

PRACTICAL APPLICATIONS OF MACHINE LEARNING USING PYTHON

Bespoke course for SEPNet – 21st and 22nd Mar 2023

1	BASICS OF PYTHON RECAP	11
1.1	Downloading Python and Setting Up Your IDE.....	12
1.1.1	Installing Python.....	13
1.1.2	Choosing and Installing An integrated development environment (IDE)	14
1.1.3	How to install packages	20
1.1.4	Configuring Your IDE for Python Development	22
1.1.5	Quiz	25
1.2	Python Fundamentals – Data Types	27
1.2.1	Numbers.....	27
1.2.2	Assignments, Strings and Types.....	28
1.2.3	Lists, Tuples and Dictionaries.....	29
1.2.4	Classes.....	30
1.2.5	Quiz – Python Fundamentals – Data Types	31
1.3	Python Fundamentals – Logic	33
1.3.1	If/Else	33
1.3.2	Loops.....	34
1.3.3	Functions.....	35
1.3.4	Quiz	36
1.4	Essential Python Libraries.....	37
1.4.1	NumPy.....	37
1.4.2	Matplotlib.....	38
1.4.3	Pandas.....	42
1.4.4	Scikit-learn.....	45
1.4.5	Pickle	46
1.4.6	Quiz: Essential Python Libraries.....	48
1.5	Using the Debugger and Getting Help Online	49
1.5.1	Types of Errors	49

1.5.2	Try/Catch Runtime Exceptions.....	50
1.5.3	Debugging using your IDE	51
1.5.4	Using help() and dir() Functions.....	52
1.5.5	Getting Help Online Using StackOverflow, YouTube... ..	53
1.5.6	Quiz: Using the Debugger and Getting Help Online	53
1.6	Good Coding Practices	54
1.6.1	PEP 8 – Style Guide For Python.....	56
1.6.2	Checking the Minimum Required Python Version.....	56
1.6.3	Project Documentation.....	56
1.6.4	Using Logging in Python.....	56
1.6.5	Quiz: Good Coding Practices	56
1.7	Collaborative Coding Using GIT	57
1.7.1	What is Version Control?	57
1.7.2	How to Install and Setup GIT on your PC/IDE	58
1.7.3	How to Set Up an Online GIT Repository: GitHub.....	59
1.7.4	Common GIT	60
1.7.5	Quiz: Collaborative Coding Using GIT	62
2	WHAT IS MACHINE LEARNING?	63
2.1	Overlaps with Maths and Statistics	64
2.1.1	Linear Algebra	65
2.1.2	Calculus	65
2.1.3	Probability and Statistics.....	65
2.1.4	Information Theory.....	65
2.1.5	Graph Theory:	65
2.1.6	Numerical Methods	65
2.1.7	Probability and Probability Distributions	65
2.1.8	Statistical Inference.....	65
2.1.9	Maximum Likelihood Estimation	65
2.1.10	Bayesian Inference.....	65
2.1.11	Regression Analysis.....	66
2.1.12	Time Series Analysis.....	66
2.1.13	Hypothesis Testing.....	66
2.2	Structure of Datasets – Features and Labels	66

2.2.1	What Are Features?	66
2.2.2	What Are Labels?	67
2.2.3	Understanding the Distribution of Data	67
2.2.4	Understanding the Correlation Between Features and Labels.....	67
2.2.5	Practical Example: Revisiting the Iris Dataset	67
2.3	The Perceptron Algorithm	68
2.3.1	The History	68
2.3.2	The Perceptron Algorithm	68
2.3.3	The Limitations.....	68
2.3.4	Extensions and Variants of the Perceptron Algorithm	68
2.3.5	Comparison With Other Algorithms	68
2.3.6	Practical Example: Applying the Perceptron to the Iris Dataset.....	69
2.4	Optimisation Through Iteration and Evaluation – Gradient Decent	69
2.4.1	Optimization Techniques	69
2.4.2	Understanding Gradient Descent	69
2.4.3	Optimization Evaluation.....	69
2.4.4	Practical Example: Charting Navigation of Perceptron Weights.....	69
2.5	Bias-Variance Trade-Off the Curse of Overfitting	70
2.5.1	Understanding Bias and Variance	70
2.5.2	Understanding Model Complexity	70
2.5.3	Handling Overfitting:.....	70
2.6	Main Types of Machine Learning Techniques	71
2.6.1	Supervised Learning	71
2.6.2	Unsupervised Learning.....	71
2.6.3	Semi-supervised Learning	71
2.6.4	Reinforcement Learning.....	71
3	SUPERVISED LEARNING: CLASSIFICATION	72
3.1	Logistic Regression	73
3.1.1	Introduction to Logistic Regression	73
3.1.2	The Logistic Function	73
3.1.3	Linear Regression vs Logistic Regression	73
3.1.4	Maximum Likelihood Estimation	73
3.1.5	Regularization	73

