Info 213

HW#2 & 3 (Practical part)

Due date: December 19, 2021

1- Load the **CarPrice_Prediction.csv** data set and perform the following steps:

a) Identifying the missed data for each feature.
b) Data cleaning if data is missed. In this case, you can remove the subjects with missed values.
c) Plot the correlation matrix among all variables.
d) Use 90% data as training and 10% test.
e) Data scaling, using "standardization" function.
f) Determine the dimensionality of data for test and train sets.
g) Using a random forest regression model (default settings), compute the prediction accuracy on training and test sets (R2, MAE, MSE). Plot the regression lines for predicted values vs. real values and histogram of predicted values vs. real values.
h) Is there any overfiring on this prediction model?
i) Change "n_estimators" from 25 to 1000 (step size 25) and analyse the response of prediction model for underfitting and overfitting issues.
j) What is your suggestion to select the optimal "estimators" for this prediction model? Why?

Note: "Price" is target in this data.

2- In this part we are going to identify the most important features using three different procedures: correlation method, linear regression coefficients, and random forest feature importance option. Load the **CarPrice_Prediction.csv** data set and perform the following steps:

a) Use 90% data as training and 10% test.
b) Data scaling, using "standardization" function.
c) Plot the correlation matrix among training features.
d) Sort the features based on correlation values between each feature and target.
e) Sort the features based on a linear regression algorithm.
f) Sort the features based on a random forest algorithm.
g) Show the sorted features based on three different methods in a table (Dataframe)
k) Select the top 3 best features using random **forest algorithm** and make a prediction task to estimate the outputs on test set.

Note: Sorting features should be done based on training set.

3- Compare the reliability of KNN (k=3), linear regression, Random forest (n_estimators =100) and decision tree algorithms on the **CarPrice_Prediction.csv** data set using 10-fold cross-validation strategy. Plot the prediction performance (R2, MAE and MSE) among different models. Which model have a better result on this dataset?

Info 213

HW#2&3 (Theory part)

Due date: December 19, 2021

The following table shows a data for a binary classification task. You are expected to answer to the following questions:

ID	Refund	Marital status	Tax (K$)	Class
0	Yes	Single	120	No
1	No	Married	80	Yes
2	No	Married	200	Yes
3	Yes	Single	200	Yes
4	Yes	Married	120	No
5	No	Single	120	Yes
6	Yes	Married	200	Yes
7	No	Married	80	Yes
8	Yes	Single	80	No
9	Yes	Single	120	Yes
10	No	Married	200	Yes

1- Calculate the entropy for "class".
2- Calculate the information Gain for all features (i.e., Refund, Marital status and Tax)
3- In a decision tree structure, which feature should be considered as root?