Assignment 3

Name: Shounak Rangwala

Netid: snr85

1. Clustering is a machine learning algorithm that is used mainly in unsupervised learning techniques. When we are given a dataset, we may not know underlying relationships the individual datapoints have with each other. We may be aware of some but there may be others ( not so obvious) that have more significance. TO find these we run clustering algorithms. The machine goes over these data points and identifies different parameters to judge them on and sorts them into various clusters. Members of each cluster exhibit similar attributes.

The uses of this are:

1. Search result grouping on browsers
2. Anomaly detection in experiment results
3. Social network analysis

(Source: <https://developers.google.com/machine-learning/clustering/overview> )

1. K-means clustering is a popular clustering algorithm. In this algorithm we define the value of k. This is the number of clusters we want to separate the data into.

The process is as follows:

1. Select K random points as cluster centers.
2. Assign data points to nearest cluster center based on a distance metric (Hamilton/Euclidean/Manhattan)
3. Calculate the mean (centroid) of the instances in each cluster to come up with newer centers.
4. Repeat steps 2, 3

(Source: previous course slides)

Application: city planning – identifying groups of houses bases on the type of house, value and geographical location.

1. The pic below shows the clustering. I preprocessed the data to scale it down to between 0 and 1. I used the entire dataset to run the clustering on. The number of clusters were fixed to 5 in this case however we can go to any number of clusters.

Chart, scatter chart

Description automatically generated

1. Logistic regression: This is a classification algorithm that deals with classifying data into discrete classes. Usually this is a binary variable that we are predicting by assessing the probabilities of different factors that affect it. It transforms its output through a logistic sigmoid function to return a probability value which is used to map it into discrete classes.

Logistic regression can be used as a layer above any known regression method like linear regression or such. Basically we need the probabilities to be in the range of 0-1. This is done through the sigmoid function which goes like this:

P(z) =

The z is the output of the linear regression. Theis can be done after gradient descent, regularization and all those methods. After you have these probabilities, you can map them to different classes based on the different rules you decide. So for example if the probability ids between 0-0.4: the output is “dog” , 0.4-0.8: the output is “cat”, 0.8-1: the output is “mouse”.

Applications:

Logistic regression can be used in predicting the possibilities of certain diseases in people based on different parameters like (age, sex, body mass index, results of various blood tests, etc.)

Difference between logistic regression and linear regression is that in logistic regression, we take the possibilities (which are continuous quantities) and map them into different classes. While in linear regression, we predict continuous values and do not segregate them into different classes based on probabilities.

1. The image below shows the accuracy of the logistic regression. I had difficulty in running the model because it was not converging in the number of iterations set by default. Increasing the number of iterations solved the issue. The score was 0.986 accuracy.

Text

Description automatically generated

(Source: <https://github.com/mGalarnyk/Python_Tutorials/blob/master/Sklearn/Logistic_Regression/LogisticRegression_toy_digits_Codementor.ipynb> )

1. For each attribute,

For each value of the attribute, make a rule as follows:

count how often each class appears

find the most frequent class

make the rule assign that class to this attribute-value

Calculate the error rate of the rules

Choose the rules with the smallest error rate

Advantages:

1. It is simple, quick and baseline in nature so gives you a good intuition of what to expect from more complex algorithms
2. It usually gives an answer that is reflective of the final answer

Disadvantages:

1. It just partitions the data once according to some rule that is most frequent, this cannot give you the better understanding of the relations/rules that exist between different attributes of the data
2. Not reliable for regression analysis where the output is a stream of continuous values
3. A decision tree that performs extremely well on given data is most probably overfitted. Which means that it not only is potraying the underlying trend (as it should) but also accommodating the outliers which are just noise. And the rules it sets up for these outliers are the reason this decision tree will not perform well on different data.
4. Information gain is the criteria for deciding upon the attribute to be used in further splitting the node of the decision tree such that the decision tree is as small as possible.

InfoGain(attribute1) = Info(ParentNode) – Info(Node[i] for i in no.of classes of attribute1)

The info() function above is the entropy value which is given by :

Info(a,b) = entropy(a/(a+b),b/(a+b))

Entropy(p1,p2) = -p1logp1 – p2logp2

The logs are in the base of 2

So if there are multiple attributes based on which you can split the node, we select the attribute such that the new nodes formed have an info value that will give you the most gain i.e. the difference in the info values is the greatest

Diagram

Description automatically generated

In the above example; the parent node has 14 values that have independent attributes of outlook,temperature, humidity and windy. The 14 values of the class we are observing are 9 Yes’s and 5 No’s.

The parent node has Info(9,5) = 0.940 based on calculation above

-[9/14 \* log(9/14) + 5/14\*log(5/14)]

In case of the outlook:

The child nodes have Info(2,3) = 0.971, Info(4,0)=0, Info(3,2)=0.971.

So the total Info of the child layer of the tree is :

Info([2,3],[4,0],[3,2]) = 5/14 \* 0.971 + 4/14\* 0 + 5/14\* 0.971 = 0.693

So the Gain(outlook) = 0.940-0.693 = 0.237

Similarly we calculate the Gains for the other attributes and we get the following:

Gain(temperature) = 0.029

Gain(humidity) = 0.153

Gain(windy) = 0.048

The maximum gain is from the outlook attribute so we will split the node based on the outlook attribute.

1. The total number of different instances are : 3\*4\*2\*5\*3 = 360 instances
2. Instance based learning is where the training examples are stored in the system as it is. Then for each test example, we compare it to all the stored training data and find the closest training data point. We assign the test data the same class as that training data point. This is different from conventional machine learning because we are not learning the weights by back propagation to predict the values, we are making a huge look-up table.

Instance based learning has similar applications as that of KNN (k nearest neighbours). This is because we always have to store some of the training data in memory while calulating the neighbour distances. 2 applicatoins of IBL:

1. Predicting stock market prices
2. Data compression
3. Data mining has serious ethical ramifications when the data being used has the potential of revealing personal secrets about individuals that they would generally not reveal themselves. Data mining used on government records has had the potential of identifying state employees and revealing their social security numbers to the public which is obviously a crime.

Sometimes the bias which exists in society gets exposed during the data mining process. This is something we must not do. For example, stereotyping people based on race and gender and sexual orientation while approving loans etc is clearly wrong and it must be ensured that this kind of data is not collected for analhsis otherwise underlying bias may affect the decisions unfortunately.

2 situations where data mining would be advantageous would be:

1. National security issues: sometimes tapping communications which are meant to be private has led to prevention of terrorist attacks.
2. Improving shopping experience: analysing shopper tendencies helps rearrange products next to each other so that in increases ease of shopping and also sales.

2 situations where this is not permissible:

1. Redlining areas
2. Giving up confidentiality of students library records
3. Takeaways from the session:
   1. There will always be new technologies- just focus on stenghtening your basics. So many begineers focus on learning the technology/software but don’t work on their basics so they are often left lacking.
   2. Data analytics are used widely from banks to pharmaceuticals etc. Big data is being used every where and all business decisionsare being made increasingly by the help of data. It is important to get into the industry with sufficient knowledge about the newest developments in both business and technology.
   3. Always have a network of contacts especially in times like these when you could be laid off anytime. One of the alumni was laid off due to COVID but because of his network as part of the MBS program he was able to secure another job.
   4. More projects you have on your resume- better are your prospects of landing a job. This is true mostly for new students who have not had that much of work experience.
   5. All the professors and the alumni are super open to questions and queries and are very helpful.

A screen shot of a person

Description automatically generated