#### 

#### Machine Learning [(HDip Data Analytics - Sept 2020 cohort)](https://moodle.cct.ie/course/view.php?id=1715)

#### 

Lurdes Teixeira Power sba19131

# **Table of Contents**

[Machine Learning (HDip Data Analytics - Sept 2020 cohort)](#_1dxv19qxybmr) 0

[**Table of Contents**](#_8a4ivxp3sn97) **1**

[**1.0 Question 1**](#_hp2e29qn79pi)

[**2.0 Question 2**](#_41lb6uitug5g)

[**3.0 Question 3**](#_h0t652qhr5zu)

[**Part a**](#_33wvp8fvlokl)

[**Part b**](#_ct6nu6z9eskw)

[**4.0 Question 4**](#_3kfnz6l4sofd)

[**Part a**](#_spjjnz6dogke)

[**Part b**](#_2csubc4067fp)

[**5.0 References**](#_4h7rukj3v25r) **7**

[**6.0 Appendix**](#_9a3oiptq6p2) **7**

[**6.1 Datasets**](#_atpwk51yl1vh) **7**

[**6.2 List of Figures**](#_ajtyvufidvis)

# **Question 1**

There are four major types of machine learning approaches . They are supervised, unsupervised, semi-supervised and Reinforcement Learning.

There are therefore usually a variety of ways to resolve a particular machine learning question.

Supervised learning learns the relationships within the data . Inputs and outputs are fed into an algorithm. When new unseen data is fed into the algorithm , it is able to predict what the output is likely to be. For example, if we have a dataset representing the speed of cars, one may want to find the relationship between the speed of the car and the number of accidents. If we feed inputs of high speeds to the algorithm and a high amount of accidents at those speeds, the algorithm should be able to predict that if it receives high speed data , then we should see an increase in the accidents.The speed versus accident example is a regression problem. A classification example is a forest with a mix of trees and we are trying to classify the type of tree into separate categories like pine or eucalyptus. Supervised learning is suited to both types of problems but it requires a lot of data to be able to work correctly.

In unsupervised learning, only input data is fed into the algorithm. The output is not given. The algorithm is left to find its own patterns. If we take the identifying trees in a forest example, but this time, we do not tell the algorithm that the trees are either pine or eucalyptus . We let the algorithm decide how to divide the values into categories based on their features like the type of leaf.

Semi-supervised learning mixes labelled and unlabelled data. An example is with cancer scans, a small number of scans are labelled and the algorithm applies the labels to other unlabelled scans.

Reinforcement learning works by using a system positive feedback, which turns into a reward. This approach is often used in gaming where scores are given as a reward when playing the game correctly.

A logistic regression , a supervised model would predict churn rates.The split depends on the data available. We need to keep a significant amount for testing, for a small dataset I would split 60/20/20 60 for training, 20 for validation and 20 for testing. More training data builds a better model, while keeping the rest of validation and testing is better for accuracy.

Total : 401 words

# **Question 2**

The poor accuracy could be explained by a couple of reasons.

The main one is overfitting. Overfitting would usually give good training results but poor testing ones. Using k-fold cross-validation would be an option to improve the situation, while another possible solution would be to scrap the original split and change it to a different split.

According to an article by Jason Brownlee, sometimes, overfitting is even possible when the “distributions are imbalanced”, even after cross validation. This usually leads to poor testing accuracy. A solution here would be to try to implement another cross validation method that keeps the “same class distribution in each subset”. This is called stratified k-fold cross-validation. This method is similar to k-fold cross- validation but includes stratification. According to the scikit learn page, this method preserves “ the percentage of samples for each class”.This means that the train-test-split would have good accuracy and we would avoid the need to use random sampling method, because stratified sampling is used.

Random sampling is simple and easy to apply to small populations, but with random sampling , one could easily have a misrepresented data sample. Stratified data sampling , the data is divided into groups called “strata”. The sample is drawn from these groups.

