**INTRODUCTION TO MACHINE LEARNING**

**(NPFL054)**

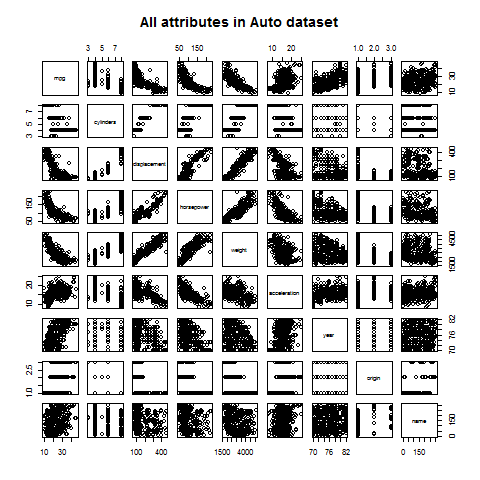
**A template for Homework #2**

**Name: Lukáš Jendele**

**School year: 2016/17**

* **Provide answers for the exercises 1. (a) - (c), 2.(c), 2.(d.1-2), 2.(e.1-2)**
* **For each exercise, your answer cannot exceed one sheet of paper.**

**1. (a) Scatterplot matrix, correlation matrix [5 pts]**

****



**1. (b) Multiple linear regression [10 pts]**

Explanation of the results:

**Intercept**: If all other attributes were zero and the car was American, then its mpg would be ~  ‑17.954 miles per gallon.

**Cylinders**: If we fix all the other attributes, then for every unit change in the number of cylinders, mpg decreases by 0.48. I.e. If we add a cylinder, we expect the car to drive almost half a mile less per gallon.

**Displacement:** Assuming all the other attributes fixed, for every increment of the engine displacement by a cubic inch, we expect the car to drive 0.023 miles further per gallon.

**Horsepower:** Provided that we fixed all other attributes, the car is expected to drive 0.018 mile less per gallon of fuel for every unit increment of horsepower.

**Weight:** If we fix all other attributes, then we expect the car to drive 0.007 mile per gallon of fuel for each extra pound of the car.

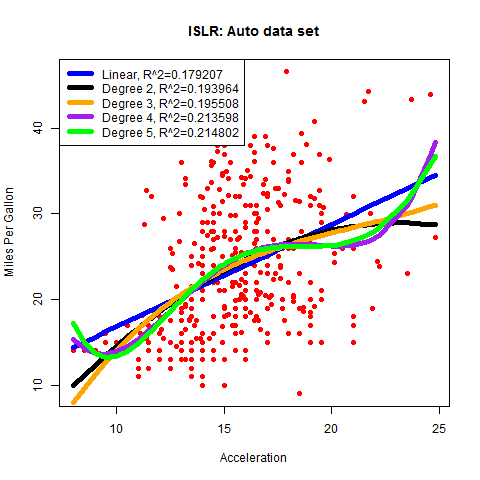
**Acceleration:** Assuming all the other attributes fixed, if the car can accelerate to 60 mph a second faster, then it is expected to drive 0.079 miles further per gallon of fuel.

**Year:** If all other attributes are fixed, for every year the car is newer, it is expected the car will drive 0.77 miles further per gallon.

**Origin:** Assuming other attributes fixed, European car are expected to drive 2.63 miles further than American cars and Japanese cars are expected to drive 2.85 miles further per gallon of fuel.



**1. (c) Polynomial regression [10 pts]**



**2. (c) Trivial classifier [10 pts]**

**Train *mpg01* frequency:**

|  |  |  |
| --- | --- | --- |
| 0 | 1 | Sum |
| 153 | 161 | 314 |

**Therefore:**

[1] "Trivial classifier always classifies as: 1"

[1] "Accuracy of trivial classifier on test dataset is: 44.871795%"

[1] "The entropy of mpg01 on train dataset is: 0.999532"

[1] "The entropy of mpg01 on test dataset is: 0.992399"

**Entropy of *mpg01*:**

[1] "The entropy of mpg01 on the whole dataset is: 1.000000"

[1] "The entropy of mpg01 on train dataset is: 0.999532"

[1] "The entropy of mpg01 on test dataset is: 0.992399"

**2. (d.1) Logistic regression: training and test error rate, confusion matrix [5 pts]**

**Training error rate:**

[1] "Train error rate of logistic regression is: 5.732484%"

[1] "Confusion matrix for test dataset:"

0 1 Sum

0 37 6 43

1 6 29 35

Sum 43 35 78

[1] "Test error rate of logistic regression is: 15.384615%"

**2. (d.2) Logistic regression: interpretation of the hypothesis parameters [10 pts]**

**Coefficients of the logistic regression:**

>coef(m.glm)

(Intercept) cylinders displacement horsepower weight

-21.486018200 -0.451765618 0.016974667 -0.033987572 -0.007071517

acceleration year origin.strEUR origin.strJPN

-0.025093073 0.576008838 1.867630758 0.643030729

>round(exp(coef(m.glm)), 3)

(Intercept) cylinders displacement horsepower weight

0.000 0.637 1.017 0.967 0.993

acceleration year origin.strEUR origin.strJPN

0.975 1.779 6.473 1.902

**Explanation:**

The interpretation is similar to linear regression, the only difference is that now, we have logits of probability. When explaining an attribute, I am assuming that all other attributes are fixed.

