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

Figure 4.1 is showing the the training data used to train a K-NN classifier (left) and the training (blue) and test (orange) performances obtained by using different values for K (right).

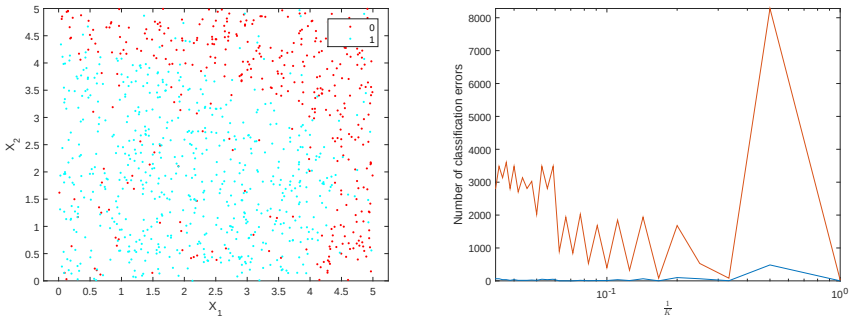

Figure 4.1: Dataset and corresponding error for different K in the K-NN classifier.

Which of the following would most likely 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 the following:

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

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 to evaluate the performances 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.

1. What are the problems and issues arising if you use this methodology?
2. Which procedure a ML scientist should follow?

Exercise 4.9

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

1. Using the Lasso penalty increases the variance of the fits.
2. It does variable selection implicitly.
3. Using the Lasso penalty increases the bias of the fits.
4. Sparse models are generally easier to interpret.

Provide motivation for your answer.

Exercise 4.10

Consider the following statements and tell if they are true or false. Motivate your answers.

1. The computation of the bias-variance decomposition is possible only theoretically. No algorithm provides an explicit decomposition of the twos.
2. An error which is comparable on the training and the test, but larger than what is required by the application, means that the used method has a large variance.
3. The cross-validation error provides slightly larger estimates of the prediction error on newly seen data.
4. If a model results in being too complex, to solve the problem we need to carefully remove some of the input features.

Exercise 4.11

Tell if the following statements about the bias-variance dilemma (and related topics) are true or false. Motivate your answers.

1. If the performance on the training and the test sets are getting almost the same as we are using more and more data for training, but both are worse than desired, the model might be too simple for the task.
2. Given a fixed training set, increasing the complexity of the model always improves its generalization capabilities.
3. It is a good idea to increase the number of samples used for training if we decided to increase the model complexity.
4. Adding new features to the model might help if the model has a bias with respect to the real process generating data.
5. The use of cross-validation decreases the variance of the model.

6. The use of the error on a validation set is suggested when we have a large dataset and we want to discriminate the performance of a set of (computationally) simple models.

Exercise 4.12

Tell if the following statements about the bias-variance dilemma (and related topics) are true or false. Motivate your answers.

1. Regularization techniques are likely to increase the bias of a model.
2. If the performance on the training is matching the desired performance, while the one on a test set is not satisfactory the model might be too complex for the task.
3. Increasing the training set size always improves the model performance.
4. It is a good idea to increase the number of samples used for training if we decided to increase the model complexity.
5. The use of the error on a validation set is suggested when we have a small dataset and we want to discriminate the performance of a set of computationally expensive models.

Exercise 4.13

Tell whether the following statements about training a supervised learning model are true or false. Motivate your answers.

1. The availability of a large dataset might lead to choosing a more complex model.
2. When data is very noisy, it is not a good idea to employ regularization.
3. The larger is the training set, the smaller would be the training error.
4. Increasing the number of features would generally lead to a decrease of the training error.

Exercise 4.14

Tell if the following statements about bias-variance trade-off are true or false. Motivate your answers. Consider a regression problem with input variables x_1, x_2, and x_3, that are linearly independent.

1. In linear regression, if we replace variable x_1 with $x_1 + x_2$, we do not change the bias of the model.

2. In linear regression, if we replace variable x_1 with $x_1^2/100$, we might increase the variance of the model.

3. For an arbitrary model, if we remove variable x_2, we do not increase the variance of the model.

4. For an arbitrary model, if we add variable x_3^2, we cannot increase the bias of the model.

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. Formally, the Bias does not increase.
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. Instead, if using a more complex model is not able to decrease the bias enough and the increase in term of variance is too large, the model reducible error will increase.

