Machine learning with scikit-learn

Australian
National
University

Intro to machine learning

Australian National University

- In exploratory data analysis, where the aim is often to generate hypotheses, modern machine learning methods based on complex computational models are often used.
- There are two broad classes of machine learning approaches: (i) unsupervised and (ii) supervised.
- Unsupervised methods, such as clustering and principal components analysis, can be directly applied to a dataset to detect natural groupings of the data, with the aim to reveal useful information hidden in the data.
- Supervised methods require not only the data to analyse, but also require the data to be labelled with the true classification or value, typically provided by an expert or gold standard assay etc. This lecture focuses on supervised methods.

Supervised machine learning

Australian National University

- Using this labelled data, predictive models can be built. There are two main uses for such predictive models: (i) to predict the labels or values for future unlabelled data (e.g. train on patient samples with known pathologies, to predict future patients with unknown pathology); (ii) use the predictive model to interpret the data, by studying the ways the predictive model depends on certain features of the data. Some predictive models are inherently simpler and more interpretable than others, but by using advanced perturbative methods (not covered in this intro course) useful information can be extracted from even complex models such a neural networks.

Scikit-learn

Australian National University

- Scikit-learn (https://scikit-learn.org/stable/auto_examples/index.html) is a major python library used for machine learning, both supervised and unsupervised.
- As well as providing implementations of many machine learning methods, it also provides a framework for evaluating the predictive performance of the models.
- We will see an introduction to some of the main capabilities of the library in today's lecture, but explore the above link to get an idea of the full range of analyses supported by the library.

Brief into to supervised machine learning

Australian National University

- In these notes we will cover only some of the key topics required to understand and use predictive models safely.
- Note that it is very easy to get incorrect and biased measures of performance that can lead analyses astray if the caveats covered here are not considered. Many early studies in cancer genomics using predictive models published incorrect optimistically biased results due to misunderstanding how to evaluate these models correctly.

Suggested textbook

Australian National University

- For an in-depth treatment of the theory behind these methods, see Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. "An Introduction to Statistical Learning 2ed" for more details. (For interest only- not required for this course)
- Pdf is freely available at https://www.statlearning.com (a python version will be available this year)

The Supervised Learning Problem

Australian National University

Starting point:

- Outcome measurement Y (also called dependent variable, response, target).
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables).
- In the regression problem, Y is quantitative (e.g price, blood pressure).
- In the classification problem, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
- We have training data (x1,y1),..., (xN ,yN). These are observations (examples, instances) of these measurements.

Objectives

Australian National University

On the basis of the training data we would like to:

- Accurately predict unseen test cases.
- Understand which inputs affects the outcome, and how.
- Assess the quality of our predictions and inferences.

