure B, so this model most likely best fits the data. It is important to note that for all models, error is expected to converge at a lower value (no matter how much lower) given a larger training set because the more data a model is given, the less likely it is that the model will be affected by noise.

Question 4:

(Question)

What are the advantages and disadvantages of a very flexible (versus a less flexible) approach[3] for regression or classification ? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred ?[4]

(Answer(s))

Advantages of a very flexible model: it has lower bias; it fits better with non-linear problems and complex systems; it fits better with a large set of data. Disadvantages: it has higher variance; it overfits easily when the dataset is small OR the variance of the noise is extremely high; it has more parameters to train

As the pros and cons discussed above, a more flexible approach might be preferred when the data has non-linear characteristics, the dataset is relatively large, the variance of noise is low, the prediction is more valued than inference, and etc.

A less flexible approach might be preferred when the dataset shows linear characteristics, small size of training data, the variance of noise is extremely high, more interpretability is desirable, and etc.

Read more here: https://spr.com/data-science-back-basics-flexible/

Question 5:

(Question)

Consider a binary classification problem ($y \in \{0, 1\}$), where the iid examples

$$D = (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(N)}, y^{(N)})$$

are divided into two disjoint sets D_{train} and D_{val}.

- Suppose you fit a model h using the training set, D_{train}, and then estimated its error using the validation set, D_{val}. If the size of D_{val} was 100 (i.e. $|D_{val}| = 100$), how confident are you the true error of h is within 0.1 of its average error on D_{val}?
- Repeat the previous question where now $|D_{val}| = 200$ (i.e you have 200 examples in your validation set).
- (Do not turn in this question) When dividing the set of examples D into two sets, how large should you make D_{val} if you wanted to be 90% confident that the true error of h is within 0.05 of the average error your hypothesis makes on D_{val}.

(Answer(s))

1.

$$\delta = 2e^{-2\epsilon^2 K/(b-a)^2}$$

$$\text{100 Examples: } \delta = 2e^{-2(0.1)^2(100)}$$

$$\delta = 0.27$$

$$\text{Certainty} = 1 - \delta = 0.73$$

2.

$$\text{200 Examples: } \delta = 2e^{-2(0.1)^2(200)}$$

$$\delta = 0.04$$

$$\text{Certainty} = 1 - \delta = 0.96$$

3.

$$1 - \delta = 0.9$$

$$\delta = 0.1$$

$$0.1 = 2e^{-2\epsilon^2 K/(b-a)^2}$$

$$0.05 = e^{-2(0.05)^2 K}$$

$$K \approx 599$$

Question 6:

(Question)

Suppose you are given the following dataset, where the target variable is MED:

RM	RAD	DIS	MED
6.6	1	4.0	24.0
6.4	2	5.0	21.6
7.2	2	5.0	34.7
6.4	2	5.0	21.6
7.2	2	5.0	34.7

Using the data above, write the *equation* derived in the lecture notes (slide 17 in Topic 3 Regularization Grad spring 2022 PDF') to compute the closed form solution for ridge regression where $\lambda = 0.1$. You do not need to actually calculate the coefficient vector - just set up the formula using the numbers given above.

(Answer(s))

$$w_{ridge} = (X^T X + N\lambda I')^{-1} X^T y$$

$$w_{ridge} = \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 6.6 & 6.4 & 7.2 & 6.4 & 7.2 \\ 1 & 2 & 2 & 2 & 2 \\ 4.0 & 5.0 & 5.0 & 5.0 & 5.0 \end{bmatrix} \begin{bmatrix} 1 & 6.6 & 1 & 4.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \end{bmatrix} + (5)(0.1) \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 6.6 & 6.4 & 7.2 & 6.4 & 7.2 \\ 1 & 2 & 2 & 2 & 2 \\ 4.0 & 5.0 & 5.0 & 5.0 & 5.0 \end{bmatrix} \begin{bmatrix} 24.0 \\ 21.6 \\ 34.7 \\ 21.6 \\ 34.7 \end{bmatrix}$$

Question 7:

(Question)

Write the gradient descent algorithm (vectorized or not) for ridge regression

(Answer(s))

for _ in range(num_iters):

$$w = w - \alpha(\frac{2(X^T X w - X^T y)}{N} + 2\lambda I' w)$$

Question 8:

(Question)

(Do not turn in this question) For the following training examples in question 3 from homework assignment 2, write the closed form solution for ridge regression when $\lambda = 0.1$.