3.1.6	Practical Example: Applying Logistic Regression to the Iris Dataset.....	73
3.2	K-Nearest Neighbours (K-NN)	74
3.2.1	History of the Main Idea	74
3.2.2	Distance Metrics	74
3.2.3	The Curse Of Dimensionality.....	74
3.2.4	Selecting The Value of K.....	75
3.2.5	Practical Example: K-NN on the Iris Dataset	75
3.3	Naïve Bayes	75
3.3.1	Bayes Theorem.....	75
3.3.2	The Naïve Assumption	75
3.3.3	Gaussian/Multinomial/Bernoulli Naïve Bayes	75
3.3.4	Practical Example: K-NN on Spam Filtering	75
3.4	Decision Trees	76
3.4.1	The History and Main Idea	76
3.4.2	The Tree Structure	76
3.4.3	The Splitting Criteria	76
3.4.4	Building a Decision Tree	76
3.4.5	Pruning a Decision Tree	76
3.4.6	Limitations and Alternatives	76
3.4.7	Practical Example: Decision Tree on the Iris Dataset.....	77
3.5	Support Vector Machines (SVM's).....	77
3.5.1	Introduction to Support Vector Machines	77
3.5.2	Linear SVM	77
3.5.3	Non-Linear SVM	77
3.5.4	Kernel Trick	77
3.5.5	Regularization	77
3.5.6	Practical Example: SVM on the Iris Dataset	77
3.6	Neural Networks	77
3.6.1	Introduction to Neural Networks.....	78
3.6.2	Artificial Neurons – Our Good Friend the Perceptron	78
3.6.3	Feedforward Neural Networks	78
3.6.4	Activation Functions.....	78
3.6.5	Convolutional Neural Networks (CNN)	78

3.6.6	Recurrent Neural Networks (RNN).....	78
3.6.7	Practical Example: Neural Networks on the Fashion MNIST Dataset.....	78
4	Supervised Machine Learning: Regression.....	79
4.1	Linear Regression.....	79
4.1.1	Introduction To Linear Regression.....	79
4.1.2	Extending From One To Multiple Independent Variables.....	79
4.1.3	Assumptions of Linear Regression.....	79
4.1.4	Model Evaluation.....	79
4.1.5	Solving via Matrices and Gradient Descent.....	79
4.1.6	Extensions of Linear Regression.....	79
4.1.7	Practical Example: Linear Regression on Boston Housing Dataset.....	80
4.2	Non-Linearity Using SVM Regression.....	80
4.2.1	Introduction to SVM Regression.....	80
4.2.2	Maximum Margin and Soft Margin in SVM Regression.....	80
4.2.3	Using Kernels for Non-Linearity.....	80
4.2.4	Practical Example: Using SVM Regression on Boston Housing Dataset.....	80
4.3	Regression Trees.....	80
4.3.1	Introduction To Regression Trees.....	80
4.3.2	How Regression Trees Handle Non-Linearity.....	80
4.3.3	Practical Example Using Regression Trees on Boston Housing Dataset.....	80
5	Unsupervised/REINFORCEMENT learning.....	81
5.1	Clustering.....	81
5.1.1	What is Clustering?.....	81
5.1.2	How to Evaluate Performance?.....	82
5.1.3	Techniques for Effective Clustering.....	82
5.1.4	Practical Example: K-Means Clustering Cancer Genomics Dataset.....	82
5.2	Genetic Algorithms (GA).....	82
5.2.1	What Are Genetic Algorithms?.....	82
5.2.2	How To Use GA's.....	82
5.2.3	Practical Example: Solving Symbolic Regression Using GA's.....	82
5.3	Q-Learning.....	83
5.3.1	What is Reinforcement Learning?.....	83
5.3.2	What is Q-Learning?.....	83