Classification Problems

Australian National University

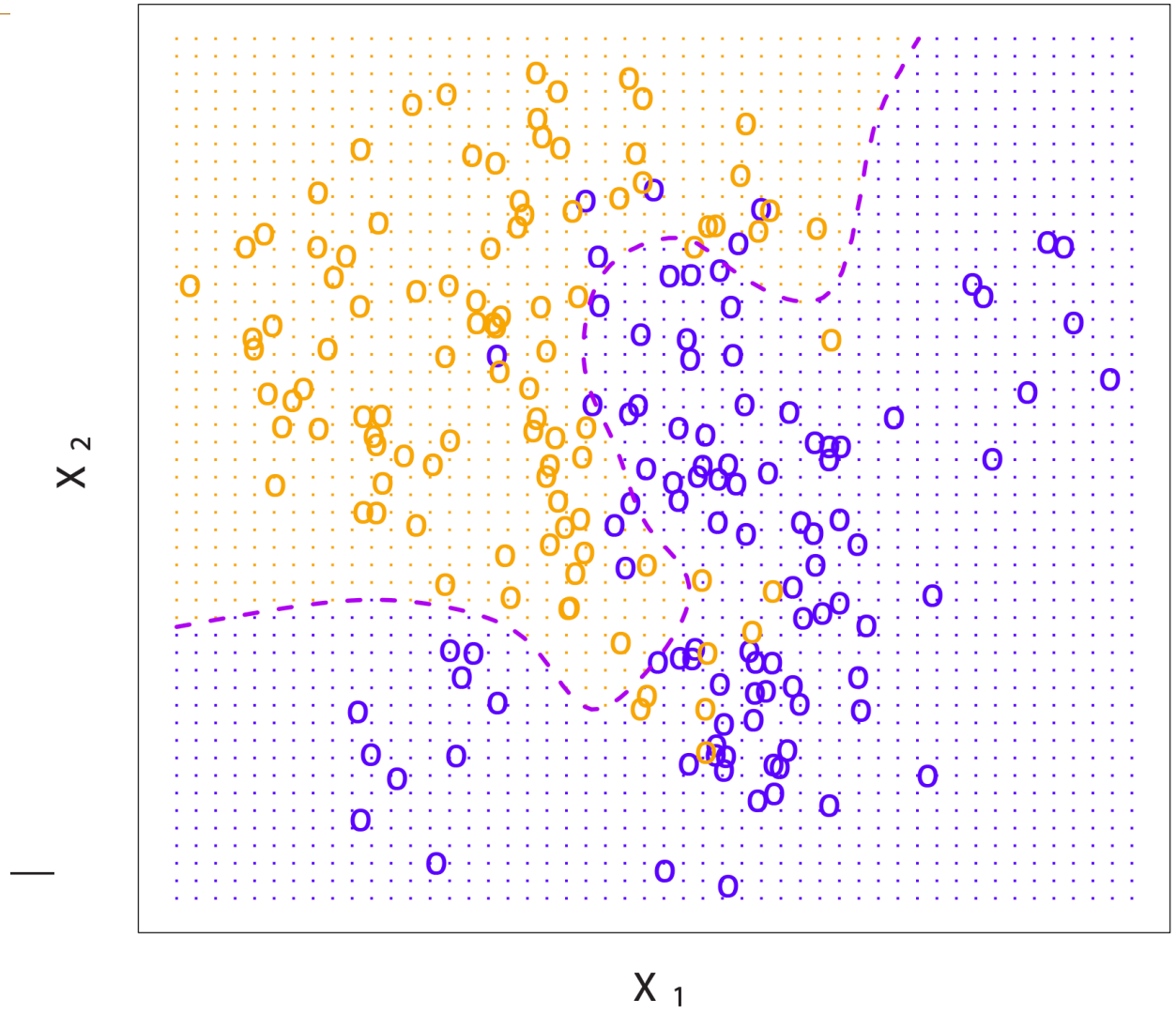
- Here the response variable Y is qualitative — e.g. lab test is one of C= (cancer, normal) (cancer=diease), digit class is one of C= {0,1,..., 9}. Our goals are to:
- Build a classifier C (X) that assigns a class label from C to a future unlabeled observation X .
- Assess the uncertainty in each classification
- Understand the roles of the different predictors among X =(X1,X2,...,X p).