Another possible cause for the inaccurate score , is that the sample size for the training or testing is too small. According to Rosie Cornish, “increasing the sample size also increases the precision”. The classification model could be good, but in order to obtain a good accuracy, one needs a reasonable sample size . Combined with random sampling, this could explain the discrepancy between test and train results.

A small sample size can also lead to biased performance of the algorithm. A possible solution here would be to look at the variance for the train and test sets for the standard deviations and means in the summary statistics.Another possible solution would be to obtain lager representative data, although this is not always possible, and more data can be costly to obtain.

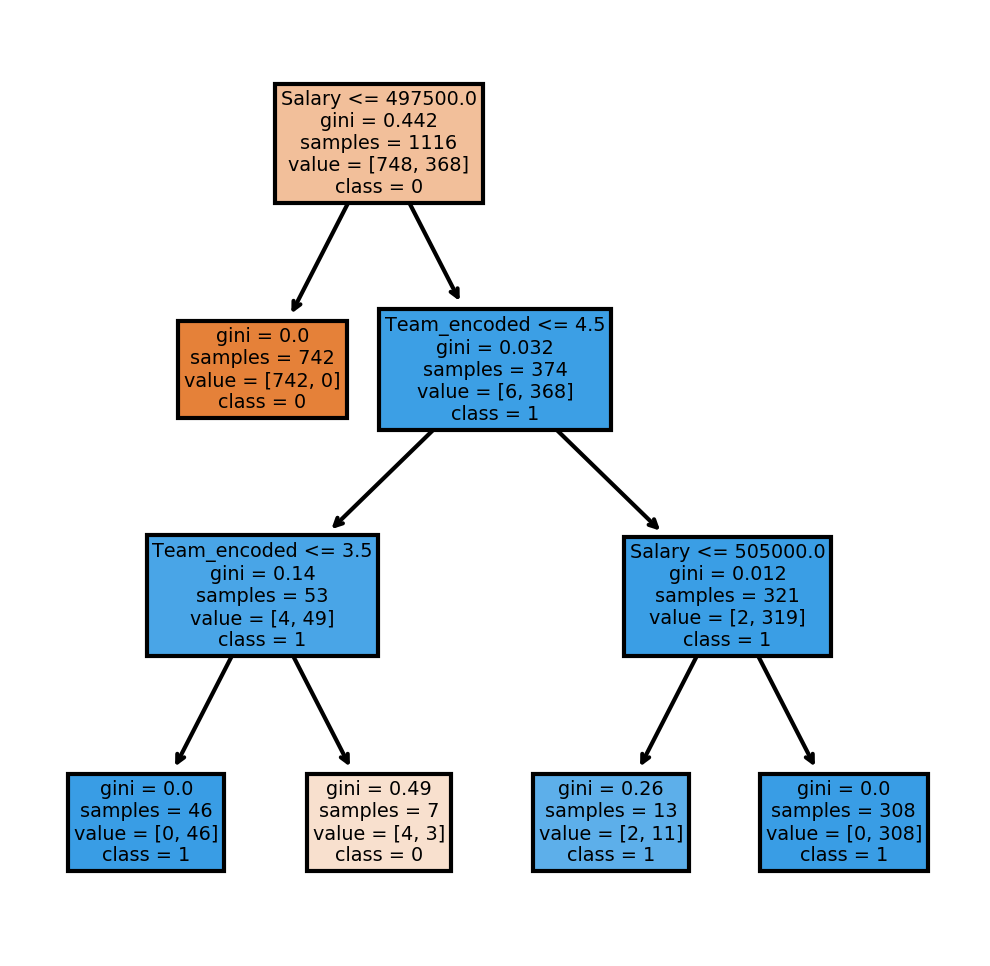
Total : 335 words

# **Question 3**

## **Part a**

Decisions trees are often used to solve classification but can also be used for regression problems. A tree consists of a root node which then splits into subnodes based on the features or attributes of the data. This splitting system comes with a set of issues such as how to split the tree in the first place or how do we stop the splitting process ? Gini Impurity can help with the first of these issues as it calculates the purity of each of the nodes. This is done by calculating the probability of the feature being incorrectly classified. Entropy or Information Gain calculates the degree of impurity and reduces impurity from the root node. Entropy is more complex to calculate because it uses logarithms, while Gini Impurity is faster to calculate.