**Cylinders:** For every extra cylinder in the car, the probability that the car's mpg is below the mean increases by 0.362.

**Displacement:** For every increment by cubic inch of the displacement, the probability that the car's mpg is above the mean increases by 0.17

**Horsepower:** For every extra horse, the probability that the car's mpg is below the mean increases by 0.033.

**Weight:** For every extra pound, the probability that the car's mpg is below the mean increases by 0.007.

**Acceleration:** For every second in the acceleration from 0 to 60 mph, the probability that the car's mpg is below the mean increases by 0.025.

**Year:** For every increment of manufacture year, the probability that the car's mpg is above the mean increases by 0.779

**Origin + Intercept:** American car (with all other attributes zero) has probability nearly zero that its mpg is above mean, i.e. its mpg is below the mean. European car has 540% higher changes of being above the mean, whereas Japanese cars have only 90% higher odds of being above the mean than the American ones.

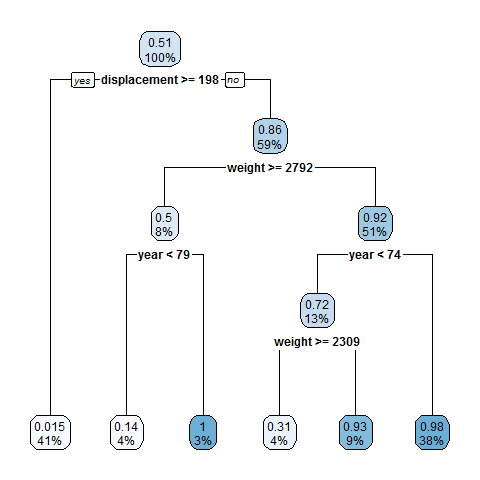
**2. (e.1) Decision trees: plot of the tree, training and test error rate [5 pts]**

**Training error rate (default cp):**

[1] "Train error rate of decision tree is: 3.821656%"

**Test error rate (default cp):**

[1] "Test error rate of decision tree is: 11.538462%"

**Plot of the decision tree:**

**2. (e.2) Decision trees: tuning the cp parameter [10 pts]**

**If cp=0 is set, and then printcp() is called, we get:**

Regression tree:

rpart(formula = mpg01 ~ cylinders + displacement + horsepower +

weight + acceleration + year + origin, data = train, cp = 0)

Variables actually used in tree construction:

[1] acceleration displacement horsepower weight year

Root node error: 78.449/314 = 0.24984

n= 314

CP nsplit rel error xerror xstd

1 0.6995184 0 1.00000 1.00627 0.003537

2 0.0506394 1 0.30048 0.38675 0.056764

3 0.0341368 3 0.19920 0.36968 0.059724

4 0.0029743 5 0.13093 0.24540 0.046803

5 0.0020337 6 0.12795 0.23886 0.043694

6 0.0019254 7 0.12592 0.24118 0.043664

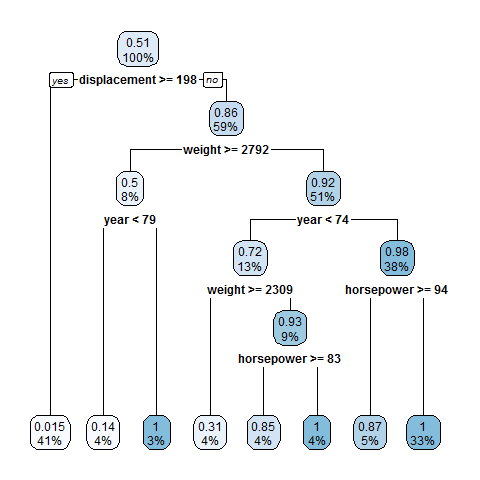
7 0.0013244 8 0.12400 0.24218 0.043781

8 0.0000000 9 0.12267 0.24101 0.043429

By looking at cross validation error *xerror*, we see that the tree predicts best at cp=0.0020337. Trees with higher cp would be underfitted and trees with lower cp would be overfitted (i.e. the tree wouldn't be able to generalize).

**The default cp is similar to the best one, therefore the accuracies are the same:**

[1] "Train error rate of decision tree is: 3.821656%"

[1] "Test error rate of decision tree is: 11.538462%"

If we look at the plot, we see that the tree has more leaves. However, when you realize that the extra leaves round to the same number as their parents, it is obvious why we are getting the same accuracy.