Assignment-2

Name: Shounak Rangwala

Netid: snr85

1. Naïve Bayes is a machine learning technique that is derived from the Bayes theorem and assumes that all the attributes of a given instance are independent from each other and can be applied in the bayes theorem formula.

It is called Naïve because in real life scenario, different attributes of an instance have probabilities that are dependent on each other and cannot be plugged into the Bayes theorem as independent probabilities.

1. Gaussian Naïve bayes is an extension of naïve bayes method. In this method we try to calculate/apply bayes theorem in predicting from a continuous range of values. Gaussian distribution is used to represent this continuous range. The probability we use in the naïve bayes is similar to the probability distribution function which gives the area under the graph. Also gaussian distribution is easy to calculate this upon because we only need to know the standard deviation and the mean to describe the plot.

In the formula below, μ is the mean and σ is the standard deviation.

Text

Description automatically generated

1. The number of mislabeled examples in the iris dataset obtained after running the Gaussian naïve bayes classifier are: 6

Graphical user interface, text, application, email

Description automatically generated

1. The MSE error is 4431.62 for single attribute linear regression (BMI)

The equation is y = 957.80\*x + 153.12

Graphical user interface, text, application

Description automatically generated

1. I used the BMI and the Bp attributes of the class instances to predict the output/target. The equation is:

Y = 802.78\*BMI + 381.115\*BP +149.48

MSE in single linear regression : 4431.622079882914

MSE in multi linear regression : 4182.341911070552

Text

Description automatically generated

1. Gradient descent is a technique used in machine learning where iteratively, the predicted values is changed in the opposite direction of the slope of the loss function. Basically we are trying to reduce the loss in the predicted value and the actual value in a given minima.

In linear regression, after we get the predicted coefficients and the slope of the function, we should apply gradient descent to the loss function. The loss function can be the MSE loss.We differentiate this function wrt the slope of the loss and get the new slope. We calculate the loss and iteratively go on with this method till loss is very small. This way we get to the best prediction model.

1. Regularization helps in preventing overfitting with gradient descent. As we try to minimize the loss function so that the predictions are the most accurate, we may have unknowingly overfitted the model. By that, I mean that the same model cannot be used for another dataset because it is trained too well on one. Regularization helps reduce the variance in the model without losing out on any important properties of the data. The only danger we run into is that too much regularization may lead to underfitting the model.

I am implementing the ridge regularization with simple linear regression on the diabetes set. As you can see, the MSE error after regularization has decreased.

Graphical user interface, text, application

Description automatically generated

1. Decision trees are a supervised machine learning technique. We start dividing the dataset into different sections based on some rules which judge the data samples’ attributes and classify them into 2 or more categories. It creates a tree like structure and thakes logarithmic time to classify new data into different classes.

Information gain is the difference between the entropy of the dataset before and after the separation by the tree. The information gain is maximum if the entropy( randomness of the data prediction) after is very small as compared to the entropy before the classification.

Gini index: it is the measure of how if a random element is incorrectly classified. Gini favours the large partition of the dataset whereas the information gain prefers the small partition of the dataset. Attribute with the lowest gini is preferred.

(Source : <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html>)

1. Advantages of Decision trees over Linear regression:
   1. Decision trees are better for finding non linear solutions while the linear regression technique gives a linear solution. Non-linear solution can give a better understanding of the dataset.
   2. Decision trees can handle collinearity better than the linear regression
   3. Decision trees are better for when the variables in the dataset are predominantly categorical and independent.

Example: More complex datasets like loan approval from a bank consider many different factors

Advantages of linear regression over decision trees:

1. Decision trees calculations are more complex than the easier linear regression line-fitting calculations
2. Linear regression takes less time complexity than decision trees.
3. Linear regression have better space complexity than decision trees

Example: More linear dataset like predicting the house price based on the size of the house

Source:

1. <https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222#:~:text=Decision%20trees%20supports%20non%20linearity,be%20having%20better%20average%20accuracy>.

2. <https://medium.com/@dhiraj8899/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a>

1. I used the DescisionTreeRegressor with a max\_depth of 3 on the diabetes dataset.

Graphical user interface, text, application, chat or text message

Description automatically generated

1. Missing Completely at Random: Every datapoint has a probability of being missing in the data set. If the probability is the same for each case it means that the data is missing completely at random. The casue of missing is unrelated to the data and the way to deal with this is to completely ingonre these missing datapoints. This leads to loss of data but it is unevitable. MCAR is the most unrealistic scenario for missing data.

Missing Not at Random: This means that the probability of missing varies for reasons that are unknown to us. To handle this we need to produce more data that about what causes the data to go missing. We could also perform what-if analysis to find how sensitive our results are in different scenarios.