In my jupyter file example, I have chosen a dataset for baseball players salaries and the model tries to predict who earns 500,000 or more. We first tried with Gini Impurity and got the classification below :



The model works and reaches an accuracy score of 0.996428 ( from a possible 1) which is very good.

Fig 1

## Then, we tried with the entropy to see if that would affect the accuracy score. We got the model in the diagram below :

Fig 2

The accuracy score is very similar if not identical to the score obtained with gini impurity . The score obtained with entropy is 0.9964285 out of a possible 1, which means that the model is very accurate.

## 

## 

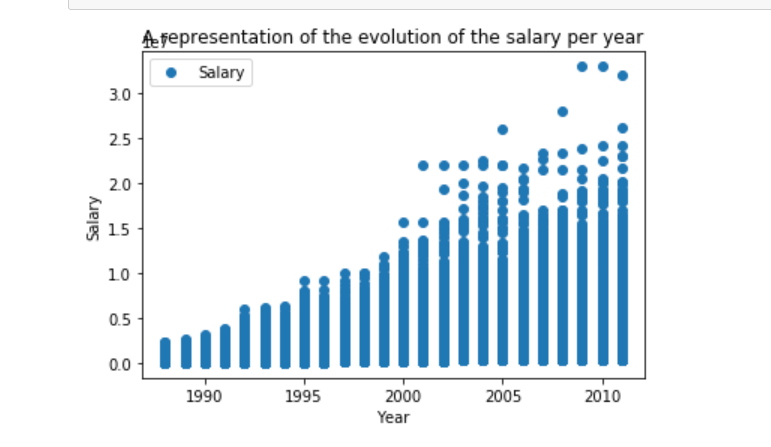
Total : 242 words

## 

## **Part b**

Regression is a different type of supervised learning. It looks at the relationship between dependent and independent variables. A regression outputs numbers and is useful when predicting numbers such as the house market prices or stock market prices , the weather (temperature) or the likelihood of an event.

L1 regularization looks at estimating the median of the data, or of the distribution. Lasso works with this L1 technique. Sometimes known as L1 penalty, because it limits the coefficients.While a L2 penalty looks does not shrink all coefficients the same way. L2 regularization looks at estimating the mean of the data. Ridge uses this technique. Both regularizations are used to avoid overfitting the model during the training process.

In the Jupyter file, I have attempted to predict the salaries of the baseball players using Linear Regression as, this type of regression is suitable for this kind of task.

I first plotted the data and the distribution

Fig 3



Fig 4

When the model was created I wanted to see the predicted values , so I decided to plot the actual and predicted values in a histogram . The graph clearly shows an imbalance between actual and predicted values .

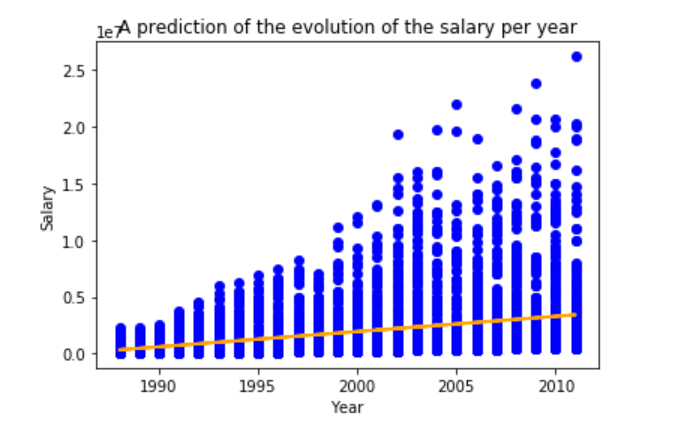
At the end, I wanted to show the final results I had obtained , so I decided to plot the prediction obtained in another graph . This graph confirms that the model is not predicting accurately. As the orange line should run along the values represented in blue .

