“Squashing Bugs Not Snakes”: EMS506U Crib Sheet

By Muhie Al Haimus, Silvia Valls Santafe & Dr Rehan Shah

Introduction

The purpose of this crib sheet is to provide a quick overview of the core concepts needed to understand Data Science and Machine Learning, and a brief reference for libraries used within the EMS506U course including: Matplotlib, Tensorflow, and Pandas.

Matplotlib

Thinking back to EMS412U and EMS418U, you would have used the *matplotlib* library to plot simple line charts. The great thing about *matplotlib* is that it is extremely versatile, especially within the realm of data science, as it also contains loads more chart types such as heatmaps and boxplots for data visualisation, which can be particularly handy when you want to inspect the performance of machine learning models.

Pandas

Imagine a fully Python syntactic-based version of excel, well you don’t have as Pandas is just that! Unlike excel that uses both a graphical user interface and a basic programming language to do calculations on large datasets, Pandas has its own set of basic functions, methods and classes to handle all of the hard work, for example it contains functions to calculate the mean, standard deviation (SD) and variance from a dataset straight out of the box! Just specify which data you want to calculate these statistics for and run the code!

TensorFlow

TensorFlow is an open-source tool in Python used to build and run machine learning models with ease. It streamlines the process of creating deep learning models and convolutional neural networks.

Example TensorFlow Codes

```
import tensorflow as tf

mnist = tf.keras.datasets.mnist
(x_train, y_train), (x_test, y_test) =
mnist.load_data()
x_train, x_test = x_train / 255.0,
x_test / 255.0

model = tf.keras.models.Sequential([
tf.keras.layers.Flatten(input_shape =
(28, 28)),
tf.keras.layers.Dense(128,
activation='relu'),
tf.keras.layers.Dropout(0.2),
tf.keras.layers.Dense(10) ])
```

Machine Learning

Machine Learning (ML) is about building mathematical models that learn patterns from data and try to find a function that maps inputs X to outputs y:

$$f_\theta(X) \approx y$$

Datasets in ML

In ML, data is split into separate subsets.

- **Training set**: used to fit the model.
- **Validation set**: used for model selection.
- **Test set**: used for final evaluation.

This separation helps assess how well the model generalises to unseen data.

Evaluation Metrics

Evaluation metrics are used to quantify how well a model performs on unseen data.

- **Accuracy**: proportion of correct predictions.
- **Precision and Recall**: performance on positive class.
- **Mean Squared Error (MSE)**: error magnitude in regression.

The choice of metric depends on the type of problem and the application.

Types of learning

ML problems are classified into different learning types.

- **Supervised learning**: data with labels.
- **Unsupervised learning**: data without labels.
- **Reinforcement learning**: learning from rewards.

Models

A model is a mathematical structure used to represent patterns in data.

- **Linear regression**: simple linear relationships
- **Decision trees**: rule-based decision models
- **Fully connected neural networks**: non-linear function approximation
- **Convolutional neural networks (CNNs)**: data with spatial structure

Useful Links

Pandas Documentation
Pandas Tutorial Videos
Matplotlib Documentation
Matplotlib + Pandas Video
Tensorflow Documentation
Tensorflow Videos