5.3.3	Practical Example: Cart-Pole Balancing.....	83
6	Effective ML Training Techniques.....	84
6.1	Train/Test Splitting.....	85
6.1.1	Importance of Train/Test Splitting.....	85
6.1.2	Determining the Split Ratio.....	85
6.1.3	Handling Imbalanced Classes.....	85
6.1.4	Keeping a Validation Set	86
6.1.5	Data Leakage.....	86
6.1.6	Random Seed	86
6.1.7	Practical Example: How to Perform Train/Test Splitting	86
6.2	Cross Validation	86
6.2.1	What is Cross Validation?	86
6.2.2	Choosing The Number of Folds.....	86
6.2.3	Special Uses of Cross Validation	87
6.2.4	Practical Example: How to Perform Cross Validation	87
6.3	Normalisation & Standardisation.....	87
6.3.1	What is Normalisation?.....	87
6.3.2	What is Standardisation?	87
6.3.3	Practical Example: Before and After Normalisation/Standardisation	88
6.4	Feature Selection	88
6.4.1	What is Feature Selection?	88
6.4.2	Main Types of Feature Selection.....	88
6.4.3	Common Feature Selection Techniques	88
6.4.4	Practical Example: Feature Selection on Wikipedia Dump.....	89
6.5	Feature Extraction.....	89
6.5.1	What is Feature Extraction?.....	89
6.5.2	Linear Algebra Based Techniques	89
6.5.3	Feature Extraction Techniques for Text and Image Datasets.....	89
6.5.4	Practical Example: PCA on Artificial Data Make Blobs.....	90
6.6	Confusion Matrix.....	90
6.6.1	What is a Confusion Matrix?.....	91
6.6.2	ROC (Receiver Operator Characteristic) Curves.....	91
6.6.3	Practical Example: Compare ROC Curves of Learners.....	91

6.7	Bootstrapping – Bagging and Boosting	91
6.7.1	What is Bootstrapping?.....	91
6.7.2	Bagging Methods	92
6.7.3	Boosting Methods.....	92
6.7.4	Practical Example: Random Forest Decision Trees	92
7	Common Challenges IN ML	93
7.1	Imbalanced Datasets	94
7.1.1	The Problem of Imbalanced Data	94
7.1.2	Techniques for Dealing with Imbalanced Data	94
7.1.3	Using Different Metrics To Evaluate Classifier Performance	94
7.1.4	Using Ensemble Methods	94
7.1.5	Practical Example: Fraud Data Set Prediction.....	94
7.2	Outliers.....	94
7.2.1	What is an Outlier?	94
7.2.2	How to Detect Outliers	94
7.2.3	How To Handle Outliers.....	95
7.2.4	Impact of ML Models	95
7.2.5	Using Domain Knowledge	95
7.2.6	Practical Example: Finding Outliers in the Tips Dataset.....	95
7.3	Missing Data	95
7.3.1	The Problem of Missing Data	95
7.3.2	How to Deal with Missing Data.....	95
7.3.3	Model Performance Metrics for Missing Data.....	95
7.3.4	Advanced Techniques for Missing Data.....	95
7.3.5	Practical Example: Missing Data with the Titanic Dataset.....	96
7.4	Small Datasets	96
7.4.1	The Problem of Small Datasets	96
7.4.2	How to Deal with Small Datasets	96
7.4.3	ML Techniques for Small Datasets.....	96
7.4.4	Practical Example: Weather Dataset?????	96
7.5	Collinearity of Features	96
7.5.1	Why is Collinearity a Problem?	96
7.5.2	How to Detect Collinearity.....	96

