Bonus Assignment 1

MSPA PREDICT 422-DL-56 LEC

Darryl Buswell

1 Introduction

This document presents results of the first bonus assessment for the Masters of Science in Predictive Analytics course: PREDICT 422. This assessment required the student to perform cluster-wise regression for automobile data from the rpart package.

2 Bonus Assessment

2.1 Load the Dataset

As a first step, we load the 'car.test.frame' data from the rpart package and observe the structure of the data.

We can see that the dataset includes a number of numeric and factor type variables:

- Price: a numeric vector giving the list price in US dollars of a standard model
- Country: of origin, a factor with levels France, Germany, Japan , Japan/USA, Korea, Mexico, Sweden and USA
- Reliability: a numeric vector coded 1 to 5
- Mileage: fuel consumption miles per US gallon, as tested.
- Type: a factor with levels Compact Large Medium Small Sporty Van
- Weight: kerb weight in pounds
- Disp.: the engine capacity (displacement) in litres
- HP: the net horsepower of the vehicle.

'Price' will be our response variable, whilst 'Mileage', 'Weight', 'Disp', 'HP' will form our set of predictor variables.

2.2 Model Fit: Single Cluster

For this assessment, we leverage the FlexMix package in order to perform cluster-wise regression. FlexMix implements a general framework for finite mixtures of regression models, with parameter estimation being performed using the EM algorithm. We first fit a cluster-wise regression using a single cluster (the full set of observations). The model fit is shown below.

```
##
## Call:
## flexmix(formula = Price ~ Mileage + Weight + Disp. + HP,
##     data = car.test.frame, k = 1)
##
##        prior size post>0 ratio
## Comp.1     1   60     60     1
##
## 'log Lik.' -555.404 (df=6)
## AIC: 1122.808   BIC: 1135.374
```

We can also observe the parameter estimates over the cluster by using the 'parameters' function.

```
##                       Comp.1
## coef.(Intercept)  822.181804
```

```
## coef.Mileage     -165.400112
## coef.Weight         4.602914
## coef.Disp.        -40.567694
## coef.HP            70.908074
## sigma            2642.448365
```

We note the polarity of the majority of coefficients estimate seems reasonable, with an increase in fuel consumption translating to a decrease in price, or an increase in horse power resulting in an increase in price. The weight and displacement variables are more difficult to justify however.

2.3 Model Fit: Two Clusters

We can repeat this exercise by refitting the cluster-wise regression with two clusters. The model fit is shown below.

```
##
## Call:
## flexmix(formula = Price ~ Mileage + Weight + Disp. + HP,
##     data = car.test.frame, k = 2)
##
##        prior size post>0 ratio
## Comp.1   0.7   46     55 0.836
## Comp.2   0.3   14     55 0.255
##
## 'log Lik.' -538.7509 (df=13)
## AIC: 1103.502   BIC: 1130.728
```

We can see that the dataset has been broken into two samples of 46 and 14 observations in length. We note the improvement in AIC and BIC values reported for the above model versus the single cluster model previously fit.

Again, we can also observe the parameter estimates over both clusters by using the 'parameters' function.

```
##                       Comp.1
## coef.(Intercept) 6101.909096
## coef.Mileage     -193.960149
## coef.Weight         3.561308
## coef.Disp.        -12.628352
## coef.HP            13.304388
## sigma            1173.502924
```

```
##                       Comp.2
## coef.(Intercept) 2867.77610
## coef.Mileage     -463.78457
## coef.Weight        11.91847
## coef.Disp.        -91.74390
## coef.HP            27.43067
## sigma            2151.60300
```

We can see that although the polarity of coefficient is consistent over both clusters, the magnitude of those coefficients does have some variation.