Introduction to Machine Learning (NPFL054)

HW #2

The exercises relate to the Auto data set, which is part of the ISLR package. They are a modification of the exercises 122/9 and 171/11 published in [1].

1) Perform multiple linear regression

[5]

1. Consider mpg as the target value. Perform a multiple linear regression using all the attributes except name. Print the results. Provide an interpretation of each hypothesis parameter in the model.
2. Perform polynomial regression to predict mpg using acceleration. Plot the polynomial fits for the polynomial degrees 1 to 5 and report the associated R^2 values.

2) Develop a model to predict whether a given car gets high or low gas mileage.

1. Create a binary attribute, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. Create a single data set d containing both mpg01 and the other Auto attributes except mpg. Compute entropy of mpg01. [4]
2. Split the data d into a training set train and a test set test 80:20. [2]
3. **Make a trivial classifier** (without using the features) and evaluate it on the test set. Compute its accuracy. [2]
4. **Perform logistic regression** on train in order to predict mpg01 using all the features except name. [4]
   (a) Compute the training error rate. Produce a confusion matrix comparing the true test target values to the predicted test target values. Compute the test error rate.
   (b) Provide an interpretation of each hypothesis parameter in the model.
5. **Perform decision tree algorithm** on train to predict mpg01 using all the features except name. [5]
   (a) Create a plot of the tree. Compute the training error rate. Compute the test error rate.
   (b) Tune the cp parameter. Choose the *best* value of cp, and evaluate your model again. What is the *best* value of cp? Why? Explain it explicitly. Compute the accuracy of the model with your *best* cp.

6. **Perform Naive Bayes algorithm** on `train` to predict `mpg01` using all the features except `name`. Test it on `test`. Compute precision and recall. [3]

7. Randomly split `train` into eight folds to perform 8-fold cross validation. **Perform** ***k*****-NN** with several values of *k*, in order to predict `mpg01` using all the features except `name`. Plot 8-fold cross-validation error rate for different values of *k*. [5]

How to submit your assignment

- Write your R code to get answers for the exercises and name it `YourLastName_YourFirstName_hw2.R`

- Write your answers into the template file `hw2.odt` posted at the course webpage. Do not change the structure of this file. Save the file as `YourLastName_YourFirstName_hw2.odt` and then export it as `YourLastName_YourFirstName_hw2.pdf`.

- E-mail both files `YourLastName_YourFirstName_hw2.[R|pdf]` to the contact person specified in the homework assignment.

References

[1] James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated. 2014.