Classification: some details

Australian National University

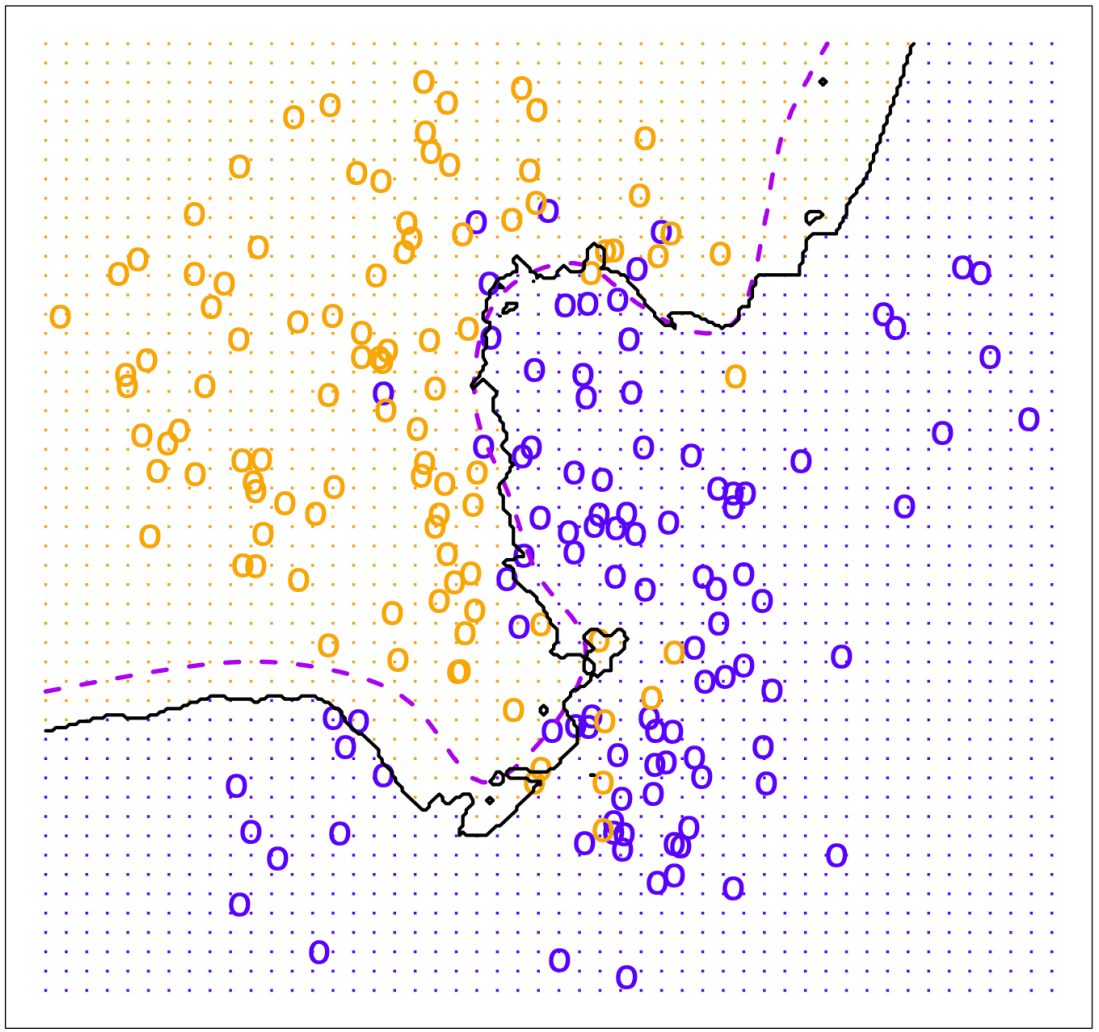
- Typically we measure the performance of the classifier using the misclassification error rate.
- (Other ways of measuring predictive performance such as AUC have advantages, but are not discussed in this intro lecture)

