**Martin Byrne sba21243 Feb 2021 Machine Learning CA 3**

**Question 1: What are the major types of machine learning approaches? Briefly explain them with at least one example for each type.**

The 3 basic paradigms of machine learning are Supervised Learning, Unsupervised Learning and Reinforcement Learning.

Starting with Supervised Learning, this type of Machine Learning has supervision which enables intervention and correction of mistakes if the algorithm being used is incorrect, it also uses already labelled data. An example is to predict house prices, we can use of data on different types of houses and use results to predict the price of new houses.

Then, Unsupervised Learning which looks at unlabelled data and tries to find clusters or groups. A good example is looking at customer data and determining if there are any segments which will benefit things like targeted marketing campaigns.

Finally, is Reinforcement Learning, there is no data and there is no teaching involved either as it is about how an algorithm will interact with the environment and make decisions. One example is Youtube and video recommendations which evolve as more videos are watched.

**Assume you are working on a data set that deals with churning of a customers using a Machine Learning model. Which framework you would like to use to solve this problem?**

If examining customer churn using ML, we could use the CRISP-DM framework, Fig 1 shows the framework process. CRISP-DM is an acronym for Cross Industry Standard Process for Data Mining and is a widely used framework in Machine Learning projects. We will give an overview of how to apply it to our problem.

Firstly, understand the business issue, which is customer churn, then we need to look at the data coming from the customers that are being lost, this will need to be prepared so that proper analysis can be performed eg trends of specific areas that are losing customers.

Then ML models such as regression etc can be applied to test understanding of why this churn is happening and to test strategies for reversing this trend and how they are performing. We can evaluate their performance in test, then after evaluation and necessary fine tuning these can be deployed in the form of strategies to both prevent further and reverse existing churn.

Diagram

Description automatically generated

Fig 1 - CRISP-DM Process (Datasciencecentral.com, 2011)

**How much data will you allocate for the training, validation, and test sets? Briefly describe and justify your approach. (400 words, 20 marks)**

There are several different test train splits that can be used however the split that I would use assuming a reasonably large dataset would be 70/30 with the 30 being split into 15 for test and 15 for validation however this could change for a smaller dataset to enable a resampling on limited data samples. The reason being that we need enough training and test data to avoid fitting issues. There is also the need for algorithms to be performant so splits would need to be flexible to allow for tuning for better performance.

**Word Count (Not including questions, tables, calculations, and figure declarations 412)**

**Question 2: Suppose you are working on a classification problem. For validation purposes, you have randomly sampled the training data set into train and validation. The accuracy is high during the validation of modelling results. When you have tested Machine Learning (ML) model on unseen data, the testing accuracy is poor. Can you explain the reasons for this poor accuracy and explain further steps to improve this accuracy of your ML model. (400 words, 20 marks) Assessment**

The reason why we would see poor accuracy on testing despite high accuracy levels seen during our training is called overfitting, this is where we have customised our model too closely to the specificities of a single training set which makes the model much less generalised than it should be so that it could be used on unseen data.

For example, in Fig 2 we are selling products to customers,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AGE | GENDER | INCOME | RELATIONSHIP | PREVIOUSLY PURCHASED |
| 30 | M | 50K | ENGAGED | N |
| 35 | F | 65K | MARRIED | Y |
| 45 | F | 40K | SINGLE | N |
| 52 | M | 45K | MARRIED | Y |
| 21 | F | 28K | SINGLE | N |
| 39 | F | 22K | MARRIED | Y |
| 61 | M | 80K | DIVORCED | Y |
| 72 | M | 20K | MARRIED | N |
| 27 | M | 42K | SINGLE | N |
| 56 | F | 70K | DIVORCED | Y |

Fig 2 Example Table

however, if we try to identify too much with certain criteria such as aiming for previous purchasers who earn over 30k, are not married and are less than 60 on the training dataset then this is too complex, and we will have trouble achieving good performance on a new dataset as it is difficult to generalise.

There are some steps which we can take to rectify overfitting, the first is to train with more data if available, this may not always be the case in the real world however if possible, use more data. In the above example there are only 10 examples but if possible, it should be many more.

If this does not help, then start examining the dataset to see if there are features that may not be relevant and can be removed. Some of this can be done automatically but there are times when it will need to be done manually. For instance, in the above table, we may discover that being a previous purchaser is not that important and this factor can be removed from consideration to improve fitting.

The next action that we can take to improve our fitting problem is what is referred to as early stopping, if we have a number of iterations this can improve our model, however there is a point when that improvement stops and the ability to generalise diminishes, knowing and stopping before this point is reached prevents the deterioration from reaching the point of overfitting.

