Technical Report

Breast Cancer Data Testing using Decision Tree, Random Forest, and Self-Training Method.

1. Machine Learning

Machine learning is a field of artificial intelligence (AI) that involves developing algorithms and statistical models that enable computers to learn and improve from experience without being explicitly programmed. In other words, the computer is trained on a large amount of data, and the algorithms learn to identify patterns and relationships within the data to make predictions or decisions.

Machine learning methods are algorithms or statistical models that enable machines to learn from data and improve their performance on a task without being explicitly programmed. These methods use techniques such as supervised learning, unsupervised learning, reinforcement learning, and deep learning to analyze data, recognize patterns, and make predictions or decisions.

Machine learning has a wide range of applications across many industries. Some common examples of machine learning in use include:

a. Image and speech recognition: Machine learning algorithms can be used to identify objects in images or transcribe speech into text.
b. Fraud detection: Machine learning can help identify fraudulent activity by analyzing patterns in transaction data.
c. Recommendation systems: Machine learning can be used to recommend products or services based on a user's past behavior or preferences.
d. Medical diagnosis: Machine learning algorithms can be trained on medical data to help diagnose diseases or identify potential health risks.
e. Financial forecasting: Machine learning can be used to make predictions about stock prices or other financial indicators.

Overall, machine learning is a powerful tool that can be used to automate decision-making processes and gain insights from data that would be difficult or impossible for humans to identify.

2. Common Machine Learning models

Machine learning models are mathematical representations that are trained on data to make predictions or decisions. They are algorithms that can learn from data without being explicitly programmed, and can generalize their learnings to make accurate predictions on new, unseen data.

Machine learning models can be used for a variety of tasks, including classification, regression, clustering, and more. They are typically trained on a large dataset that is split into a training set and a testing set. The training set is used to teach the model how to make predictions or decisions, while the testing set is used to evaluate the model's performance on new data.

There are many common models used in machine learning, each with its own strengths and weaknesses. Here are some of the most widely used models:

a. Linear regression: A model used to predict a numerical output based on one or more numerical input variables. It assumes that there is a linear relationship between the inputs and the output.
b. Logistic regression: A model used to predict a binary output (e.g. yes/no) based on one or more input variables. It uses a sigmoid function to map the input variables to the output probability.
c. Decision trees: A model that creates a tree-like structure to make decisions based on a series of conditions. It splits the data into smaller subsets based on the input variables until a decision can be made.
d. Random forests: An ensemble model that creates multiple decision trees and combines their predictions to make a final prediction. It reduces the risk of overfitting and can handle high-dimensional data.
e. Support vector machines: A model that finds the best hyperplane to separate two classes of data. It can be used for both classification and regression tasks.
f. Naive Bayes: A probabilistic model that predicts the probability of an outcome based on the input variables. It assumes that the input variables are independent of each other.
g. K-Nearest Neighbors (KNN): A model that predicts the output based on the k-nearest data points in the training set. It can be used for both classification and regression tasks.
h. Neural networks: A model that simulates the behavior of the human brain to make predictions. It consists of layers of interconnected neurons that can learn from data.
i. Gradient boosting: An ensemble model that combines multiple weak models into a strong model. It iteratively improves the model by fitting the residuals of the previous model.
j. Clustering algorithms such as K-means and hierarchical clustering: Models used to group similar data points together based on their characteristics. They can be used for exploratory data analysis or to identify patterns in the data.

Each of these models can be used for a variety of tasks, such as classification, regression, clustering, and more. Choosing the right model depends on the specific problem being solved and the characteristics of the data being used. It's common for data scientists to experiment with multiple models before selecting the one that works best for a given task.

3. Machine Learning models for Classification

There is no one-size-fits-all best machine learning model for classification tasks, as the performance of a model depends on various factors such as the size and quality of the data, the complexity of the problem, and the computational resources available. However, some models have been found to perform well for classification tasks in certain scenarios. Here are some commonly used models for classification tasks:

a. Logistic regression: A simple and efficient model that works well when the data is linearly separable and the classes are well-separated.

b. Support vector machines: A powerful and versatile model that works well with both linear and non-linear data and can handle high-dimensional data.
c. Decision trees and random forests: These models are good for handling categorical and continuous data, and can be easily interpreted by humans.
d. Naive Bayes: A fast and efficient model that works well when the independence assumption between the input variables holds, and is commonly used for text classification tasks.
e. Neural networks: A highly flexible and powerful model that can handle complex data and achieve state-of-the-art performance on various classification tasks.
f. K-Nearest Neighbors: A simple but effective model that works well with small datasets and when the decision boundary is not well defined.

It's important to note that the choice of model depends on the specific characteristics of the data and the problem being solved, and it's common for data scientists to experiment with multiple models before selecting the one that works best for a given task.

4. Public dataset for breast cancer

Breast Cancer Wisconsin (Diagnostic) Data Set this dataset contains information about breast cancer patients, and the task is to classify whether a tumor is malignant (cancerous) or benign (non-cancerous) based on various features.

The dataset includes 569 samples, with 212 malignant and 357 benign cases. Each sample has 30 features, including radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, fractal dimension, and more. These features are computed from digitized images of fine needle aspirate (FNA) of breast mass, which are then used to make a diagnosis of malignant or benign tumor.

This dataset is a public dataset that is freely available and can be used for various machine learning tasks related to breast cancer diagnosis. There are many other public datasets available for various machine learning tasks, such as the famous MNIST dataset for handwritten digit recognition, the Iris dataset for flower classification, and the Boston Housing dataset for regression tasks.

Public datasets are important for machine learning research because they allow researchers to compare their models against others in the field and evaluate their performance on a standard set of data. They also help promote transparency and reproducibility in research by allowing others to reproduce and verify the results.

