Machine Learning Lab A3 - G1

2K19/CO/396 SURAJ KUMAR

Experiment – 6

Aim:

Heart Disease - Classifications (Logistic Regression, KNN, Naive Bayes Classifier, Random Forest Classifier/Decision Tree)

Softwares used:

Jupyter Notebook.

Introduction:

We have a data which classified if patients have heart disease or not according to features in it. We will try to use this data to create a model which tries predict if a patient has this disease or not. We will use logistic regression (classification) algorithm.

MATERIAL AND METHODS

KNN

KNN is a nonparametric, unsupervised machine learning algorithm. Nonparametric algorithm means there are no assumptions made on underlying data during training; the model does not summarize the training data. Nonparametric algorithms such as KNN [18] are beneficial because, generally, the practical data does not adhere to theoretical assumptions made (e.g., linearly separable, Gaussian mixtures). KNN tries to minimize the intra-cluster distance and maximize the inter-cluster distance.

RANDOM FOREST CLASSIFIER

A random forest classifier is a decision tree-based ensemble machine learning algorithm. It is a bootstrap aggregation algorithm. In bootstrap aggregation, also known as bagging, we create many random subsamples of the dataset with replacement and pass this data into randomly selected uncorrelated decision trees, also known as base learners [31]. These base learners work in parallel and produce their output. This step is known as bootstrap. The final output of the algorithm is the output determined by max voting on the outputs of all the individual base learners. This step is known as aggregation [32]. Figure 14.3 shows the process of the random forest classifier.

LOGISTIC REGRESSION (LR)

Logistic regression is a machine learning algorithm, named after the logistic function, used at its core. The logistic function is also referred to as the sigmoid function. Logistic function is a curve with “S” shape, similar to tanh(x), but exists only between 0 and 1, whereas tanh(x) exists between 1 and −1. Logistic function can take any real number and map it into a value between 0 and 1, but never exactly 0 or 1.

DECISION TREE

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

NAÏVE BAYES

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes’ Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.