San José State University
Computer Science Department
CS156, Introduction to Artificial Intelligence, Spring 2022

Homework #5

Objective:

This homework's objective is to implement a decision tree model to predict airline customer satisfaction (satisfied or not).

Details:

For this assignment you will be using the data obtained from US airline satisfaction survey. This dataset contains a variety of factors that might affect customer satisfaction with airline service. This dataset is publicly available from kaggle dataset repository. You can find more information about this dataset here:
https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction

Use input file homework5_input_data.csv for this assignment. The rows of this csv-formatted file are different records of customer interactions with airlines and the columns are various input variables for this dataset. The last column titled "satisfaction" is the output/response variable you will be predicting based on all the other input variables.

Implement a decision tree model to predict the "satisfied" or "neutral or dissatisfied" using all the other variables. This will be a binary classifier because there are only 2 classes available in the response variable. As for the independent variables, you will have to convert those columns that contain categorical data to one-hot encoding in order to use numeric input to train your model. If input values are not numeric, scikitlearn libraries will give an error if you try to use those values to train a classifier. You have to convert these categorical variables to numeric values. However, notice that not all independent variables contain categorical data. The columns that will need to be converted to one-hot encoding are: Gender, Customer Type, Type of Travel, Class.

Remember that in general there are two ways to deal with categorical data (for this assignment I am asking you to use the second approach, which is one-hot encoding):

1. You can map each category to a numeric value and replace each categorical value with its mapped numeric value.
2. You can create new binary variable column for each category available in the original column. In this scenario you will end up with as many new columns as the number of categories in the original column and the values will be 0 or 1 to indicate if the observation contains the particular category value in the original column. This approach is called one-hot encoding.

Here is a useful post to help you understand options for categorical variable encoding:
https://pbpython.com/categorical-encoding.html

You will have to split your input data into two subsets. The first will contain variables that need to be converted to one-hot encoding. The second one already contains numeric data.

To convert categorical columns to one-hot encoding you can use panda's get_dummies() function. Let's say your raw data is contained in the pandas dataframe called X1. You can use the following code to convert categorical data to numeric data:

```
X1_numeric = pd.get_dummies(X1, columns=X1.columns, prefix=X1.columns)
X1_numeric.head()
```

	Gender_Female	Gender_Male	Customer Type_Loyal Customer	Customer Type_disloyal Customer	Type of Travel_Business travel	Type of Travel_Personal Travel	Class_Business	Class_Eco
0	0	1	1	0	0	1	0	0
1	0	1	0	1	1	0	1	0
2	1	0	1	0	1	0	1	0
3	1	0	1	0	1	0	1	0
4	0	1	1	0	1	0	1	0

...

Once the conversion is done, use panda's concat() function to merge the two subsets back into a single input data set. Use that merged dataset to perform training-test set split.

You can use the code from my notebook examples as a reference to help you get started:
- DecisionTrees.Breast.ipynb

<u>Your submission should include the following:</u>

1. Load the dataset.
2. Convert categorical variables to numeric format as described above.
3. Break the data into the training and test datasets.
4. Train a decision tree model (DecisionTreeClassifier) to predict the class variable. Report (print out) 5-fold cross-validation accuracies (for all 5 folds as well as the mean accuracy).
5. Train a decision tree model on all the training data and report prediction accuracy on the test data.
6. Plot two confusion matrices for test set predictions (one non-normalized and one normalized). You can choose to use the same implementation of plotting a confusion matrices as I showed in my examples or include a different implementation. If you use code examples from the internet then make sure to site your sources in your notebook.

Submission:

Submit both the notebook file (.ipynb) and the pdf of the same notebook (.pdf) for your homework submission. To obtain a pdf of a notebook you can export or print it to a pdf format. Don't submit a notebook with just code and no output. I need to see that you were able to run your code and produce the output. Make sure submit by 11:59pm on the due date listed in Canvas. Submit your solution via Canvas.

If you have any questions, message me or the grader or both:
Yulia.Newton@sjsu.edu
gursimransingh@sjsu.edu

Grading:

I will return the grades as fast as we can grade this homework. Normally it should not take more than a few weeks.

A total of 10 points are possible for this homework assignment.