Fig 5

A logistic regression is used to solve classification problems and relies on categorical data. I applied it to the baseball dataset with very good results. No I am no longer predicting a continuous output but I am predicting based on categories such as team or position.

The accuracy score is 0.99285 out of a possible 1 and a 0.99 in the classification report out of a possible 1 also, so the model seems to be working.

Total 315 words

# 

# 

# 

# **Question 4**

## **Part a**

Gaming in general is an area of growth for reinforcement learning, and an automated chess game is now operated by bots, who learn from the moves you make and from past games. Other areas have successfully implemented Reinforcement learning such as robotics for industrial automation. An article from Robotics Online Marketing Team from 2018 explains how a robot is fed images of parts and can then recognize the real parts for example.

In active reinforcement learning, the algorithm needs to decide on its own actions and sequences of what it needs to do; where as in passive reinforcement learning, the algorithm is given a sequence of actions to evaluate. Each of them have their own techniques . Passive learning uses “Direct Utility Estimation”, “Adaptive Dynamic Programming (ADP)” or “Temporal Difference Learning (TD)”, while active learning uses “ADP with exploration function” and “Q-Learning”. Direct Utility Estimation, a sequence of trials is recorded, and an “estimation is based on sample values”.ADP tries to learn the model from the environment by evaluating the rewards or all other expected rewards of each stage, adjusting in each stage. TD is able to learn by adjusting to the next stage only.The difference between ADP and ADP with exploration function is that it adjusts weights by explores areas that haven’t been searched and lowers weights to already explored areas. Q-learning is a type of TD that is simpler to compute but slower than ADP. It learns Q-value functions.

Total : 242 words

# **Part b**

Natural Language Processing (NLP) is part of Artificial Intelligence, its aim is for machines to be able to understand and interpret the human language. A powerful method of text manipulation , NLP now has an important role in human interactions, from being able to identify and auto translate text and correct grammar to being at the source of Android app design. Humans now interact far more with machines that they even realise : calls in a call center are now often handed by chatbots while Siri or Alexa provides us with the day’s weather.

Social media has now become one of the fields where NLP is key. A key portal for businesses, Facebook is now more than a platform for friends to catch up. Businesses can analyse a variety of things thanks to NLP, from sentimental analyses on the product reviews , which can help the business know how a product fares after sell or before launch; demographics and trends can be analysed within groups, providing data for research. For business NLP in social media is a tool that can be key to an advertising campaign or analysis of marketing emails or surveys, providing the business with an insight of the customer’s opinions , whether they are negative or positive.

In the Jupyter example, I have selected the tweets from an airline as the dataset for the sentiment analysis. I have performed text analytics (tokenization, stemming, vectorization,) in order to turn the text into values to be analysed .

I then applied TF-IDF and calculated TF-IDF for each word within a sentence.

I have also plotted the words that appear more frequently for positive and negative feedback.

The plot shows that the negative feedback seems to be more predominant amongst the tweets.

Fig 6

Total 287 words

# **5.0 References**

Australian bureau of statistics - sampling methods

<https://www.abs.gov.au/ausstats/abs@.nsf/0/A493A524D0C5D1A0CA2571FE007D69E2>

Stratified k-fold cross-validation

<https://www.geeksforgeeks.org/stratified-k-fold-cross-validation/>

Scikit learn stratified k fold cross validation

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html#:~:text=Stratified%20K%2DFolds%20cross%2Dvalidator,of%20samples%20for%20each%20class>.

