4 Bias-Variance Dilemma

Exercise 4.1

While you fit a Linear Model to your data set. You are thinking about changing the Linear Model to a Quadratic one (i.e., a Linear Model with quadratic features $\phi(x) = [1,\ x,\ x^2]$). Which of the following is most likely true:

1. Using the Quadratic Model will decrease your Irreducible Error;
2. Using the Quadratic Model will decrease the Bias of your model;
3. Using the Quadratic Model will decrease the Variance of your model;
4. Using the Quadratic Model will decrease your Reducible Error.

Provide motivations to your answers.

Exercise 4.2

Which of the following is/are the benefits of the sparsity imposed by the Lasso?

1. Sparse models are generally more easy to interpret;
2. The Lasso does variable selection by default;
3. Using the Lasso penalty helps to decrease the bias of the fits;
4. Using the Lasso penalty helps to decrease the variance of the fits.

Provide motivation for your answer.

Exercise 4.3

We estimate the regression coefficients in a linear regression model by minimizing ridge regression for a particular value of λ. For each of the following, describe the behaviour of the following elements as we increase λ from 0 (e.g., remains constant, increases, decreases, increase and then decrease):

1. The training RSS;

2. The test RSS;
3. The variance;
4. The squared bias;
5. The irreducible error.

Exercise 4.4

Suppose that Figure 4.1 is showing the Test error of K-NN obtained by using different values for K.

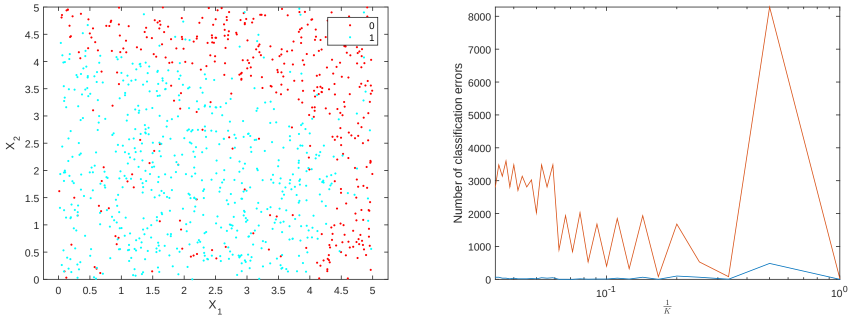

Figure 4.1: Dataset and corresponding error for different K in the K-NN classifier.

Which of the following is most likely true of what would happen to the Test Error curve as we move $\frac{1}{K}$ further above 1?

1. The Test Errors will increase;
2. The Test Errors will decrease;
3. Not enough information is given to decide;
4. It does not make sense to have $\frac{1}{K} > 1$;

Exercise 4.5

Comment on advantages and drawbacks of the following choices:

1. Increase the model complexity and fix number of samples;
2. Increase the number of the samples and fix model complexity.

Exercise 4.6

Assume to have two different linear models working on the same dataset of $N = 100$ samples.

- The first model has $k_1 = 2$ input, considers linear features and has a residual sum of squares of $RSS_1 = 0.5$ on a validation set;
- The second model has $k_2 = 8$ input, considers only quadratic features and has a residual sum of squares of $RSS_2 = 0.3$ on a validation set;

Which one would you choose? Why? Recall that the F-test for statistics for distinguish between linear models is:

$$\hat{F} = \frac{N - p_2}{p_2 - p_1} \frac{RSS_1 - RSS_2}{RSS_2} \sim F(p_2 - p_1, N - p_2),$$

where p_1 and p_2 are the two parameters of the two models and $F(a, b)$ is the Fisher distribution with a and b degrees of freedom.

Exercise 4.7

Which techniques would you consider for evaluate the performance of a set of different models in the case we have:

1. A small dataset and a set of simple models;
2. A small dataset and a set of complex models;
3. A large dataset and a set of simple models;
4. A large dataset and a trainer with parallel computing abilities.

Justify you choices.

Exercise 4.8

Suppose you have a dataset and you decided to use all the samples to train your model, including the selection of the parameters of your model and the features you want to consider. What are the problems and issues arising if you use this methodology?

Which procedure a ML scientist should follow?

Answers

Answer of exercise 4.1

1. NO Changing the model does not influence the Irreducible Error.
2. YES We are considering a larger hypothesis space thus it is most likely that the Bias will decrease.
3. NO A more complex model is likely to increase the variance w.r.t. a simpler one.
4. MAYBE If the model is able to reduce the bias and, at the same time, the variance is not increasing too much, we are providing a more accurate model.