There are also other techniques available such as regularisation or using ensembles to try and avoid overfitting. It also must be mentioned that we could try and detect overfitting prior to test but sometimes that is not an option so we should try as many techniques as possible to tackle this issue.

It should also be stated that in trying to avoid overfitting we should avoid falling into underfitting, there is a sweet spot between the two which offers both good generalisation and good performance.

**Word Count (Not including questions, tables, calculations, and figure declarations 403)**

**Question 3: a) Briefly describe the design issues that the data scientists faced during the development of models based on Decision Trees.**

**Diagram, schematic

Description automatically generated**

Fig 3 Decision Tree (https://www.facebook.com/kdnuggets, 2020)

There are different design issues the data scientists face when developing models using decision trees, an example of a tree is in Fig 3.

Several of these are related to splitting, there are several questions about splitting the first of which is the best split, splits should be made on the best attributes to split on but that is not always the case. The next concern is how these attributes should properly be split in order to avoid unnecessary parts of the tree. Finally, is when should you stop splitting, splitting grows the tree and leads into overfitting if the tree becomes too complex. These questions are why we have techniques such as Gini Impurity checks to help us find the best answers.

**Demonstrate the difference between Gini Impurity and Entropy in a Decision Tree Models.**

Gini Impurity and Entropy are both impurity measures for classification in decision tree models but there are some differences in them. First difference is the fact that using logarithms in calculations makes using Entropy both slower and more complex when doing calculations, so Gini is faster. Second there are differences in intervals, Gini is 0 to 0.5 on the Y axis and Entropy is 0 to 1.

**Provide an example that demonstrate the usage of Gini index and Entropy on the decision trees. (300 words, 15 marks)**

An example of using both Gini Impurity and Entropy is shown in a table in Fig 4. This is just an example table that has been made to demonstrate the decision whether to go for a walk depending on the weather.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DAY | Weather Outlook | Temperature | Humidity | Wind | Go For Walk |
| 1 | Sunny | Warm | Low | Breeze | Yes |
| 2 | Rain | Cool | Low | Gale | No |
| 3 | Rain | Cool | High | Breeze | Yes |
| 4 | Sunny | Warm | High | Breeze | Yes |
| 5 | Sunny | Cool | Low | Gale | No |
| 6 | Rain | Warm | High | Gale | No |
| 7 | Sunny | Warm | High | Breeze | Yes |
| 8 | Sunny | Cool | Low | Breeze | Yes |
| 9 | Rain | Warm | High | Breeze | Yes |
| 10 | Sunny | Cool | High | Gale | No |

Fig 4 Example Table

Fig 5 shows the formulas for both calculations.

Text, application

Description automatically generated

We can look at the Gini Index for Weather and Go For Walk, I used the Miniwebtool (Miniwebtool.com, 2021) for calculations

Sunny = 6/10

Rain = 4/10

Sunny & Go For Walk 4/6

Rain & Go For Walk 2/6

Gini Index = 1 – ((4/6) ^ 2 + (2/6) ^2)

4/6 = 0.67

2/6 = 0.33

0.67 ^ 2 = 0.4489

0.33 ^ 2 = 0 .1089

0.4489 + 0.1089 = 0.5578

1 – 0.5578 = 0.4422

Then we can look at using Entropy on the same numbers for Sunny & Go For Walk

0.67 log = - 0. 5777

0.67 \* - 0. 5777 = -0.3736

0.33 log = -1.5994

0.33 \* - -1.5994= -0.5278

-0.5278 - -0.3736 = -0.9014

As can be seen these calculations lead to different outcomes, these calculations can also be done for all the columns in the table, and they would need the same method applied.

**Word Count (Not including questions, tables, calculations, and figure declarations. 300)**

**b) What is Regression? Explain the difference between L1 and L2 regularizations. Briefly discuss and highlight the difference between Linear and Logistic Regression by considering any dataset of your choice. (300 words, 15 marks)**

Regression is a set of statistical techniques that are used to find a relationship or correlation between variables that are called dependent variables and independent variables. These are also referred to as target and predictor variables.

The L1 and L2 regularisations are parameter control mechanisms, they are usually applied to different regression methods. L2 is applied when using a regression type called Ridge regression and the purpose is to shrink coefficients evenly, these coefficients are used for both training and fitting an additional constraint. L1 on the other hand is used by Lasso regression which is an alternative to Ridge, this is used to restrict coefficients so they are close to 0, this has the consequence of ensuring that when using Lasso that some coefficients are 0 exactly.