Example: K-nearest neighbors in two dimensions

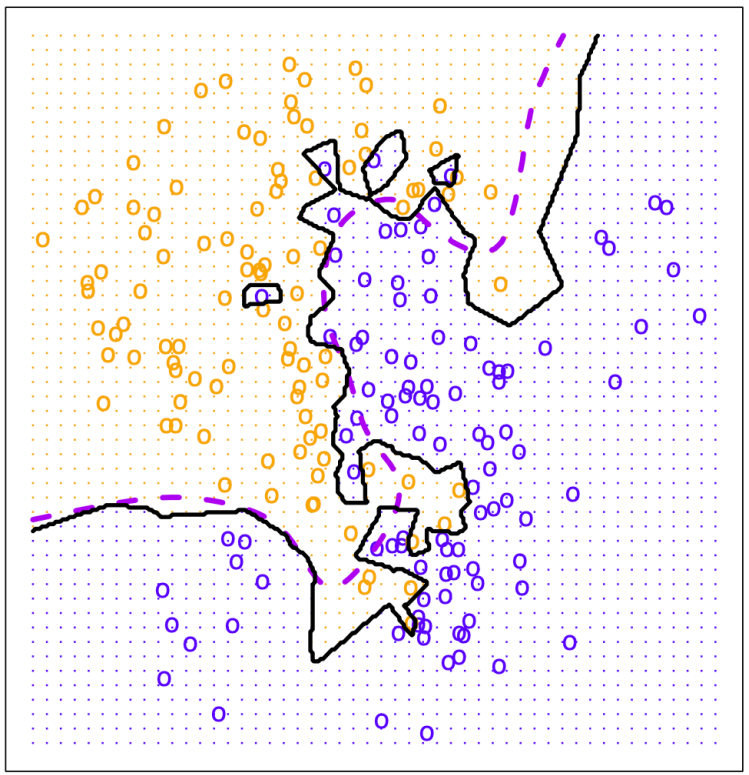


Australian National University

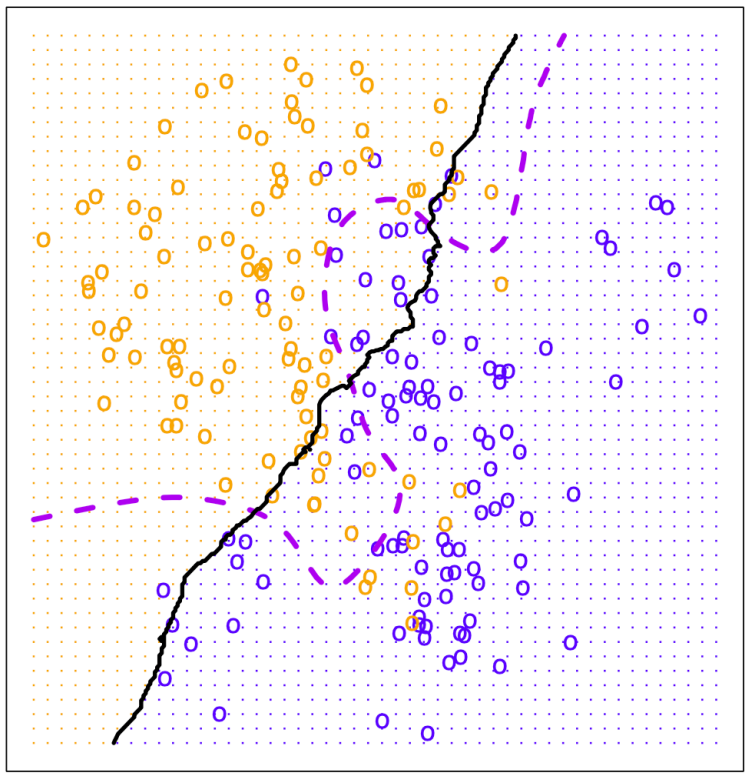
KNN: K=10

