11/21/2021

Report and Dataset

AI & Machine Learning – CA 2

Bridgeman Thomas
LETTERKENNY INSTITUTE OF TECHNOLOGY

Contents

Introduction

Artificial Intelligence and Machine Learning are used virtually everywhere in the modern age. This is because it can analyse large amounts of data and apply different techniques to predict outcomes more effectively than humans can or simply predict outcomes have no hope in predicting. For this assignment, I have chosen two different datasets to study and apply the machine learning techniques that I have learned this semester.

Dataset

The dataset I have chosen is the 'Milk Grading' dataset from Kaggle (Prudhvi GNV 2021). The domain of this dataset can be classified as environment, agriculture and food. This dataset is under the EU OPD Legal Notice and is therefore available for public use. I believe that storing this data on Google Drive may be the most beneficial but storing and reading as a CSV file would also work.

The data in this dataset was manually collected and used to predict the quality of milk. There are 7 variables for each sample of milk in the dataset – pH, Temperature, Taste, Odour, Fat, Turbidity (a measure of clarity) and Colour. These 7 variables are used to get a Quality rating of the milk. The inspiration for the people who made the dataset was to use machine learning in the dairy industry. The inspiration for me was my love of milk and I also have an interest in the future of the dairy industry as it is a large part of the Irish agricultural sector and has many jobs relying on it. With the recent covid-19 pandemic, food wastage was brought into the limelight where farmers unable to sell their produce to restaurants and hotels ended up having to discard much of their food such as milk. A huge amount of milk was already wasted every year due to it going off. My main goal is to apply machine learning to samples of milk as if it is used in the industry it could help decrease the amount of milk discarded each year, saving people money and putting less pressure on farmers and helping the environment. The results from the dataset could also give an insight into the amount of milk that is not drinkable each year and force us to come up with better methods of storing and preserving it. (Yaffe-Bellany and Corkery 2020)

Cleaning

I didn’t have much cleaning to do with this dataset as I am using all data present in the table and the data itself had all correct and no null values. I did however check the data for null values to verify this fact. As well as that I put my columns of pH, Temperature and Colour on a standard scalar for sklearn. This was added to its data frame and the other data of odour, taste, fat, and turbidity was added to another data frame before I combined them. The standard scalar removes the mean and scales it to unit variance. This allows me to have all my data on a similar level and will look better upon graphing data as the points used will be all-around 0 and 1. The results from my data will hopefully be more accurate because of this. As stated the rest of my data was already suitable for use however, if needed I could have cleaned the data using pandas to remove rows with null or incorrect values.

Machine Learning Techniques

To get results I will have to use machine learning techniques. These are known methods used by AI models to get accurate predictions for the data being used. They have been developed and tested over the years and are now able to be done efficiently with libraries. For this dataset, I plan to use several datasets that should help give me adequate predictions. Decision trees are the first technique I plan to use, and I plan on using the Random Forest techniques alongside it. The next technique I would like to use would be the K-Fold Cross Validation and compare it to a Train/Test

model to see if the results end up being the same. The final technique I plan on using is K-Nearest Neighbour to see clusters of data.

Random Forest and Decision Trees

A decision tree is a structure often based on the design of a flowchart and used to decide the outcome or output of a series of events. It is given the name tree as the internal nodes are considered tests and the external nodes often called leaf nodes are categories. An event or each question will decide what the next event or the output will be. Decision trees may not always be the most optimal tree as when it works its way down the tree, it picks the path that will result in the least entropy or randomness. But they have proved to be both useful and successful in many industries especially in job searching. (Nilsson 1996)

Random Forests are used as decision trees that often are overfitted or correspond too closely to their training data and struggle in adding new data. Therefore, random forests are used, this constructs alternate decision trees and the tree selected most is used. The data is then used in multiple models on subsets of the training data, and it then combines the predictions of all models. This is often called Bagging or Bootstrap Aggregating. (Louppe 2015)

For my Milk Grading dataset, I will take my clean dataset and import tree from sklearn and use it to construct the decision tree with my columns headers and the data associated with them. Several other Python libraries will be imported then to display an image of the decision tree. I should then be able to use a Random Forest Classifier and enter in my data values and get output telling me the quality of the milk.