Answer of exercise 4.2

1. YES since they provide a clear distinction between those input which are more meaningful (with non–zero parameters) and those which are less relevant (zero parameters);
2. YES since it only includes in the model those input which are meaningful for the model;
3. NO since we are reducing the number of parameters considered in the model, thus the hypothesis space is reduced. This will likely increases the bias of the obtained model;
4. YES since the regularization action helps in avoid overfitting and using unnecessary complex models.

Answer of exercise 4.3

1. INCREASES: by increasing λ, we are forced to use simpler models. This means that training RSS will steadily increase because we are less able to fit the training data exactly.
2. DECRESASES AND THEN INCREASES: at first, the test RSS improves (decreases), because we are less likely to overfit our training data. Eventually, we will start fitting models that are too simple to capture the true effects and the test RSS will start to increase.
3. DECREASES: increasing λ will imply that we are fitting simpler models, which reduces the variance of the fits.

4. INCREASES: by increasing λ we fit simpler models, which likely have larger squared bias.
5. REMAINS CONSTANT: increasing will have no effect on irreducible error, since it has nothing to do with the model we fit.

Answer of exercise 4.4

Clearly the K-NN classifier can only contemplate integer values for the parameter K, thus it does not make sense to decrease the K below 1. The only true statement is the fourth one.

Answer of exercise 4.5

1. Increase the model complexity:
   - Advantages: the Hypothesis space we are considering is larger, thus it may happen that the bias becomes smaller than before.
   - Drawbacks: a more complex model will likely have an increased variance.
2. Increase the number of samples:
   - Advantages: we decrease the variance of the model we are considering.
   - Drawbacks: we need to obtain these samples (which may be expensive) and the training phase might become more and more time consuming (as well as the test one if you are considering non-parametric methods).

Answer of exercise 4.6

The first model has $p_1 = k_1 + 1 = 3$ parameters and the second one $p_2 = k_2 + \frac{k_2(k_2-1)}{2} + 1 = 8 + 28 + 1 = 37$ parameters. To evaluate if two regressive models, under the assumption that the noise is i.i.d. zero mean Gaussian random variable, we can use an F-test. In the specific, we know that the test statistic is:

$$\hat{F} = \frac{N - p_2}{p_2 - p_1} \frac{RSS_1 - RSS_2}{RSS_2} \sim F(p_2 - p_1, N - p_2),$$

is distributed as an F distribution, with $(p_2 - p_1, N - p_2)$ degrees of freedom. Thus, we can evaluate the p-value of the statistic $\hat{F}$ computed on the two aforementioned models:

$$\hat{F} = \frac{100 - 37}{37 - 3} \frac{0.5 - 0.3}{0.3} = 1.2353,$$

whose p-value is 0.2311. Thus there is no statistical evidence at confidence level lower than $\alpha = 0.2311$ that the second model is better than the first one.

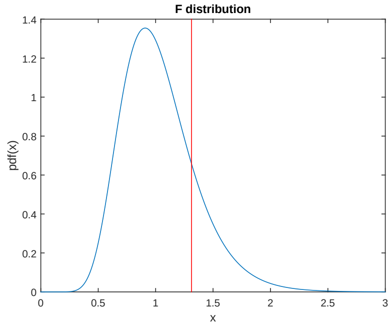

Figure 4.2: Representation of the F-statistic $\hat{F}$ in the case analysed.

Answer of exercise 4.7

1. LOO On a small dataset and for simple models the LOO procedure is not too computationally complex, thus it provides a good approximation (almost unbiased) estimation of the test error.

2. AIC (Adjustment techniques) In this case the training on a smaller dataset may lead to overfitting and thus does not provide any information on the model providing good performance on a newly seen samples. If the model is complex, it might be the case that the LOO procedure is still not feasible.

3. CV You should be able to provide a stable estimates to select your model, but at the same time you can not perform LOO for computational complexity reasons.

4. LOO In this case we are able to perform multiple training at the same time, thus the time required for LOO can be reduced by k times, where k is the number of parallel process you can run at the same time.

Answer of exercise 4.8

If we are not considering a validation set to select the model we will most likely select the most complex model among the ones considered, which could lead to overfitting the training set. We could avoid to use a validation set with early stopping techniques (if we are using a gradient descend technique) or by considering an adjustment technique.

Moreover, we do not have any clue about the error we are likely to have in the case of a newly seen data. Thus, we are not able to provide any performance on the goodness of the prediction our model is going to provide.

A proper ML procedure would consider the split of your data into 3 sets (training, validation, test), where the training set is used to estimate the model, the validation set to select the model and the test set to evaluate the error on unforeseen data. Equivalently, if we use crossvalidation, LOO or adjustment techniques, we should at least save some data to test the performance of the considered method.