# Introduction to Machine Learning (NPFL054)

## HW #1

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

- 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 values of Adjusted  $R^2$ .

# 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.
- 2. Split the data d into a training set train and a test set test 80:20.
  - 3. Make a trivial classifier (without using the features) and evaluate it on the test set. Compute its accuracy.
  - 4. Perform logistic regression on train in order to predict mpg01 using all the features except name. Use a threshold of 0.5 to cut the predicted probabilities to make class predictions.
    - (a) Compute the training error rate.
    - (b) Produce a confusion matrix comparing the true test target values to the predicted test target values. Compute the test error rate, Sensitivity, and Specificity.
    - (c) Provide an interpretation of each hypothesis parameter in the model.
  - 5. In the previous exercise you used a threshold of 0.5. Re-run the experiment from the previous exercise with different threshold values, namely 0.1, 0.3, 0.6, 0.9.
    - (a) For each threshold value, produce a confusion matrix for comparing the true test target values to the predicted test target values and compute the Precision, Recall, and F1measure.

[5]

[4]

# [3]

[5]

[5]

[3]

(b) Provide an interpretation of the values of the given performance measures.

6. Perform decision tree algorithm on train to predict mpg01 using all the features except name.

- (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.
- 7. Final comparison NOT OBLIGATORY Compare the best models trained in the previous exercises 4., 5., and 6. Which one could be considered as the best?

#### How to submit your assignment

- Write your R code to get answers for the exercises and name it YourLastName\_YourFirstName\_hw1.R
- Write your answers into the template file hw1.odt posted at the course webpage. Do not change the structure of this file. Save the file as YourLastName\_YourFirstName\_hw1.odt and then export it as YourLastName\_YourFirstName\_hw1.pdf.
- E-mail both files YourLastName\_YourFirstName\_hw1. [R|odt|pdf] to the contact person specified in the homework assignment.

## References

 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.

[5]

[0]