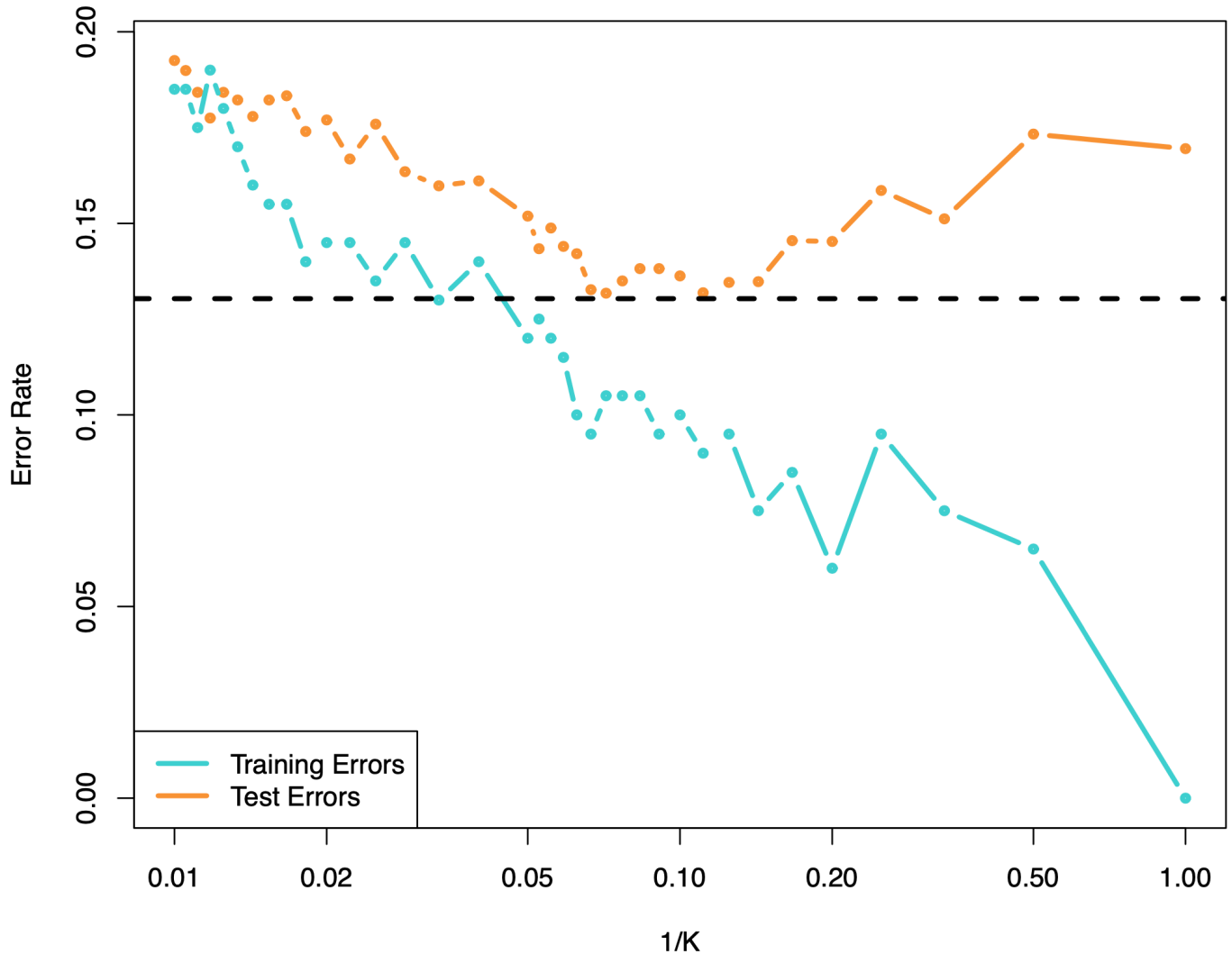


KNN: K = 1



KNN: K = 100




Error Rate
0.00
0.05
0.10
0.15
0.20
0.01
0.02
0.05
0.10
0.20
0.50
1.00
1/K
Training Errors
Test Errors

