Assignment 2

EE769

Introduction to machine learning

Submitted by

Akash H Kapase

193100010

Under the Guidance of

Prof. Amit Sethi

Contents

1. Problem statement:

Implement any classifier using library functions to predict whether an employee will leave the company or not. Try out various data pre-processing techniques using pandas, scikit-learn or any other library you wish to use to get a higher ranking on the Kaggle leaderboard.

Given: Training data set with target value and testing data set without target value

To find: Prediction of output or target value (Employee will leave the company or not. Attrition) of testing data set.

2. Observations from data set

a. Observations about problem statement

It is multiple features one output problem with features as

Input features:

Age,,BusinessTravel,DailyRate,Department,DistanceFromHome,Education,EducationField,EmployeeCount,EmployeeNumber,EnvironmentSatisfaction,Gender,HourlyRate,JobInvolvement,JobLevel,JobRole,JobSatisfaction,MaritalStatus,MonthlyIncome,MonthlyRate,NumCompaniesWorked,OverTime,PercentSalaryHike,PerformanceRating,RelationshipSatisfaction,StockOptionLevel,TotalWorkingYears,TrainingTimesLastYear,WorkLifeBalance,YearsAtCompany,YearsInCurrentRole,YearsSinceLastPromotion,YearsWithCurrManager,ID

Output feature : attrition

As output attrition can have only two values (binary 0 or 1)

It is binary classification problem with multiple features

b. From training data set

-From training data set it is observed that

Most of output values are 0 and hence solution will bias towards 0. Very few entries have 1 value.

-Also not all features have numerical data

Some of the features like BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, OverTim .have categorical data e.g Geneder has values Male , female ..etc.

-Not all features have impact on output attrition such as Employee Number ID etc.

These features are excluded while defining feature set for x values of training and testing data set

- All values are given no any missing value for any feature

c. Distribution of data in each feature

From the distribution of data in each feature and its impact on output it is observed that decision tree can be used as classification algorithm for the given problem. Some features are important which have high impact in prediction of attrition. Decision tree can handle this.

3. Pre-processing of data

Input from kaggle is in csv format and to deal with such file for data manipulation and analysis pandas library is imported.It take input file in csv format and store output in same format.

Before applying classification algorithm given data should be pre-processed as it is not in the form require for classification algorithm

As discussed in observations not all features have numerical data required for general classification algorithm so data is in categorical form which is to be converted into numerical data.

Here two methods are tried for this

a. Using pandas and dummies

```
X = pd.get_dummies(train_data[features])
```

It converts all categorical form features into numerical type by assigning numerical values. It divides particular feature into number same as type of entries under that feature and assign binary value(0 or 1) under classified features.

E.g under feature marital status there are three types of entries married, single and divorced so after using this function divides features into three features named married, single and divorced and assign 0 or 1 value 1 value if it present and 0 value if it does not present.

b. Using SciKit learn

It uses label encoder to convert categorical data to numerical value by assigning numerical value to each datatype based on number of types of entries under particular feature.

E.g for above feature marital status it will keep same feature and assign 1 2 3 value to type of entries and put respective value instead of entry type.

1. married

2. single

3. divorced

Both methods gives good result . Though panda method is simple ,accuracy obtained by scikit learn algorithm is better

So it is used for conversion.

Age given in numerical form but it not useful in such format for classifier so it id grouped in 4 group as....

Age <30.   -1

Age >= 30 &< 40.   -2

Age >= 40 &< 50. -3

Age >= 40.   -4

4. Approach towards the solutions

Decision tree is used as explained in observations It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. It take care of features which have high impact on prediction as we require here.

Individual decision tree exhibit high variance and tend to over fit the model. Hence ensemble method is used. The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve robustness over a single estimator.

Averaging method of ensemble is simple and reduces the variance that obtained through individual decision tree. Here prediction of individual estimator is averaged.

Two types of classifier of this type is used are random forest classifier and Extra tree classifier. As both methods achieve a reduced variance by combining diverse trees, sometimes at the cost of a slight increase in bias. In practice the variance reduction is often significant hence yielding an overall better model.

In random forests, a random subset of candidate features is used. In Extra tree randomness goes one step further, thresholds are drawn at random for each candidate feature and the best of these randomly-generated thresholds is picked as the splitting rule which reduces variance bit more. Hence out of two Extra tree classifier reduces variance more than random forest classifier which can be seen from result shown below

Type of classifier	Accuracy
Individual decision tree classifier	0.78282
Random forest classifier	0.87373
Extra tree classifier	0.89393

5. Result and Learning

Result is shown in above table. Best accuracy is obtained by Extra tree classifier and is 0.89393

1. Selection of classifier for particular problem:

Based on given problem statement and data set which classifier will best fit the model with minimum variance.

2. Conversion of categorical data to numerical data using pandas and SciKit learn

How to deal with categorical data set for classification using pandas and SciKit learn libraries in Python along with various features of this library like label encoder etc..

3. Random Forest and Extra Tree algorithm

How random forest and extra tree algorithm fit best classification model into given data set with minimum variance.

Various commands related to this classifier algorithm

4. Data set in csv file format

How to import and save csv file using pandas library