

## COMP 333 — Week 10 Welcome

# Machine Learning

In Week 10 (this week) and Week 11 we cover Machine Learning.

EDA builds *models* to capture the insight  
to predict outcomes in new situations  
as aids to decision-making.

Machine Learning is one way to build models  
— in a data-driven way —  
the algorithms learn from the data.

Feature engineering is a key contributor  
to the success of machine learning.

The simple methods to build models  
such as curve-fitting  
can be as informative as complex ML algorithms.  
So start with simple approaches to model building  
before moving on to ML.

This week (week 10) we cover

- ▶ ML Example 1

Builds a Decision Tree

for the iris dataset

and explains the working of the decision tree algorithm

- ▶ ML Example 2

Shows the ML Process

builds and compares six models

for the iris dataset

using the Python `scikit-learn` library.

Example 2 is an example of the full EDA process.

- ▶ Machine Learning (Introduction)

Types of ML: Unsupervised, Supervised, Semi-Supervised, Reinforcement.

Regression, Classification, Prediction.

Next week (Week 11) we cover

► Machine Learning Process

Splitting the dataset into training set + test set.

Setting evaluation metrics.

Training the model using the training set.

Cross-validation to evaluate different models,  
and tune (hyper)parameters of ML algorithms.

Evaluation of final model using (independent) test set.

► Machine Learning Algorithms

k-means clustering, hierarchical clustering.

linear regression

logistic regression (for classification)

k-Nearest Neighbour classification

decision trees, random forest

support vector machine

artificial neural network

► Guidelines to ML

Machine Learning is a major discipline in its own right.

**You will not become an expert on ML in this course.**

You should know the following:

- ▶ What is machine learning  
ML terminology
- ▶ Where does ML fit in Exploratory Data Analysis
- ▶ What is a model
- ▶ What kind of models does ML build
- ▶ Where does feature engineering fit in ML
- ▶ What is unsupervised machine learning
- ▶ What is supervised machine learning  
label, class  
binary classifier  
multi-class classifier  
multi-label classifier
- ▶ What is regression, classification, prediction
- ▶ The process of building and evaluating a ML model  
dataset, training set, test set, cross-validation,  
k-fold cross validation, leave-one-out cross-validation (LOOCV)
- ▶ Evaluation metrics in ML  
true and false positive and negative, TP, TN, FP, FN  
confusion matrix  
precision, recall  
accuracy  
specificity, sensitivity  
F-measure  
Matthews Correlation Coefficient (MCC)

You do **not** need to know:

- ▶ how the ML algorithms work
- ▶ how to handle imbalanced data
- ▶ how to generate data
- ▶ the theory or statistics behind the ML algorithms
- ▶ technical details of the issues, such as
  - over-fitting,
  - independence (additivity, homoscedasticity),
  - regularization
- ▶ semi-supervised learning
- ▶ reinforcement learning
- ▶ deep learning

**Most importantly, you must know how to use**

the Python `scikit-learn` library

to build and evaluate models,

as shown in Example 2.

READ the files marked READ.

Do Labs 10 and 11 with scikit-learn.

Work through Example 1 and Example 2

For these topics, it is worthwhile to also read/watch the supplementary material.

All the best, Greg.