A decision tree was the first technique I thought of when I saw the data. This is because each of the columns of data affects the final column which is the actual quality of milk. There are far too many variables present to create an accurate decision tree myself but by using machine learning I hope it can provide this information for me. I expect this to provide me with a decision tree and I can use this to test my inputs for each of the columns to get an output on the quality of the milk.

Train/Test and K-Fold Cross-Validation

Train/Test is a machine learning technique that measures the accuracy of a model. Data is divided usually 80 or 70 per cent into the training set and 20 or 30 per cent into the testing set. The training data is used to create the model while the tester shows us the accuracy of the model. For it to work effectively, both data subsets must be selected randomly and contain a wide range of representatives and feature outliers in the data.

K-Fold Cross Validation is like train/test and can be used as an alternative to it. It works by splitting the data into K randomly assigned segments often called folds. One segment is used as the test data set and after each of the remaining K-1 segments have been trained, the performance is measured against the test set. The average of the K-1 r-squared scores is taken and tells us the accuracy of the model. (Refaeilzadeh *et al.* 2016)

Applying this technique to my Milk Grading dataset, I will once again import several methods from sklearn. I will then split my data into training and training and testing datasets. I will likely try with both a 70/30 split and an 80/20 split to see which produces better results but should in theory be very similar anyways. After this, I will apply K-Fold to the data using cross_val_score on my data and get the average for each fold of data. Again, here I will try several values for the number of folds to get the most accurate results for my dataset. The mean value of the folds is then calculated, and we can compare this to our train/test value. An SVC model will have to be created for this work and can be

either linear or polynomial and once again, both will be used to see which gives the more accurate representation of my data.

I am using train/test and k-fold cross-validation on the data so I can get some accurate readings or scores on the data. The ideal result here would be to get a value as close to 1 as possible. I expect the results to be good as there is a lot of data present for it to work with producing more accurate results.

K-Nearest Neighbour

K-Nearest Neighbour or KNN is a supervised machine learning technique used for both classification and regression. It works on the principle that similar things will exist near or in close proximity to each other. The programme will have to be run several times in an attempt to get the value for K that reduces the number of errors present. The nearest neighbours in the model will contribute more to the average than the values further away from your K-value. (Nilsson 1996)

For my dataset of Milk Grading, I believe I will have to have my data in a DataFrame which should already be the case. I will use some of the same code I used above to get a train/test split of my data and might be able to create a method here because of code re-use. I will then import KNeighboursClassifier from sklearn and use its methods to train the data and then test it. I should then be able to produce a graph that will tell me the best value to use for K. Once I have the best value for K I can create methods that get the distance and get the neighbours.

K-Nearest Neighbours was chosen as I believe it could be interesting to see how different sets of milk ratings relate to each other. I expect the results to be all grouped in a similar area with a few outliers for data with high temperatures or pH etc. I hope that it produces accurate data I can learn from once again.

Conclusion

Once again, my main goal for this project is to use the machine learning techniques stated in the report to help predict the quality of milk so that methods to prevent milk waste from occurring become more common. I also hope that I become familiar with the machine learning techniques I have to decide to use and that they help in providing me with interesting data predictions.

References

Louppe, G. (2015) 'Understanding Random Forests: From Theory to Practice, the *University of Liège*, available: http://arxiv.org/abs/1407.7502 [accessed 22 Nov 2021].

Nilsson, N.J. (1996) *Introduction to Machine Learning. An Early Draft of a Proposed Textbook* [online], Stanford University, available: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=PJguPTEAAAAJ&citation_for_view=PJguPTEAAAAJ:35N4QoGY0k4C [accessed 21 Nov 2021].

Prudhvi GNV (2021) Milk Grading [online], *Kaggle*, available: https://kaggle.com/prudhvignv/milk-grading [accessed 22 Nov 2021].

Refaeilzadeh, P., Tang, L., Liu, H. (2016) 'Cross-Validation', in Liu, L. and Özsu, M.T., eds., *Encyclopedia of Database Systems*, Springer: New York, NY, 1–7, available: https://doi.org/10.1007/978-1-4899-7993-3_565-2 [accessed 22 Nov 2021].

Yaffe-Bellany, D., Corkery, M. (2020) 'Dumped Milk, Smashed Eggs, Plowed Vegetables: Food Waste of the Pandemic', *The New York Times*, 11 Apr, available: https://www.nytimes.com/2020/04/11/business/coronavirus-destroying-food.html [accessed 25 Nov 2021].