Assessment 5: ADTA 5230.501 Module 5 Assignment, Chaudhary Yog 95

Q1.

Python library used for naive Bayes models is scikit-learn. As following Naïve Bayes Classifier:
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import train_test_split
from sklearn import metrics
and library installed in python environment pip: pip install scikit-learn.

Q2.

Each record is classified using the exact Bayes classifier. As following Bayes classifier in the process.

- Modeling the joint probabilities Distribution
- Bayes Theorem
- Classification Decision
- Handling Continuous Features
- Handling Discrete Features
- Computational aspects.

The Bayes classification requires knowing the true distributions for real-world applications due to the curse of the difficulty of estimating the true underlying distributions.

Q3.

The conditional probability of event A given event B is the probability that event A occurs given that event B has already occurred. As donated as P(A|B) and is calculating using the formula: P(A|B) = P (A ∩ B)/P(B)

Q4.

Yes, a Bayesian classifier can be used with numerical predictors. The most common approach is to discretize the numerical predictors into discrete intervals or bins. As following Naïve Bayes classifier with numerical predictors:

- Model Assumption.
- Parameters Estimation.
- Probabilities density function.
- Class Predictions.
- Implementation in Python.

Q5.

The Delayed Flights example involves predicting whether a flight will be delayed based on factors such as departure time, airline, and weather conditions. It uses a naive Bayes classifier to estimate the probability of delay given these factors. The example demonstrates how to train the classifier using historical flight data and evaluate its performance.

For Example, Flight Delays:
Day of Week: Code as 1 = Monday, 2= Tuesday, …... 7 = Sunday.
Schedule department: Broken down into 18 intervals between 6:am and 10: p.m.
Original: airport codes: DFW (Dallas Fort Worth International Airport), ORD (O'Hare International Airport), LAX (Los Angeles International Airport)
Destination: Airport Codes: DFW, ORD, LAX
Carrier: Flight airline code: AA (American Airlines), US(United), US (US Airway), QA (Quater Airlines), DL(Delta),

Q6.

The advantages of the naive Bayes classifier include simplicity, fast training and prediction, and good performance on large datasets. However, it assumes independence between features, which can be a limitation in some cases also it Can be sensitive to outliers.

The advantages of Naïve Bayes Classifiers are as follows:

- Simplicity and Efficiency
- Performance.
- Scalability.
- Handling of missing Values.
- Probabilities interpretations.
- Handles Continuous and Discrete date.

Shortcomings:

- Strong Features.
- Data Scarcity
- Complex Models.
- Correlated Features.
- Estimates Errors.

Q7.

Yes, a data analyst can use CART (Classification and Regression Trees) for both classification and regression tasks. CART is a versatile algorithm that constructs binary decision trees, which are a series of questions leading to a decision or prediction.

- For classification tasks.
- For regression tasks.

Both CART models are widely used due to their simplicity, and the fact that they can handle both numerical and categorical data. which can be advantageous in many real-world scenarios. However, decision trees can be prone to overfitting, especially if they are allowed to grow deep.

Q8.

Trees are based on a hierarchical structure composed of nodes, where each node represents a decision based on a feature or attribute.

- Hierarchical Structure.
- Splitting Criteria.
- Recursive Partitioning.
- Feature Selection.
- Pruning.
- Stopping Criteria.
- Tree Induction Algorithms.
-

Q9.

The construction of decision trees, including methods like CART (Classification and Regression Trees), is underpinned by two key ideas, which involve splitting the data into subsets based on features, and impurity measures, which quantify the homogeneity of a subset with respect to the target variable.

- Recursive Partitioning
- Greedy Approach

Both decision tree concept greedy optimization all trees to be built effectively for classification and regression tasks.

Q10.

The methodology behind decision trees was developed by Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen in 1984. Each methodology has different improvements such as different categorical variables. The field of decision trees is part of a large research area in machine learning and AI.

Q11.

The two types of nodes in a tree are the following:

- **internal nodes** (decision nodes)

- **leaf nodes** (terminal nodes).

Those nodes that represent the entire dataset before are referred to as the “root node” which technically decides the node.

Q12.

New records are classified, it are “dropped” down the tree. When it has dropped all the way down to a terminal node, we can assign its class simply by taking a “Vote” of all the training data that belongs to the terminal node, when the tree was grown. The class with the highest vote is assigned to the records. As following systematic process.

- Root node
- Continue Down the tree.
- Reach a leaf Node.
- Assign the class.

The process can be completed, which is one of the advantages of a decision tree to make a prediction once the tree has been constructed.

Q13.

The "Riding Mowers" example is a classic case study in classification within the realm of business analytics. While I don't have access to proprietary databases or specific case studies, the typical setup for a "Riding Mowers" scenario would involve a company that manufactures and sells riding lawnmowers and wishes to better understand its market segmentation and customer based on factors such as income, lot size, and presence of children. The classification task helps identify target customers for marketing campaigns. Hare a hypothesis summary

- Objective
- Content.
- Data
- Analysis.
- Outcome

The classification models can be instruments in the decision-making process by turning data. Which is the prediction of potential buyers' profiles for riding mowers.

Q14.

Gini impurity:

Gini impurity and entropy are two metrics that help us construct a decision tree, which is essentially a sequence of questions that lead us to a conclusion or classification. Both are used to quantify how much disorder or "impurity" exists at a given node in the tree – that is, a place where a decision is made to split the data.

In Decision Trees: When building a classification tree, the Gini impurity is used to evaluate splits. The tree algorithm will select the splits that result in the lowest weighted Gini impurity for the two child nodes. In other words, it chooses the split that reduces the uncertainty the most after the split.

Entropy:
Entropy is a measure from information theory that quantifies the amount of uncertainty or randomness in the data. It's used in the context of decision trees to measure the impurity of an arbitrary collection of examples. The higher the entropy, the more the information content.

In Decision Tree: The concept of entropy is used to calculate the Information Gain, which is the reduction in entropy before and after the dataset is split on an attribute. When constructing a decision tree, the splits are chosen to maximize the Information Gain, effectively reducing the randomness or impurity in the nodes following the split.

Both measures serve a similar purpose in the construction of a decision–making tree. The tree-building algorithm will compare the impurity of a node. Which attributes result in the largest decreases in impurity the algorithm iteratively constructs a tree to classify the training data.

Q15.

In a tree, the method for a numerical outcome involves partitioning the range of the outcome variable into intervals or bins and assigning each interval to a branch or leaf node. The predicted outcome for a new record falling within a particular interval is based on the majority outcome of training records within that interval. As following of how regression trees work:

- Starting at the root
- Choosing the Best split.
- Prediction from leaf nodes.
- Handing overfitting.

The regression tree provided an easily interpretable model for predicting numerical outcomes. Which is an ensemble of multiple decision trees.

Q16.

Decision trees are popular methods in machine learning and statistics due to their case of use. As following Advantages of Decision Trees.

- Interpretability.
- Feature Selection.
- Handling Non-linear Relationships
- Versatility.

Weakness of Decision Trees.

- Overfitting.

- Instability.
- Optimality.
- Problems with Unbalanced Data
- Decision Boundary Limitation.

The advantages of a tree include interpretability, Trees are scalable to large datasets, handling both numerical and categorical data, and the ability to capture non-linear relationships. However, weaknesses include instability (sensitivity to small changes in data), overfitting if not properly pruned, and difficulty in handling missing values. These methods combine multiple decisions tree to produce a more accurate model.