Final Examination Report
On
Decision Tree, K-Nearest Neighbors

CSE-0408 Summer 2021

Rebecca Sultana
Dept. of CSE
State University of Bangladesh (SUB)
Dhaka, Bangladesh
rebecca.sultana.riya@gmail.com

***Abstract*—This is my academic projects.**

Assignment 1. Classification and Learning [Decision Tree]

I. Introduction

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

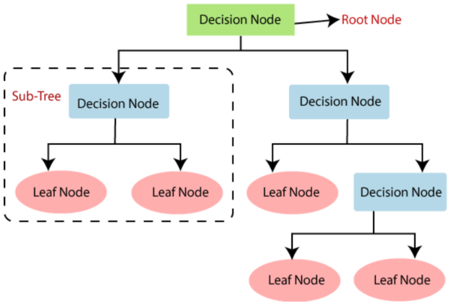
II. Types of Decision Trees

It can be of two types:
1.Categorical Variable Decision Tree: Decision Tree which has a categorical target variable then it called a Categorical variable decision tree.
2.Continuous Variable Decision Tree: Decision Tree has a continuous target variable then it is called Continuous Variable Decision Tree.

III. Classification Algorithm

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree.



The complete process can be better understood using the below algorithm:
Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).
Step-3: Divide the S into subsets that contains possible values for the best attributes.
Step-4: Generate the decision tree node, which contains the best attribute.
Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

IV. Advantages

1. Compared to other algorithms decision trees requires less effort for data preparation during pre-processing.
2. A decision tree does not require normalization and scaling of data.
3. Missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.

V. Drawbacks

1. A small change in the data can cause a large change in the structure of the decision tree causing instability.
2. For a Decision tree sometimes calculation can go far more complex compared to other algorithms.
3. Decision tree training is relatively expensive as the complexity and time has taken are more.

Assignment 2. K-Nearest Neighbor(K-NN) Algorithm

VI. Introduction

The k-nearest neighbors (K-NN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.

VII. Why do we need a K-NN Algorithm?

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset.

VIII. How does K-NN work?

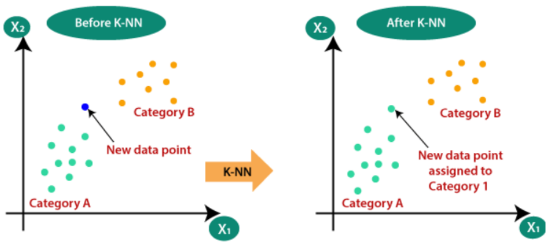
Step-1: Select the number K of the neighbors.
Step-2: Calculate the Euclidean distance of K number of neighbors.
Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
Step-4: Among these k neighbors, count the number of the data points in each category.
Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

IX. Advantages of K-NN Algorithm:

1. It is simple to implement.
2. It is robust to the noisy training data.
3. It can be more effective if the training data is large.

X. Drawbacks of K-NN Algorithm:

1. Always needs to determine the value of K which may be complex some time.
2. The computation cost is high because of calculating the distance between the data points for all the training samples.



XI. K-NN Code

Code figure in below:
importing libraries
import numpy as nm
import matplotlib.pyplot as mtp
import pandas as pd

importing datasets
data$_{s}et = pd.read_{c}sv('user_{d}ata.csv')$

Extracting Independent and dependent Variable
x= data$_{s}et.iloc[:,[2,3]].values$
$y = data_{s}et.iloc[:,4].values$

Splitting the dataset into training and test set.
from sklearn.model$_{s}electionimporttrain_{t}est_{s}plit$
$x_{t}rain, x_{t}est, y_{t}rain, y_{t}est = train_{t}est_{s}plit(x, y, test_{s}ize = 0.25, random_{s}tate = 0)$

feature Scaling
from sklearn.preprocessing import StandardScaler
st$_{x} = StandardScaler()$
$x_{t}rain = st_{x}.fit_{t}ransform(x_{t}rain)$
$x_{t}est = st_{x}.transform(x_{t}est)$

Fitting K-NN classifier to the training set
from sklearn.neighbors import KNeighborsClassifier
classifier= KNeighborsClassifier(n$_{n}eighbors = 5, metric =' minkowski', p = 2)$
$classifier.fit(x_{t}rain, y_{t}rain)$

Predicting the test set result
y$_{p}red = classifier.predict(x_{t}est)$