https://christgithub.github.io/eportfolio/

Reflexion on exploring Machine Learning

Discussing the fourth industrial revolution and the use of large language models in various industries and its impact on our life was extremely interesting. Those discussions broaden my views and thinking about the impact of smart AI powered technologies. On one hand, it highlighted and confirmed real concerns regarding the unemployment rate linked to the use of AI assistant tools. Digital forgery is also a major issue which will progress exponentially, raising legal battles with authors and artists fighting for their property copyrights. On the other hand, the progress in the medical field is definitely counterbalancing the argument, offering invaluable help to practitioners and hope to cure illnesses still incurable to this day. I am a huge believer in AI powered robotics and it is also a field I really look forward to exploring. As we progressed in the course, I discovered many passionate researchers and writers in the AI community. Both technical and ethical researchers in this field really opened my eyes about how we could greatly benefit from a super artificial intelligence but also what is at stake in pursuing the development of an AGI. To name a few, Yann Lecun, Demmis Hasabis, Cathy O'Neil, Joy Buolamwini, Geoffrey Hinton, Emely Bender, Meredith Broussard… Will shape the beginning of my journey exploring this fascinating field.

Before joining the course, I had a somewhat vague idea of machine learning. Indeed, working as a software engineer allows me to be exposed to numerous technology stacks. I have participated in projects with our data science team to integrate classifier models dedicated to understanding customers behaviour and purchasing patterns. Although, during discussions, I was always puzzled by the terminology surrounding the conversations.
Going through the details of machine learning and studying what is happening behind the scenes and even exploring the underpinning mathematics was an amazing experience.
The course builds up gradually and logically, starting with a strong introduction on exploratory data analysis.

EDA

Setting up a Google Colab notebook by importing the necessary tools for the exploratory data analysis tasks that will be performed later in the notebook. These libraries such as Pandas, Numpy, Matplotlib, Seaborn, Missingno each contain specific functions and functionalities to analyze, manipulate and visualize data. These libraries allow us to generate various types of charts: histograms, pairplots, scatter plots in order to understand the data distribution as well as the relation between the features in the dataset; Univariate and Bivariate analysis

- Peek at the dataset with head() or tail()
- Counting the dataset rows
- The shape of the dataset (rows and columns)
- The data types per columns
- Assessing how much missing data the document counts
- Viewing the mean, median, count, max,
- Visualising the skewness or kurtosis
- Finding out how many unique fields per rows
- Checking whether the type is consistent per column
- Removing columns irrelevant to the analysis
- Reshaping the dataset
- Detecting outliers in the data

Correlation

We dive into understanding the relation between data points. Correlation and regression are statistical techniques to study, quantify and describe the relationship between variables. Correlation quantifies the strength and type of relationship whereas regression expresses the relationship between one dependent variable and one or more independent variables called predictors or target variables

- Quantify and visualise the positive or negative strength of the relationship between variables using Covariance and Pearson Correlation Coefficient.
- Understand linear regression in the context of machine learning by building a predictive model in Python using a Y-intercept form equation.
- Build a polynomial model and plot the predicted outcome

Regression

In statistics, linear regression is a model that estimates the linear relationship between a target scalar response (dependent variable) and one or more independent variables (regressors).
Logistic regression is a supervised machine learning algorithm widely used for binary classification tasks, such as identifying whether an email is spam or not and diagnosing diseases by assessing the presence or absence of specific conditions based on patient test results.

Clustering
Clustering is an iterative method for grouping a collection of objects in such a manner that objects in the same cluster are more similar in some specific predefined characteristics to each other than to those in other clusters.
Leveraging the K-means algorithm, we experimented with clustering by loading the Iris dataset in order to surface different groups of flowers based on characteristics such as sepal length, sepal width, petal length and petal width.
After setting our centroids, we invoke the K-means method 'Fit'.
We use the Elbow method to determine the optimal number of clusters so that we can avoid overfitting.
By employing the sum of square error (SSE) we measure the total squared distance between each data point and the centroid of the cluster it belongs to.

Group project
During the group project, early on, everyone agreed very quickly on what role they were going to take during the whole project. During an initial Zoom call, we discussed what the goals of our project would be and we assigned a task to all.
All team members attended meetings regularly. We organised tasks using an agile approach with the support of a Trello board to allow better visibility and communication for anyone at any time. We set up a Github project repository where we used a feature branch git workflow to commit, review and merge changes to our notebook.
Every step of the project was well organised and handled effectively. As well as being an efficient collaborative process it was very exhilarating too.
Nowadays, many companies are allowing their staff to work remotely. Although I am used to working remotely, I learned that it is always challenging to articulate my thinking as well as I would in person in a group. Above all, I learned to organise myself within a group in a studying setting and follow everyone's pace and expectation which can be different from a professional setting at times when we're all tackling a learning curve at the same time.

As a team member, I help with the tasks I am familiar with such as enabling visibility of concurrent tasks throughout the project duration and setting up Collaborative Agile project approaches. Overall, it was an excellent experience during which I have learned at many levels.

ANNs

An Artificial neural network consists of connected nodes called neurons, which model the neurons in the brain. The connections between neurons serve a function similar to the brain synaptic neural connections adjusting their strength (weight) depending on stimuli. We experimented with the gradient descent function or loss function of an ANN by adjusting the learning rate which is the stride in sort taken during the descent along the function. The cost function is crucial to the learning of a model and offers a metric to measure a model's performance by quantifying the difference between predictions and actual results. The lower the value, the better the model is learning to fit the data in an optimal way. At the lowest value of the cost function, the model has reached convergence.

Applying the calculations of chained partial derivatives to measure and adjust the ANN learning performance was extremely interesting and I finally was able to apply calculus to a concrete use case.

CNNs

Working on classifying images with a convolutional neural network and measuring the performance of the CNN was probably the trickiest part of the module. Preparing the data in advance to optimise its use by the CNN was a step further than more basic EDA. Each step of the dataset optimisation was geared towards specific requirements of the CNN’s image feature detection process and the operations the layers need to perform before passing the data to the classifying layers. Choosing the correct numbers of different types of layers as well as attempting to parameterise (hyperparameters) the model was an iterative process which required logging each parameter (learning rate, epochs, activation function, patience and early stopping) and results of the validating and training scores.

Many new concepts have been learned and many experiments carried out since the group project. There was very much an emphasis on the data analysis related to the nature of the business domain we were attempting to model. A lack of knowledge certainly directed us towards a thorough

exploration of the dataset we were provided with, which as a result we didn't get to properly exploit and model using a more advanced predictive algorithm. These are experiments I look forward to carrying out in the near future. From my short experience working with Machine Learning algorithms, although they are based on extremely well thought out mathematics, as a user and despite all the knowledge required, it very much appears as a trial and error process where experimentation is key.