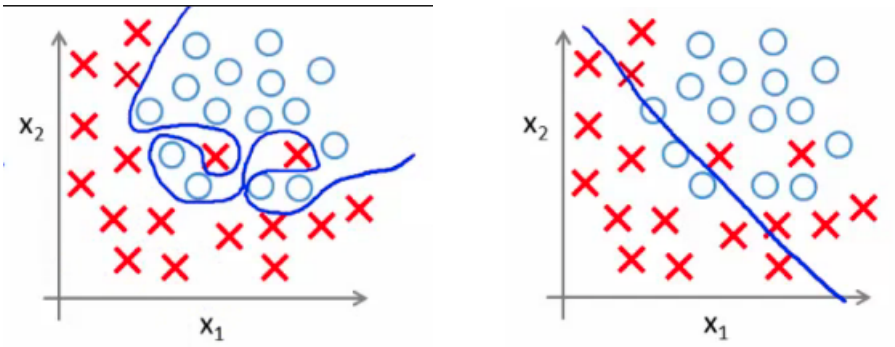
Introduction to Modelling

Exercises

1. Why is it necessary to consider a model's performance on validation or test data?

2. Examine the classification boundaries drawn below. The line drawn is the model, which aims to separate the red X's from the blue O's. Which of these models exhibits high bias? Which exhibits high variance?



3. If my model performs with a high level of accuracy on the training data, but performs substantially worse on a validation dataset, the model likely suffers from [high bias/high variance] *(circle one)*.

4. **(True/False)** When you report the accuracy/error of your model to your sponsor, you should report the accuracy/error from the training data.

5. When is cross-validation *necessary*?

6. **(True/False)** Your practicum project dataset has 100,000 observations that will be used for modeling. Leave-one-out cross validation seems like the most reasonable approach to get an accurate estimate of a model's performance.

7. When you have variables that are null or missing in over 40% of observations, what is the most likely course of action?

8. What does *missing value imputation* mean? Give one example of how you might impute a missing value.

9. What are two unsupervised approaches (i.e. two approaches that do not take into account information from a target variable) to discretizing or binning a numeric variable?

10. How can you interpret the coefficients of a regression when you have a log transformation on both the input variable and the target variable?

11. When transforming variables, what are some of the things you should keep in mind?

List of Key Terms

Training/Validation/Testing Phases

Overfitting

Underfitting

Bias-Variance Tradeoff

k-fold CrossValidation

Leave-one-out Cross Validation

Missing Value Imputation

Discretization (Binning)

Log transformation interpretation