5. Decision Tree

A decision tree is a type of supervised machine learning algorithm that is used for both classification and regression tasks. It is a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents a class label or a numerical value.

The algorithm starts with a root node that includes all the training data and recursively splits the data based on the feature that provides the most information gain, which is a measure of the reduction in entropy or impurity of the data. The process continues until all the data is classified or the stopping criteria are met.

In a classification task, the output of the decision tree is a class label for a given input feature vector, while in a regression task, the output is a numerical value. Decision

trees are easy to interpret and visualize, making them popular for use in industries where decision-making transparency is important, such as healthcare and finance.

Decision trees can also handle both categorical and numerical data and can be used with missing data. They can handle complex interactions between features and can capture nonlinear relationships between the input features and the output variable.

However, decision trees can be prone to overfitting, especially when the trees are deep and complex, and can lead to poor generalization on unseen data. To address this, ensemble methods such as random forests and boosting can be used to improve the performance and reduce the overfitting of decision trees.

In summary, decision trees are a powerful and flexible machine learning algorithm that can be used for classification and regression tasks. They are easy to interpret and visualize, can handle complex interactions, and can be used with both categorical and numerical data.

There are two important concepts related to decision trees.

a. Accuracy: Accuracy is a measure of how well a decision tree model can predict the correct class label for new, unseen samples. In the context of breast cancer classification, accuracy would measure how well the decision tree can predict whether a tumor is malignant or benign based on the input features. It is defined as the number of correct predictions divided by the total number of predictions, and is usually expressed as a percentage. For example, if a decision tree model predicts the diagnosis class correctly for 80 out of 100 samples, then the accuracy would be 80%.
b. Alpha: Alpha is a tuning parameter that controls the complexity of a decision tree model. In decision tree algorithms, the tree structure can become overly complex if the algorithm tries to fit the training data too closely, which can lead to overfitting. Overfitting occurs when the model is too complex and performs well on the training data, but poorly on new, unseen data. To prevent overfitting, the alpha parameter is used to limit the depth of the decision tree or the number of leaf nodes. A higher value of alpha will result in a simpler model with fewer splits and fewer nodes, while a lower value of alpha will result in a more complex model with more splits and more nodes. The optimal value of alpha can be determined through a process called cross-validation, where the dataset is divided into training and validation sets, and the performance of the model is evaluated for different values of alpha on the validation set. The value of alpha that results in the best performance on the validation set is then chosen as the optimal value.

6. Random Forest

Random Forest is an ensemble learning method based on decision trees. It is a supervised machine learning algorithm that is used for both classification and regression tasks. In Random Forest, multiple decision trees are created using different random subsets of the features and training samples. This helps to reduce the risk of overfitting and improve the generalization performance of the model. During the training phase, each tree in the forest is trained on a randomly selected subset of the training data, and at each split, a random subset of features is considered.

The final prediction of the Random Forest algorithm is obtained by averaging the predictions of all the individual trees. For classification tasks, the class with the highest frequency among the predictions of all the trees is chosen, while for regression tasks, the mean or median of the predictions of all the trees is calculated.

Random Forest is a popular algorithm due to its ability to handle high-dimensional data with complex interactions between the features. It can also handle missing values and outliers, and is relatively fast to train compared to other ensemble methods like Gradient Boosting. It is also robust to noise and can handle imbalanced datasets.

Random Forest is widely used in various applications, such as medical diagnosis, image classification, and financial analysis, where accurate predictions are critical. Its high accuracy, robustness, and ability to handle complex data make it a powerful and flexible machine learning algorithm.

7. Self Training

Self-training is a semi-supervised machine learning technique that is used when labeled data is limited, but a large amount of unlabeled data is available. The goal of self-training is to improve the performance of a supervised learning model by incorporating information from the unlabeled data during training.

In self-training, the supervised learning model is first trained on the small amount of labeled data. Then, the model is used to make predictions on the unlabeled data, and the predictions with the highest confidence scores are added to the labeled data. This expanded labeled dataset is then used to retrain the model, and the process is repeated until convergence.

The self-training approach assumes that the high-confidence predictions on the unlabeled data are correct, and can be treated as additional labeled data. This approach can be especially useful when obtaining labeled data is difficult or expensive, and the unlabeled data is readily available.

However, self-training can also be risky, as the model may become overconfident and produce incorrect predictions on the unlabeled data, which can lead to poor performance on the test data. Therefore, it is important to carefully select the threshold for selecting the high-confidence predictions and to evaluate the performance of the model on a separate validation or test set.

Self-training can be applied to various supervised learning models, such as decision trees, neural networks, and support vector machines. It is commonly used in applications such as natural language processing, speech recognition, and image classification, where large amounts of unlabeled data are available.

8. Conclusion

Machine learning offers a wide range of techniques for analyzing and interpreting complex datasets, including the classification of breast cancer based on input features. Some of the commonly used machine learning models for breast cancer classification include decision trees, random forests, and self-training.

Decision trees are a simple and interpretable model that can be used for breast cancer classification, and their performance can be improved by tuning the alpha

parameter to control the complexity of the tree and avoid overfitting. Random forests are an ensemble model that combines multiple decision trees to increase accuracy and reduce variance, while self-training is a semi-supervised learning technique that can leverage unlabeled data to improve model performance.

The breast cancer dataset is a common publicly available dataset that contains input features related to the diagnosis of breast tumors, including features such as tumor size, shape, margin, and texture. Machine learning techniques can be applied to this dataset to predict the diagnosis class of a breast tumor and inform clinical decision-making.

Overall, machine learning offers powerful tools for analyzing and interpreting medical data, including the diagnosis and treatment of breast cancer, and has the potential to improve patient outcomes and advance medical research.