Source: <https://stefvanbuuren.name/fimd/sec-MCAR.html>

1. This sound like we want to maximise precision at any cost. The true positives have to be high while the false negatives are not our concern. To boost this we would favour the covering algorithm like PRISM.
2. The 4 different types of learning techniques are :
3. Classification learning- the system is presented with a classified set of examples and based on this learning it will try to classify unseen examples. Eg. The play or not to play judgement based on the weather
4. Association learning- looks for associations between different attributes of the data irrespective of whether they are used in predicting a particular class value. Eg. Deriving association rules in the weather dataset that gives relations between the different attributes themselves.
5. Clustering- the model groups data points into clusters according to some underlying property that the machine learns through unsupervised learning. Eg. Clustering the iris dataset into clusters identifying different types of irises.
6. Numerical prediction- the output to be predicted is not a discrete quantity but a numerical value that can lie in a wide range of possible values. Eg. Predicting CPU performance score based on the data given
7. Types of categorical data:
   1. Nominal data – these are mutually independent and unordered in nature. Eg. Male/female.
   2. Ordinal data- these are mutually exclusive and ordered in nature. Eg. Is the temperature cold, mild, warm or hot?

Types of numerical data:

1. Discrete data – values that are whole numbers eg. Number of homeruns in a season by a batter
2. Continuoius data – values that can be taken from a range of values eg. RBI for a batter in a season
3. Data integrity: This is the averall completeness and accuracy of the data stored or used for the data analytics. If the data’s integrity is secure, then we can confidently use that data even in the future for carrying out analytics. Data integrity ensures that your data is safe from malicious attacks (hackers). Many businesses collect data and if the integrity of that dat is high, they can safely store it and use it in the future to carry out predictive analysis or any other kind of analytics without having to collect new data or worry about data cleaning.

(Source: <https://www.talend.com/resources/what-is-data-integrity/>)

4 causes of lower data quality:

1. Data incorrectly entered: Numan error by employees in entering data in the records is the most common reason to have poor data quality.
2. Data migration and conversion: When we migrate from an old to a newer, upgraded system, the data may be lost or get corrupt in the process. Sometimes the data format may not be compatible with the new system and that will lead to poor data quality.
3. Fragmentation of the system: earlier there would be just one system that would be inputting data. But with the distributed systems used increasingly in businesses these days, there is a possibility of redundant data.
4. Excessive data: Data collected is so much that the time it takes to clean and prepare the data is more than the time before it is used in analytics and hence the results of the analytics will be compromised and then the data quality will be poor.

(Source: <https://docs.dhis2.org/master/en/implementer/html/reasons-for-poor-data-quality.html#:~:text=There%20are%20many%20potential%20reasons,between%20different%20paper%20forms>

<https://www.blazent.com/top-5-causes-poor-data-quality/>)

Ways to prevent poor data quality are:

1. Fix the data source : Most corrupt, poor quality data used by businesses is used from sources that haven’t properly collected/ cleaned the data. If we want to continue using data from these sources then the only way to improve data quality is by ensuring the source collects better data.
2. Fix data in the ETL phase: Even if the data cannot be improved at the source, we can try to process the data and clean it during the extract, transform and load stage before we add the data into the database and use if for further applications.
3. Apply entity resolution: If we have multiple entries of the same data in our database, this method helps us identify which entries are interrelated so that we can elect one of them and remove the redundancy and improve operation efficiency.\

(Source: <https://www.redpointglobal.com/blog/4-ways-solve-data-quality-issues/>)

Characteristics of good quality data:

1. High accuracy
2. High completeness
3. High reliability

(Source: <https://www.blazent.com/seven-characteristics-define-quality-data/>)

1. Overfitting is when you train your model on the training data so well that it would accommodate for all oddities in that data and show very little error. But when you apply that model to another dataset, it throws a huge error because it only knows the training data too well and cannot predict other data.

To prevent this:

We do regularization . This is a technique we use for regression where we penalize large coefficients to avoid overfitting by adding a hyperparameter of our own. By changing the value of the hyperparameter, we can regularize the model so that it performs better with any dataset.

1. Question : what is the difference between lasso and ridge regularization?

Answer: Ridge regularization happens when the hyperparameter is multiplied with the squares of the coefficients in the linear regression equation. So if the coefficients are very large during ridge regularization they are penalized more in the cost function. This makes regularization quicker.

Lasso regularization happens similar to ridge regularization but instead of using the squares of the coefficients, we use the absolute values of the coefficients. The advantage this provides is that we can make come coefficients to be 0 and this helps with feature selection in linear regression.

Question: what is the difference between regression trees and classification trees?

Answer: While both come under the umbrella of decision trees, regression trees deal with continuous variables like linear regression does. Classification trees deal with discrete categories/variables. So the rule of splitting used in the decision tree would be an equality in case of classification and the rule would be an inequality (<,>) in case of regression because in regression tree we try to see if different independent variables are in certain particular intervals or not.