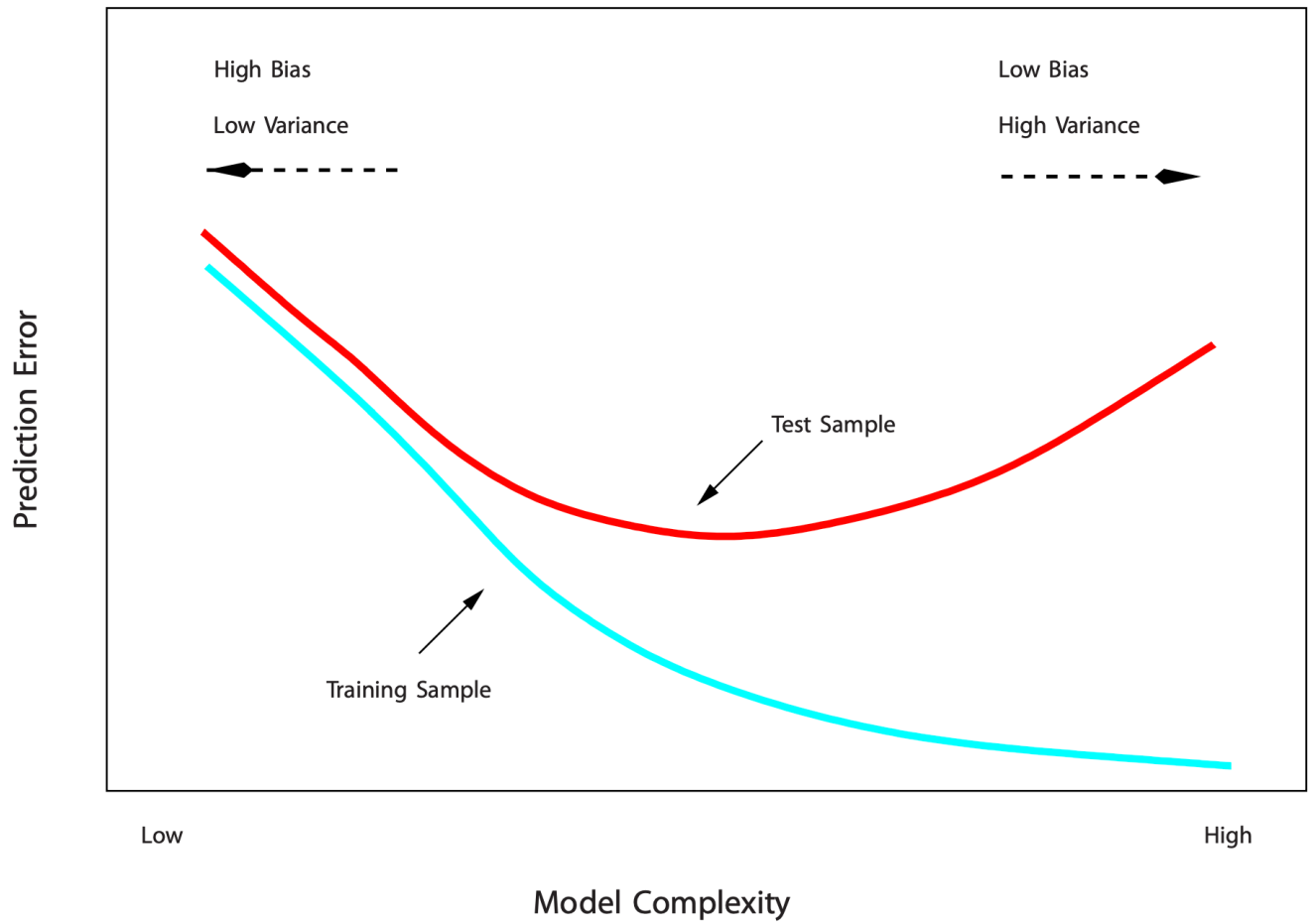
Training Error versus Test error

Australian National University

- Recall the distinction between the test error (measured on an independent test set) and the training error (measured on the same data used to train the model):
- The test error is the average error that results from using a statistical learning method to predict the response on a new observation, one that was not used in training the method.
- In contrast, the training error can be easily calculated by applying the statistical learning method to the observations used in its training.
- But the training error rate often is quite different from the test error rate, and in particular the former can dramatically underestimate the latter.

