02 - Classification

The aim of classification is to be able to predict the *nominal value* of a class based on the values of a set of *numeric* and or *nominal* attributes. The aim of regression is to be able to predict the *numeric value* of a dependent variable based on a set of *numeric* attributes.

When training a classifier on a particular dataset it is important to avoid over-fitting. This is the process of creating a classifier which is overly specialised on the training data, i.e., it performs well on the data it is trained with, but very poorly on a new unseen data-set. This is done by using a test set to test the performance of the model. As the model is trained both the training error and testing error will decline, but eventually the test error will start to increase as the model becomes over-fitted, at this point the training process should be stopped (Figure 1).

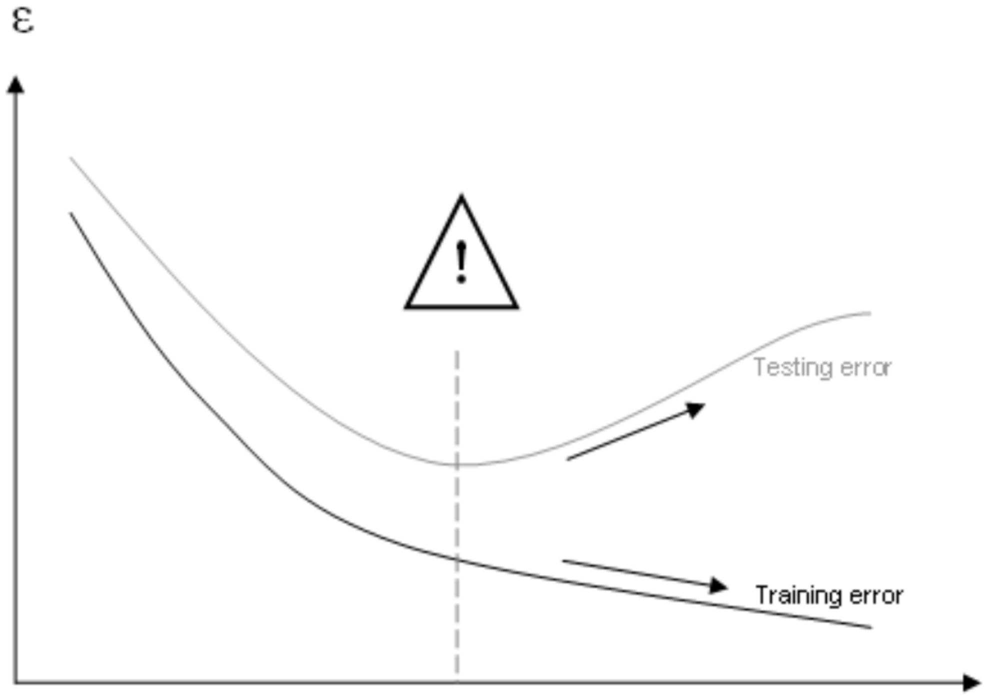

Figure 1: Over-fitting

The distinct stages of designing a classification model are:

1. Collect the raw data.

2. Clean the data (e.g., outlier removal, missing data removal etc.).

3. Pre-process the data (e.g., normalization, standardization, etc.).

4. Determine the type of problem (i.e., classification or regression).

5. Pick an appropriate classifier (e.g., decision tree, Bayes network, etc.).

6. Choose some default parameters for the classifier, the choice of classifier and parameters affect the prediction performance.

7. Pick a training/testing strategy (e.g., percentage split, cross-validation etc.).

8. Train the classifier using your training/testing strategy.

9. Analyse the performance of your model.

10. If your results are unsatisfactory consider altering your model (i.e., changing the classifier, its parameters, and/or your training/testing strategy) and re-training/testing.

11. If your results are satisfactory validate your model on an unseen set of pre-processed data.

In this lab we will focus on classification problems and carry out steps 5 to 9 using default the parameters and training/testing strategy. We will explore some of the other steps in other lab sessions.

Exercises

1. Fire up the Weka software, launch the explorer window. The default tab is the ‘Pre-process’ tab. Click on the ‘Open File’ button and select the iris dataset.
2. Select the ‘Classify’ tab. Under ‘Classifier’ Choose ‘Naïve Bayes’. What type of classifier is this?
3. Under the ‘Test options ’section leave the default selection ‘Cross validation’. This will be explained in a later lecture.
4. Click on ‘Start’. The ‘Classifier output’ pane will be updated. Look at the ‘Confusion Matrix’ at the bottom. What is it telling you? Make sure you understand it. How good do you think the learned classifier is?
5. Under ‘Results list’ you should see a line showing when the model was trained. Right click on it and select ‘Visualise classifier errors. On the plot, errors are shown as squares. How well do you think the classifier performed?
6. Repeat steps 2 to 5, but use ‘J48’ under ‘trees’.
7. Once you have done steps 2 to 5 for J48, right click the output under ‘Result list’ and select ‘Visualize tree’. How do you interpret the decision tree?’
8. Repeat steps 2 to 5, but use ‘JRip’ under ‘rules’.
9. Once you have done steps 2 to 5 for JRip, examine the rules in the ‘Classifier output’ pane. How do you interpret the rules?
10. Make sure you understand the confusion matrix.
11. Make a note of which is the ‘best’ classifier from the three that you have tried.
12. Open the diabetes data set.
13. Repeat steps 2 to 5 using the three classifiers mentioned above. Is it the still the same classifier that comes up with the ‘best’ results?