There are many types of regression however 2 of the most popular are Linear Regression and Logisitic Regression.

Diagram

Description automatically generated

Fig 6 (ODSC - Open Data Science, 2019)

Linear Regression and Logistic Regression have several differences, linear is used to solve regression problems while logistic is used in classification problems. This then leads on to linear being used on continuous values whereas logistic is used on categorical values. They have different looking lines, linear has the best fit line, logistic has the S curve which shows the classified examples see Fig 6.

If we consider a dataset that looks at house prices, any set of the many available can be looked at. The differences between Linear and Logistic regressions becomes much easier to understand, house prices will track for example the relationship between floor space and house prices so that a prediction can be made about new houses using Linear Regression. Logistic can be applied to the type of house that is being examined, for example Bungalow, Townhouse, Detached etc. It can also be used for example on a house location as location is a categorical value as it is either in a location or not.

**Word Count (Not including questions, tables, calculations, and figure declarations. 314)**

**Question 4: a) Explain the role of Reinforcement Learning (RL) models on a larger canvas of Machine learning and identify the suitable areas for the implementation of reinforcement learning models. Distinguish the active and passive RL approaches. (250 words, 10 marks)**

Looking at the role of Reinforcement Learning how it relates to the wider picture of Machine Learning is general necessitates understanding of what RL is in the first place. This is a type of Machine Learning that essentially learns by trying something and any errors that occur.

Then look at differences between RL and other ML techniques, RL takes the successes and failures as rewards and punishments from learn in an environment, the RL agent then tries to have as many rewards as possible. This differs from Unsupervised Learning where the goal is to find similarities or otherwise between data points and Supervised Learning which tries to learn some generalisations from a set of examples so that they maybe applied elsewhere.

There are several different areas for implementing RL models, these include self-driving vehicles which could learn things like what are the best paths to take while driving to avoid bumps etc. Another area where RL could be implemented is in Automation such as Engineering to control things like how much energy is being used and how to reduce this.

We must compare active and passive approaches to RL, active RL has an agent that must make a decision on what it has to do as there is no fixed policy in place that can be used whereas in the passive form of RL there is a fixed policy for the agent to act on. There are different algorithms available, and they will depend on whether the activity is model-free which is the real world or model-based which is simulation based.

**Word Count (Not including questions, tables, calculations, and figure declarations. 261)**

b) **What is role of Natural Language Processing (NLP) in Machine Learning? Highlight the impact of NLP for social media data processing.**

Natural Language Processing has an increasing role within Machine Learning, simply put there is an enormous demand for allowing the interpretation of human language. Once we can accurately interpret language then comes the possibility of using automation such as chatbots to keep websites operating 24 hours a day when staff might only be there for half that.

Natural Language Processing has had a large impact on Social Media data processing especially on certain networks, Twitter being the prime example. Tweets have been used to test and improve sentiment analysis models as they need to be examined for context, subtext, sarcasm and other things that are more understandable when using spoken communication.

**Consider any dataset or paragraph comprised of (at least 200 words) of your own choice and show that the outcomes of the text analytics operations (for example, tokenization, normalization, stemming and etc.) for text classification. Perform sentiment analysis on the chosen dataset and calculate TF-IDF. (400 words, 20 marks)**

For the dataset operations I decided to use the TextBlob library (Readthedocs.io, 2020) as it contains all the functionality that I need to use for the example. All code is in the attached notebook and screenshots are from that. The first thing that we look at is imports, see Fig 7 below.

Text, letter

Description automatically generated

Fig 7 Notebook Imports

Then we import our data which is this case is authored by Thomas Klikauer and published ZNet (Klikauer, 2021), Fig 7

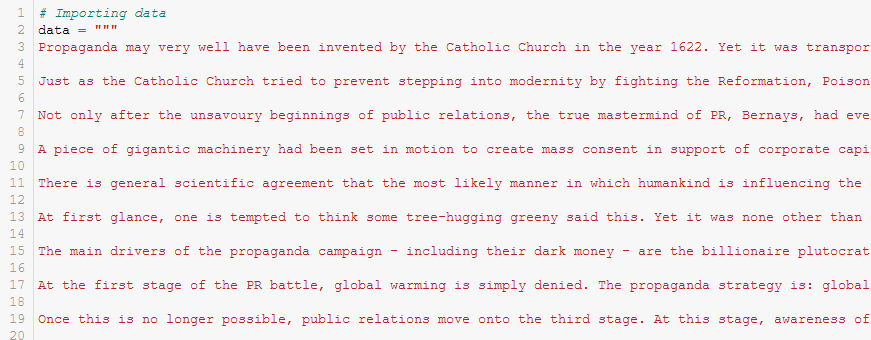


Fig 7 Data Import

After our data import then we create a TextBlob object and tokenise our text into either sentences or words, see Fig 8