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 distributed, we can use an F-test to state if the two RSSs are significantly different from each other. 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.

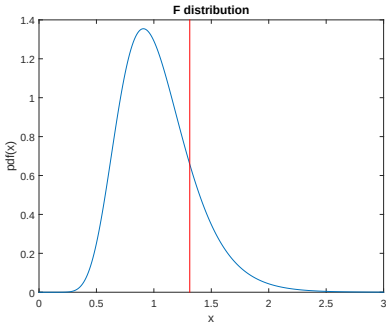

Figure 4.2: Representation of the F-statistic $\hat{F}$ in the case analysed.

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

1. 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.

2. 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.

Answer of exercise 4.9

1. NONE OF THEM: regularization decreases the variance in a model.

2. BENEFIT: since the solution of the Lasso optimization constraints some coefficients to zero it selects automatically those features which are more relevant to the problem.

3. DRAWBACK: even if by regularizing we are decreasing the variance, at the same time we are also likely to introduce some bias due to the fact that the model space is too simple.

4. BENEFIT: having less features which are relevant, it is easier to understand the dynamics of the problem we are solving.

Answer of exercise 4.10

1. TRUE: it is possible only if we know the true process. Some method have an explicit decomposition (e.g., KNN).

2. FALSE: if the two errors are comparable the variance of the model is low. Conversely, since we are not able to reach the desired performance it might be because the real process is more complex than the model. Therefore, we might be in a case where a large bias is present.

3. TRUE: on average it provides a slightly pessimistic estimates of the test error.

4. FALSE: this is an option. We might also resort to regularization (e.g., ridge regression) or to feature extraction (e.g., PCA).

Answer of exercise 4.11

1. TRUE: increasing training data will reduce the variance of the model, hence the error on both training set and test set are likely to depend on bias (in fact, being

the performance bad also on training set we are not overfitting the noise in data). Using a more complex model seems the proper choice.

2. FALSE: the more complex is the model, the more data I usually need to train it in order to avoid overfitting.

3. TRUE: the more complex is the model, the more data I usually need to train it in order to avoid overfitting.

4. TRUE: the bias might be due to some features the model is not using as input.

5. TRUE/FALSE: cross-validation is used only to assess the performance of the model and does not reduce their variance, however through a proper assessment I can hopefully select a model with lower variance.

6. FALSE: with computationally simple models, cross-validation is suggested to get a better estimate of the test error.

Answer of exercise 4.12

1. TRUE: regularization acts as constraints or penalty on parameters variability and, hence, reduces the variance of the model and at the same time increases its bias.

2. TRUE: we might have provided a model which is overfitting the available data. Using a simpler model seems a proper choice.

3. FALSE: if the model selected presents a bias w.r.t. the real process, increasing the training set size would provide only limited benefits. In general, the more data we have, the more we are reducing the variance of the model.

4. TRUE: the more complex is the model, the more data I usually need to train it in order to avoid overfitting.

5. FALSE: with computationally hard models and only a few data one should prefer the use of an adjustment technique, like AIC, to get an estimate of the test error.

Answer of exercise 4.13

1. TRUE If one has more samples, he/she might also consider more complex model to be used on the available data. Recall that the availability of a large amount of data does not mean that the model should be necessarily complex. Instead, the complexity should be appropriate to the process generating data.

2. FALSE Regularization techniques are able to avoid overfitting, therefore they are less prone to model the noise present in the data.

3. FALSE (generally) If we assume that we fix the model, as we include more samples we have that the generally the training error is increasing or remains the same. In some specific case it might happen that the error decreases.

4. TRUE We are more and more prone to overfitting as we increase the number of features, which leads to a decrease of the training error, but to lower generalization capabilities.

Answer of exercise 4.14

1. TRUE. Indeed, since the variables are linearly independent, the linear models representable with $\{x_1, x_2, x_3\}$ are the same as the one representable with $\{x_1 + x_2, x_2, x_3\}$.

2. TRUE. It might be the case that the variance is larger, since the linear models representable with $\{x_1, x_2, x_3\}$ are not comparable with the ones representable with $\{x_1^2/100, x_2, x_3\}$.

3. TRUE. Indeed, regardless the used model, the space of representable models with a subset of the variables, i.e., $\{x_1, x_3\}$ is contained in the space of representable models with all the variables.

4. TRUE. Indeed, adding variables enlarges (or, possibly, leaves unchanged) the space of representable models.