Lesson 10

Starting Machine Learning

Understand your data

- See if the numbers are in the expected range
- Check the distribution if it is expected
- Look out for trends and correlations

Common Matplotlib commands (self explanatory)

```
import matplotlib.pyplot as plt

plt.xlabel('some string')
plt.ylabel('some string')
plt.title('some string')
plt.grid(True)
plt.show()
```

The Machine Learning Process (Binary Classification)

Decide your objective

Doctors diagnosed breast cancer tumours (benign or malignant) based on subjective evaluation of images of tumour cells.

What happens if a doctor says a tumour is benign, but it turns out to be malignant?

Scientists asked:
can we extract measurements of those cells using image processing and use machine learning to predict the cancer?

Collect the necessary data, process and clean it

Your data may need to be

- Collected from scratch
- Assembled from different sources

Your data would have to be examined for

- Outliers (are these values wrong? Or just too large?)
- Missing entries (why are they missing?)
- Invalid entries

By using an existing dataset, we skip this step.

Understand the data

- Plotting suitable graphs
- Five-number summary

Choose a model to build
In this problem set, we use the k-Nearest Neighbours algorithm. There are many types of model that we can use. The choice of model depends on the objective.

Build a model using the existing data
You would need to determine the performance of the machine learning model on existing data first. This helps you and users to gain confidence that it can be deployed in real situations.

Clicker Question 1. Think of the breast-cancer scenario and pretend that you are the scientist. What would you want your breast cancer model to be able to **do well**?

A. Predict most malignant cancers correctly
B. Predict most benign cancers correctly
C. Both A & B

You need to divide your data into two sets:
Training set - this data is used to build the model
Test set - this data is used to evaluate the predictions of the model

The usual split is 60:40. Records are **randomly sampled** from the dataset into each set. This is achieved by the `train_test_split()` function.

A random sampling process will separate the records into the two sets. The usual split is 60:40.

Clicker Question 2. A machine learning model is supposed to make predictions on a particular column of data in the dataset. What is that column called?

A. Target
B. Feature
C. Record
D. Observation

Build your model using the training set

Use your model to predict the targets in the (?)

A. training set
B. test set

Build your confusion matrix

The results of the model are summarized in the confusion matrix. From the confusion matrix, several metrics can be calculated. Two important ones are

Accuracy – total correct predictions as a percentage of the total number of records

Sensitivity – total correct predictions on the positive case as a percentage of total number of positive cases in the records

Iterating your model

How do you know the k chosen in Question 4 is the best one? We have to iterate through the values of k to find the best one.

This means we actually need to partition our dataset into Training set, Validation set, and Test set.

Clicker Question. Which set helps us to decide the best value of k?

A. Test Set
B. Training Set
C. Validation Set

Clicker Question 3. The test set has 100 records, and 20 records are of target category “malignant” and 80 records are of target category “benign”. You build a model that ends up classifying all records as “benign”. What will the confusion matrix look like?

A	Predicted malignant	Predicted benign
Actual malignant	0	80
Actual benign	0	20

B	Predicted malignant	Predicted benign
Actual malignant	0	20
Actual benign	0	80

C	Predicted malignant	Predicted benign
Actual malignant	0	0
Actual benign	20	80

D	Predicted malignant	Predicted benign
Actual malignant	0	0
Actual benign	80	20

The Machine Learning Process (Linear Regression)

Decide your objective

You notice that features 0 and 3 of the breast cancer dataset seem to have a relationship. If it is true, you could construct an equation where values in column 3 can be calculated from column 0. You would like to investigate the extent to which this is possible.

Extract the data

Understand the data

- Plotting suitable graphs
- Five-number summary

Decide what kind of equation you need and decide your independent and dependent variables.

Let’s start with a linear equation.

You need to prove that your equation can predict existing data well.

You need to divide your data into two sets:

Training set - this data is used to build the linear equation
Test set - this data is used to check the predictions of the equation

The usual split is 60:40. Records are **randomly sampled** from the dataset into each set. This is achieved by the `train_test_split()` function.

Build your linear model using the training set

Clicker Question. Use your model to predict the (1) **in the** (2)

A. (1) independent variable (2) training set
B. (1) dependent variable (2) training set
C. (1) independent variable (2) test set
D. (1) dependent variable (2) test set

Calculate metrics to determine the performance

- Plot a graph
- Mean-squared error
- R2-score

Iterating the model

Is a linear model sufficient to explain the relationship between the two features ? If not, we may attempt polynomial regression.