Homework Assignment 3

In this homework, you are expected to create a univariate feature selection schema using measures of impurity. We have learned two measures of impurity: entropy (***H(t, D)***), Gini Index (***Gini(t, D)***). These measures essentially show how random the distribution of a particular variable is or how well the values of a particular variable are separated. We also learned the Information Gain (IG) measure, which is a measure of the reduction in the overall entropy (or whichever measure of impurity is used).

In this homework, your task is to create a feature selection method called: ***IUFS*** (impurity-based univariate feature selection), which takes

- a dataset as (a pandas data frame, denoted as ***dataset***),
- the name of the target variable (denoted as ***target***),
- a measure of impurity (denoted as ***measure***) and
- the number of features to be selected (denoted as ***k***)

and returns the names (or indices) of the k columns to be selected based on the information gain.

Notes:

1. The input variable ***measure*** can only get 'entropy' or 'gini' as values. Otherwise, method terminates.
2. You can assume that your dataset consists of categorical variables.
3. You can find the definitions of measures of entropy and Gini index, as well as information gain in your textbook and slides. For entropy use base 2 for the logarithm.
4. You will be provided a starter code. You can use the car evaluation dataset provided in the starter code for your testing. Feel free to use other datasets with categorical variables (both target and descriptive features).

Bonus (+10 points):
Add a fifth parameter to IUFS method, called ***gain*** which can take 'IG' or 'GR' as values and will determine which selection measure to use between information gain (if gain=='IG') and gain ratio (if gain=='GR').

Question 1 (20 pts)
Suppose that you are given the task of creating a k-Nearest Neighbor (kNN) model to predict the class labels in the Iris dataset.
kNN classifiers work simply by searching the data space and finding the nearest neighbors of a given query instance. In other words, the kNN algorithm assumes that similar things exist in close proximity and they are near to each other. Therefore, if we can find the k-nearest neighbors for a given query instance, we can estimate a likelihood for its label. kNN models are usually referred to as lazy learners, because they simply memorize the provided dataset and do not necessarily learn a model. The prediction is purely based on the data and no actual model is learned. In addition to that, your boss wants you to use Euclidean distance to perform the nearest neighbor search and predictions. Your boss also wants you to use k=3.
Based on this information, what can you say about the inductive bias for this prediction task? Briefly describe at least one restriction and one preference bias for the kNN classifier that you will train for this task.

Question 2 (20 pts)
Indicate the data types and scales of attributes (i.e., ID, sepal length (cm), sepal width (cm), petal length (cm), petal width (cm), species) in Iris dataset. (Both the data types and the data scales.)

Question 3 (20 pts)
Machine learning is often referred to as an ill-posed problem. *What does this mean?*

Question 4 (40 pts)
An online movie streaming company has a business problem of growing **customer churn**[1]. Customers are cancelling their subscriptions to join a competitor and the company wants to decrease the customer churn.
Your task is

[1] Customer churn is a fancy way of saying the percentage of customers that stopped using your company's product or service during a certain time frame.

1. Propose a predictive data analytics solution which can be used to address this situation.
2. Describe the predictive model that will be built (*Note: not the model itself but the task*)
3. Describe how this model will be used by the business
4. How using this model can solve the original customer churn problem.

Relate the above mentioned steps to the major tasks in CRISP-DM process model.