Machine learning mastery cross validation for imbalance classification

<https://machinelearningmastery.com/cross-validation-for-imbalanced-classification/>

K fold cross validation

<https://machinelearningmastery.com/k-fold-cross-validation/>

Statistics: An introduction to sample size calculations

<https://www.statstutor.ac.uk/resources/uploaded/sample-size.pdf>

Sample sizes for classification models

<https://arxiv.org/abs/1211.1323>

Applying artificial intelligence and machine learning in robotics

<https://www.automate.org/blogs/applying-artificial-intelligence-and-machine-learning-in-robotics>

What is reinforcement learning

<https://www.geeksforgeeks.org/what-is-reinforcement-learning/>

Explaining reinforcement learning active and passive

<https://www.kdnuggets.com/2018/06/explaining-reinforcement-learning-active-passive.html#:~:text=Both%20active%20and%20passive%20reinforcement,that%20it%20can%20act%20on>.

Gini Index For Decision Trees

<https://blog.quantinsti.com/gini-index/>

Regression

<https://www.investopedia.com/terms/r/regression.asp>

NLP, AI, and Machine Learning, what’s the Difference?

<https://monkeylearn.com/blog/nlp-ai/>

Natural language processing and social media

<https://www.paldesk.com/natural-language-processing-and-social-media/>

sklearn.linear\_model.LogisticRegression — scikit-learn 0.24.2 documentation. (2020). Https://Scikit-Learn.Org/Stable/Modules/Generated/Sklearn.Linear\_model.LogisticRegression.Html. <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

# 

# **6.0 Appendix**

## 6.1 Datasets

Baseball

<https://mathcs.org/statistics/datasets/index.html>

Airline sentiment

<https://www.kaggle.com/welkin10/airline-sentiment>

## 

## 6.2 List of Figures

## Figure 1 Gini Index decision tree

A picture containing text, sign, different, screenshot

Description automatically generated

Figure 2 : Entropy decision tree

A picture containing text, sign, different, screenshot

Description automatically generated

Figure 3 A representation of the evolution of the salary per year

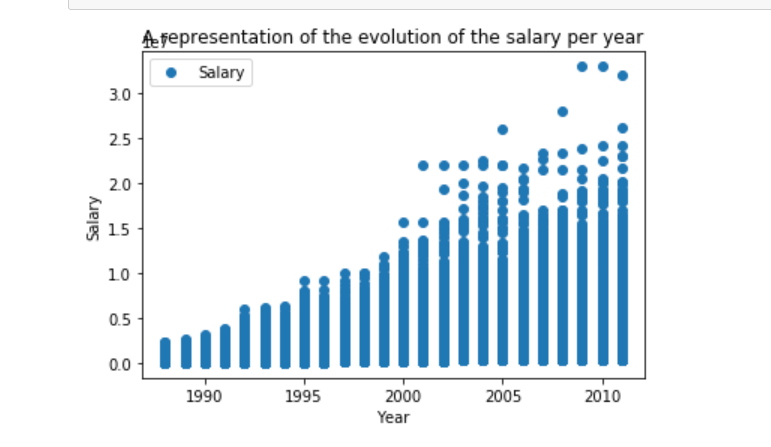


Figure 4 A prediction of evolution of the salary



Figure 5 A prediction of the evolution of the salary per Year

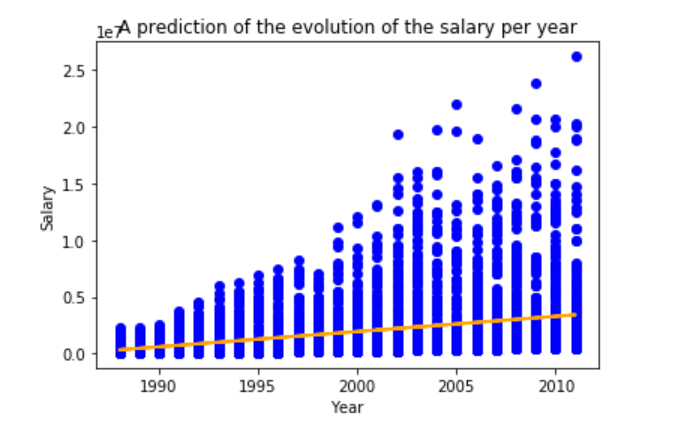


Figure 6 Sentiment Analysis for Negative values for Airline Data