Australian National University

Training- versus Test-Set Performance

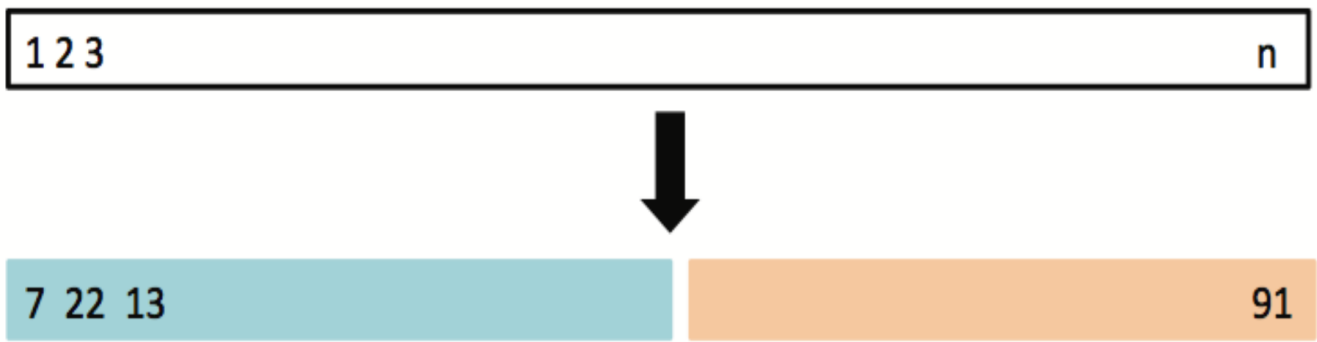


More on prediction-error estimates

Australian National University

- Best solution: a large designated independent test set. Often not available
- Here we instead consider a class of methods that estimate the test error by holding out a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations

The Validation process



A random splitting into two halves: left part is training set, right part is validation set

Validation-set approach

Australian National University

- Here we randomly divide the available set of samples into two parts: a training set and a validation or hold-out set.
- The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set.
- The resulting validation-set error provides an estimate of the test error. This is typically assessed using MSE in the case of a quantitative response and misclassification rate in the case of a qualitative (discrete) response.

Drawbacks of validation set approach

Australian National University

The validation estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.

- In the validation approach, only a subset of the observations — those that are included in the training set rather than in the validation set — are used to fit the model.
- This suggests that the validation set error may tend to overestimate the test error for the model fit on the entire data set.

K -fold Cross-validation

Australian National University

Widely used approach for estimating test error.

- Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.
- Idea is to randomly divide the data into K equal-sized parts. We leave out part k, fit the model to the other
  $K - 1$ parts (combined), and then obtain predictions for the left-out kth part.
- This is done in turn for each part $k = 1, 2, \ldots K$, and then the results are combined.

K -fold Cross-validation in detail

Australian National University

Divide data into K roughly equal-sized parts (K = 5 here)

1	2	3	4	5
Validation	Train	Train	Train	Train

Cross-validation: right and wrong

Australian National University

Consider a simple classifier applied to some two-class data:

- 1. Starting with 5000 predictors and 50 samples, find the 100 predictors having the largest correlation with the class labels.
- 2. We then apply a classifier such as logistic regression, using only these 100 predictors.
- How do we estimate the test set performance of this classifier?
- Can we apply cross-validation in step 2, forgetting about step 1?
- NO!
- This would ignore the fact that in Step 1, the procedure has already seen the labels of the training data, and made use of them. This is a form of training and must be included in the validation process.
- It is easy to simulate realistic data with the class labels independent of the outcome, so that true test error =50%, but the CV error estimate that ignores Step 1 is zero!
- See "An Introduction to Statistical Learning 2ed" for details

Summary

Australian National University

When evaluating the predictive performance of your model:

- Do not estimate it on the same data your trained on.
- If you have selected a subset of features that gave the best performance, you need to measure performance on an independent data set.
- And similarly if you have selected the best hyperparameters based on performance of the data.

- See "An Introduction to Statistical Learning 2ed” for details