ISSS610
Applied Machine Learning

Assignment 2

[Question 1] Algorithm for Building Decision Tree

Day	Outlook	Temp.	Humidity	Wind	Play (Tennis)
D1	Sunny	85	85	Weak	No
D2	Sunny	80	90	Strong	No
D3	Overcast	83	86	Weak	Yes
D4	Rain	70	96	Weak	Yes
D5	Rain	68	80	Weak	Yes
D6	Rain	65	70	Strong	No
D7	Overcast	64	65	Strong	Yes
D8	Sunny	72	95	Weak	No
D9	Sunny	69	70	Weak	Yes
D10	Rain	75	80	Weak	Yes
D11	Sunny	75	70	Strong	Yes
D12	Overcast	72	90	Strong	Yes
D13	Overcast	81	75	Weak	Yes
D14	Rain	71	91	Strong	No

Given the toy dataset above, please build a decision tree to predict "**play**". Requirements:

1. Full derivation and explanation of how the branch decision is made at each node
2. Use Gain Ratio as branching criterion
3. Use the below algorithm to deal with continuous variables

```
for (each continuous feature A){
   Sort the examples according to their value for A;
   for (each ordered pair, X_i, X_{i+1}, in the sorted list)
      if (the category of X_i and X_{i+1} are different)
          find the midpoint of X_i and X_{i+1} denoted as c_i  to
          define threshold A < c_i
}
```

4. Evaluation of the built decision tree (use the training data)

[Question 2] New York Taxi Tip Prediction

In this question, you are to build a decision tree regression model to predict three things a taxi trip in New York City, based on other trip information, such as geo location, time of the trip, etc.

Three targets are:

- `tip_paid`: Whether the passenger paid a tip to the driver
- `tip_amount`: How much did the passenger to the driver
- `fare_amount`: How much was the taxi fare.

Two data sets are provided: `taxi-train.csv, taxi-test.csv`

Note: the target variables in `taxi-test.csv` are hidden. You need to run your prediction models and provide the best guesses to the target variables.

Your tasks:

1. Split the taxi-train.csv into train, valid and test dataset.
2. Perform exploratory data analysis on train and valid dataset.
3. Perform feature engineering on train and valid dataset.
    a. Take care of time zone transformation. The local time zone is US/East while the given time zone is in UTC.
    b. Use one-hot encoding for categorical variables with small number of distinct values.
    c. Bucketize long tail values and perform one-hot encoding for categorical variables with large number of distinct values.
    d. Come out with strategies to deal with location and time information.
    e. Try reasonable feature cross
    f. Any other things?
4. Build a decision tree classifier/regressor with default setting
    a. Evaluate the performance of the decision tree classifier: precision, recall, AUC, etc.
    b. Evaluate the performance of the decision tree regressor: RMSE
    c. Demonstrate and explain whether overfitting can be observed
5. The documentation of `sklearn.tree.DecisionTreeClassifier` can be found here.
   The documentation of `sklearn.tree.DecisionTreeRegressor` can be found here:

Please choose at least two parameters (e.g., criterion and max_depth) for hyperparameter tuning

    a. Explain why the tuning of chosen parameters can help improve the model performance
    b. Through your experiment, give your best combinations of the chosen parameters
    c. Compare the tuned model with the one using default setting in previous section

[Question 3] Ensemble learning on Census Income Prediction

With the same dataset and data pre-processing used in Question 2, please build ensemble models to answer the following questions.

1. Build `RandomForestClassifier/Regressor`, `AdaBoostClassifier/Regressor` and `GradientBoostingClassifier/Regressor` for income prediction with the objective metric you've chosen above. Define your parameter space and use either `GridSearchCV` or `RandomizedSearchCV` to find the best parameter combinations. Report the objective metric on the test data, and parameter combinations for the three models.
2. List 5 important features for each of the three models you've built for each prediction task.