7.5.3	How to Deal with Collinearity	96
7.5.4	Practical Example: Body Mass Index Dataset	96
7.6	Breaking IID	97
7.6.1	What is the IID Assumption?.....	97
7.6.2	How To Detect Non-IID Data?.....	97
7.6.3	How to Deal with Non-IID Data?.....	97
8	Pipeline for Machine Learning PROJECTS.....	98
8.1	Collect the Data.....	99
8.1.1	Methods for Acquiring Data.....	99
8.1.2	Practical Example: Web Scraping Trading View.....	99
8.2	Prepare the Data	99
8.2.1	The Importance of Data Preparation	99
8.2.2	Checking Your Data with Visualisation.....	99
8.2.3	Techniques for Cleaning Data	99
8.2.4	Techniques for Encoding Data	100
8.2.5	Techniques for Continuous Features	100
8.2.6	Dealing with Imbalanced Data	100
8.2.7	Dealing with High Dimensional Data	100
8.2.8	Splitting Data For Training/Testing	100
8.2.9	Documenting Your Data.....	100
8.2.10	Practical Example: Preparing Our Web Scraped Dataset.....	100
8.3	Choose A Model	100
8.3.1	What Type of Machine Learning Model?.....	100
8.3.2	Practical Example: Regression of Time Series Scraped.....	100
8.4	Train The Model.....	100
8.4.1	Train/Test Splitting and Cross Validation.....	100
8.4.2	Optimising Training Times	100
8.4.3	Optimising Testing Times	101
8.4.4	Practical Example: Training Neural Network Model Price Regression	101
8.5	Evaluate The Model.....	101
8.5.1	Evaluation Metrics	101
8.5.2	Diving Deeper into Model Performance	101
8.5.3	Practical Example: Feature Contribution from Neural Network Weights	101

8.6	Parameter Tuning	101
8.6.1	The Importance of Parameters.....	101
8.6.2	How To Tune Parameters.....	101
8.6.3	Practical Example: Using GridSearchCV Tune Random Forest	102
8.7	Make Predictions	102
8.7.1	How To Make Predictions on Test Data.....	102
8.7.2	Practical Issues With Making Predictions	102
8.7.3	Deploying A Model In Production.....	102
8.7.4	Practical Example: Visualising how Decision Tree Makes a Prediction	102
9	Advanced ML Topics	103
9.1	Recurrent Neural Networks (RNNs) and LSTM's	104
9.1.1	The Basics of RNN	104
9.1.2	The LSTM Architecture.....	104
9.1.3	Training LSTM's	104
9.1.4	Common Issues	104
9.1.5	Practical Applications.....	104
9.2	Deep Learning	105
9.2.1	What is Deep Learning?	105
9.2.2	Implementing Deep Learning.....	105
9.3	Deep Q-Learning	105
9.3.1	What is Deep Q-Learning?	105
9.3.2	Applying Deep Q-Learning	105
9.3.3	Challenges of Deep Q-Learning.....	105
9.3.4	Improvements to Deep Q-Learning	106
9.4	Generative Adversarial Neural Networks (GANN)	106
9.4.1	Typical GANN Architecture	106
9.4.2	Applications of GANN's.....	106
9.5	Conformal and Probabilistic Prediction	106
9.5.1	What is Conformal Prediction?.....	106
9.5.2	What are Venn Estimators?	106
9.5.3	Applications Of Conformal Prediction	106
10	PRACTICAL Assignments	107
10.1	Practical – Data Wrangling with Pandas: Titanic Dataset	108

10.2	Practical – Classification Case Study: Credit Card Fraud	108
10.3	Practical – Clustering Case Study: Iris Dataset.....	108
10.4	Practical – Regression Case Study: Housing Prices	108
10.5	Practical – Image Classification Case Study: MNIST Dataset	108
11	Glossary	109
12	REFERENCES.....	116

DRAFT