DATA
Fellowship

Data Science in a Day Cheat Sheet

Below is an overview of links, libraries and instructions for today's session

Python Keywords:

- def
- if, else
- return

Python Functions:

- print()
- min()
- type()

Python Data Types:

- Integer
- float
- string

Python Packages:

- Pandas
- Seaborn
- Sklearn
- Graphviz

Pandas:

1. Purpose: Data reading, writing, analysis and manipulation
2. Some functions we'll be using:
    a. read_csv() -- to read the .csv file
    b. head() -- outputs the first 5 rows of our data frame
    c. info() -- outputs high-level information about our features
    d. describe() -- outputs basic descriptive statistics for all numeric features
    e. value_counts() -- gets number of unique values a column can have
    f. drop() -- removes columns or rows from the dataframe
    g. dropna() -- drops NAs in either the specified rows or columns
    h. fillna() -- replaces NAs in either the specified rows or columns
    i. get_dummies() -- used to one-hot encode or binarize our features

DATA
Fellowship

Seaborn:

1. Purpose: Data visualisation

# Variables	Data Type(s)	Graph name	Seaborn function	Parameters
One	Categorical	Count plot	sns.countplot()	a. X b. data
One	Numerical	Distribution plot	sns.distplot()	a. X
Two	Numerical, Numerical	Scatter plot	sns.scatterplot()	a. X b. Y c. data
Two	Categorical, Numerical	Bar chart	sns.barplot()	a. X b. Y c. data
Two	Categorical, Numerical	Box plot	sns.boxplot()	a. X b. Y c. data
Two	Categorical, Numerical	Swarm Plot	sns.swarmplot()	a. X b. Y c. data
Two	Categorical, Numerical	Violin Plot	sns.violinplot()	a. X b. Y c. data
Three	Categorical, Categorical, Numerical	N/A	sns.catplot() *OR* Add 'hue' parameter to any of the above	a. X b. Y c. data d. hue

GraphViz:

1. Purpose: Visualising decision trees

Sklearn:

1. Purpose: Machine learning
2. Some functions we'll be using:
    a. DecisionTreeClassifier() -- calls the decision tree model from Sklearn
    b. fit() -- Trains/fits a decision tree model to our data
    c. predict() -- predicts new output data based on new input data
    d. train_test_split() -- splits the data into training and testing sets
    e. accuracy_score() -- outputs the accuracy score of our model
    f. classification_report() -- outputs other metrics for model evaluation

DATA
Fellowship

General Programming Glossary:

- A variable = a placeholder or 'container' for a datum (this datum is also called the variable's *value*).
- In programming, we can only make a variable if we give it a name.
- A variable name = the name we give to a variable when we create it (e.g: the instruction 'tax = 0.25 makes a variable called 'tax', and gives it the number 0.25 as its value).
- A datum (pl. data) = a piece of information (e.g: a number, some text, the truth-values 'True' and 'False', or a sequence of such things).
- A numerical variable = a variable whose value is a number.
- An array = a sequence of data (e.g: the array [7, 4, 1] contains the data 7, 4, and 1, in that order).
- A function = a sequence of instructions, performable by a machine, that takes some data as input, manipulates it, and produces some output (e.g: a recipe is a kind of function. It takes the raw ingredients as input, manipulates them, and produces the output of the meal).
- An algorithm = a set of instructions; a recipe.
- A program = a sequence of instructions, executable by a computer, that creates or uses variables, functions and logic.
- An application = a software program that runs on a computer.
- API = Application Programming Interface, i.e, a set of (often freely available) functions which, when embedded into a program, allows that program to access the data within another application or database (e.g: the Google Maps API is a set of functions which, when embedded into the program of (for example) a small café owner's website, allows that program to access the wealth of cartographical data within Google Maps; thereby enabling a widget displaying the location of the café on its website).
- A widget = a small program, often embedded within the program of a website, enabling its user to access some service or perform some function.
- A library = a collection of code that serves as the basis for a programming language or computer program (e.g: libraries can include helpful documentation, data types, and pre-written functions).
- A process = a task performable by a computer (e.g: adding two numbers together).
- CPU = the Central Processing Unit of a computer; a chip on which the most fundamental processes - such as logic and maths - take place.
- GPU = Graphics Processing Unit; a chip that performs processes in parallel; typically used for graphics processing, but sometimes used for non-graphical processes.
- Structured data = data that are organised by some data model.

DATA

Fellowship

- Unstructured data = data that are not structured.
- A data model = a system that makes data more understandable and analysable (e.g: imagine we have a huge amount of varied information on the undergraduate students within some university. One data model might organise this data in a table, with the columns *age*, *subject studied*, *gender*, etc).
- Artificial Intelligence = the subject concerned with capturing the principles of human cognition, and aiming to mimic the phenomena of human intelligence (AI involves mainly the development of logic-based, probabilistic and statistical theories). With AI-based systems it is not always understood how a computer comes to behave as it does: though inference learns to do so.
- Machine Learning = the use of statistical techniques to give computer systems the ability to 'learn' from data, such that it's not always well understood how the computer comes to behave as it actually does.
- Supervised Learning occurs when a machine, after being presented with labeled training data comprising input/output pairs, learns a general rule that maps inputs to outputs.
- Unsupervised Learning occurs when a machine, after being presented with unlabeled training data, learns a generality.
- Reinforcement Learning occurs when a machine, by repetitively trying to complete a task and being rewarded only when it succeeds, infers a general method by which to complete that task.
- Data lake = a repository of data in its natural format.
- Data warehouse = an organised repository of data.
- Linear Regression = a linear method by which to model the relationship between two or more variables.
- A Decision Tree = a method by which to display, in a tree-like way, an algorithm involving only 'if-then' statements as instructions.
- Neural Networks = a computer system loosely modelled on the human brain and nervous system.