Text

Description automatically generated

Fig 8 Word tokenising

When we tokenise the words it will look like Fig 9

Text, letter

Description automatically generated

Fig 9 Word Tokens

We can also tokenise our data into sentences, see Fig 10

Text, letter

Description automatically generated

Fig 10 Sentence Tokens

After that then we can look at lemmatizing data and stemming our dataset (Jonathansoma.com, 2017), see Fig11.

Graphical user interface, text, application, email

Description automatically generated

Fig 11 Data Stemming and Lemmatizing

After that then we can have a look at the sentiment analysis, this is done using the built-in sentiment analysis tool, it should be stated that this is probably not good enough to use on a full real-world application as is but as it is built on the NLTK library it has some very good features for beginners, see Fig 12. Polarity is a floating-point numerical value that goes between -1 and 1 where 1 will be a positive statement and -1 will be a negative statement. Subjectivity is also a floating-point numerical value, it goes from 0 up to 1 where 0 is objective and 1 is subjective.

Graphical user interface, text, application, email

Description automatically generated

Fig 12 Sentiment Analysis

TF-IDF is an acronym which stands for Term Frequency – Inverse Document Frequency and it is a measure of how relevant a word is in a particular document (MonkeyLearn Blog, 2019). Calculating this measure involves the multiplication of 2 separate documents. We will look at the code from Fig 13 for calculating TF-IDF (Loria, 2013). We will also have to add another couple of datasets for our calculations, the second dataset is from Frances Madeson (Madeson, 2021) and the third is from The Red Nation (The Red Nation, 2021)

Graphical user interface, text, application, email

Description automatically generated

Fig 13 TF-IDF Part 1

Once the methods have been set up then the datasets can be passed in as can be seen in Fig 14

Graphical user interface, text

Description automatically generated

Fig 14 TF-IDF Part 2

One final note is that if you look at the top words for document 1 and 2 these will probably be eliminated by using the stop words package and will have a very different TF-IDF as a result because stop words will eliminate commonly used words which will skew this type of calculation.

**Word Count (Not including questions, tables, calculations, and figure declarations. 405)**

**Reference List**

Datasciencecentral.com. (2011). *CRISP-DM – a Standard Methodology to Ensure a Good Outcome*. [online] Available at: https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome [Accessed 18 May 2021].

https://www.facebook.com/kdnuggets (2020). *Decision Tree Algorithm, Explained - KDnuggets*. [online] KDnuggets. Available at: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html [Accessed 22 May 2021].

Jonathansoma.com. (2017). *TextBlob spaCy sklearn lemmas stems and vectorization*. [online] Available at: http://jonathansoma.com/lede/algorithms-2017/classes/text-analysis/textblob-spacy-sklearn-lemmas-stems-and-vectorization/ [Accessed 23 May 2021].

Klikauer, T. (2021). *» Anti-Global Warming PR*. [online] Zcomm.org. Available at: https://zcomm.org/znetarticle/1046839/ [Accessed 22 May 2021].

Madeson, F. (2021). *» Sunrise Activists Marching 400 Miles to Demand Green New Deal*. [online] Zcomm.org. Available at: https://zcomm.org/znetarticle/sunrise-activists-marching-400-miles-to-demand-green-new-deal/ [Accessed 23 May 2021].

MonkeyLearn Blog. (2019). *What is TF-IDF?* [online] Available at: https://monkeylearn.com/blog/what-is-tf-idf/ [Accessed 19 May 2021].

ODSC - Open Data Science (2019). *Logistic Regression with Python - ODSC - Open Data Science - Medium*. [online] Medium. Available at: https://medium.com/@ODSC/logistic-regression-with-python-ede39f8573c7 [Accessed 22 May 2021].

Readthedocs.io. (2020). *TextBlob: Simplified Text Processing — TextBlob 0.16.0 documentation*. [online] Available at: https://textblob.readthedocs.io/en/dev/# [Accessed 23 May 2021].

The Red Nation (2021). *» To End Fossil Fuels, End Settler Colonialism*. [online] Zcomm.org. Available at: https://zcomm.org/znetarticle/to-end-fossil-fuels-end-settler-colonialism/ [Accessed 23 May 2021].