(Answer(s))

$$w_{ridge} = (X^T X + N\lambda I')^{-1} X^T y$$

$$w_{ridge} = (\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} + (4)(0.1) \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix})^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 3 \\ 7 \end{bmatrix}$$

Question 9:

(Question)

(Do not turn in this question) For each of parts 9a through 9d, indicate whether we would generally expect the performance of a flexible (complex) hypothesis class (aka complex model class) to be better or worse than an inflexible (simple) hypothesis class (aka simple model)..[5] Justify your answer.

(a) The sample size N is extremely large, and the number of features d is small.

(b) The number of features d is extremely large, and the number of observations N is small.

(c) The relationship between the features and labels is highly non-linear ?

(d) The variance of the noise, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high.

(Answer(s))

A.

An inflexible model will better because it is truer to the true relationship.

B.

We'd expect a flexible model to do better because it is closer to the true relationship.

C.

We'd expect a flexible model to do better because it is closer to the true relationship.

D.

We'd expect the flexible model to do worse because it will begin modelling the noise to a significant extent.

Question 10:

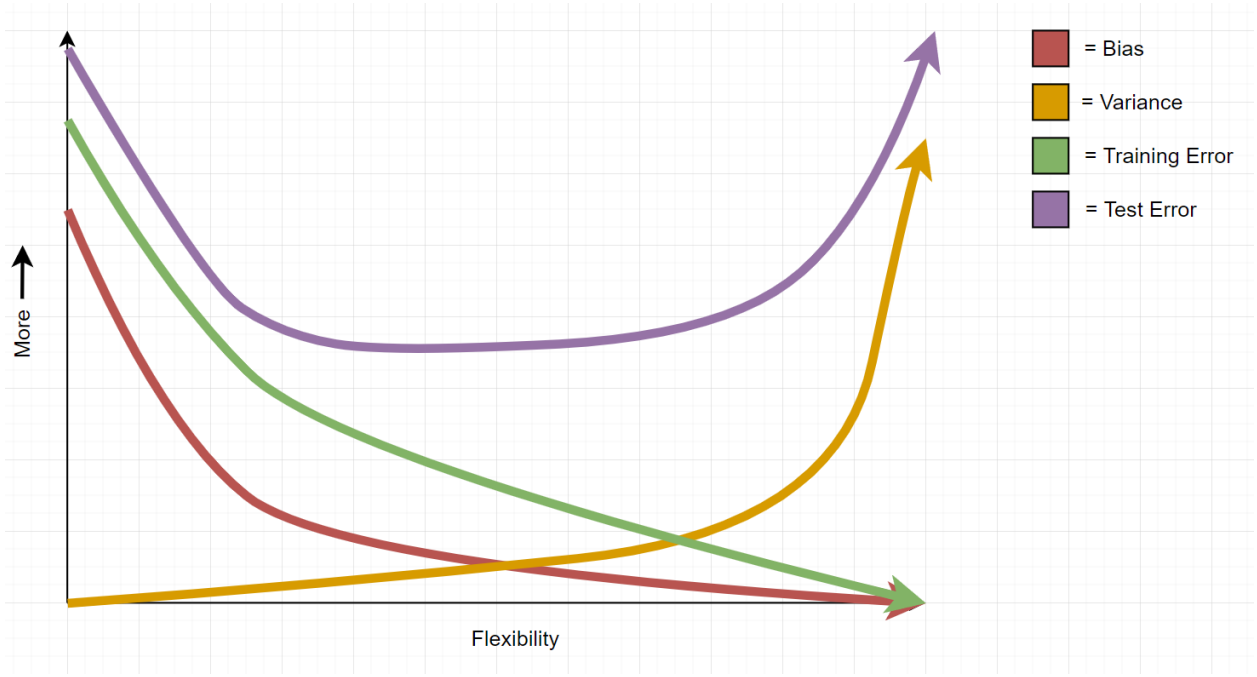
(Question)

(Do not turn in this question) Bias-variance decomposition

- Provide a sketch of typical (squared) bias, variance, training error, test error on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. 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 four curves. Make sure to label each one.[6]
- Explain why each of the curves has the shape displayed in part (a).

(Answer(s))

-



- Starting with the Bias curve: it starts off high as the flexibility is low because the model is unable to capture all the information it needs to represent the true relationship of the data. Then it gets lower as the model gains the ability to do so.

  The Variance curve: it starts off low because when the model is less complex, the amount the model will change given different training data is low. As the model gains complexity, it will change more given different training data.

  The Training Error curve: This curve is very similar to the bias curve, for similar reasons. As the flexibility of the model increases, it will be able to better fit the training model (but not necessarily for new data).

  The Test Error curve: This starts off high because models that do not have enough complexity to model the true relationship will not do well on new examples. As the model approaches the real complexity of the true relationship, error will get lower and lower. Then, once the complexity passes the real complexity of the true relationship, the error will rise again because it will model the noise in the training data.