ACSC/STAT 3740, Predictive Analytics

WINTER 2024

Toby Kenney

Homework Sheet 4

Due: Wednesday 20th March: 11:30

Note: This homework assignment is only valid for WINTER 2024. If you find this homework in a different term, please contact me to find the correct homework sheet.

Note: All data sets in this homework are simulated.

Standard Questions

1. The file `HW4Q1.txt` contains data on the relation between workers' rights and happiness. The data set contains the following variables:

Variable	Meaning
max.weekly.hours	The maximum number of hours an employee can be regularly required to work in
min.hourly.wage	The minimum hourly wage that can be paid to an employee.
paid.sick.leave	Whether employees are legally entitled to paid sick leave.
paid.parental.leave	Whether employees are legally entitled to paid parental leave.
min.holidays	The minimum number of holidays that employees are entitled to.
union.percent	The percentage of employees who belong to a labour union.
happiness	An index indicating the overall happiness of the population.

A data analyst uses the following code to fit a linear regression model to the data.

```
HW4Q1_linear<-lm(happiness~.,data=HW4Q1)
```

Use appropriate diagnostics to assess how appropriate the assumptions of the linear regression model are. What changes would you suggest making to the model to better model the data?

2. A data scientist at a company is analysing data about customer retention in the file `HW4Q2.txt`.

Variable	Meaning
previous.customer	Whether the customer has previously done business with the company.
age	The customer's age
sex	The customer's gender
spending	How much the customer spent.
service.needs	The number of hours of service the customer needed
survey.rating	The rating given by the customer.
six.month.return	Whether the customer returned within six months.

She has fitted a generalised linear model to predict whether the customer returns within 6 months, using the code in the file `HW4Q2_GLM.R`. Perform diagnostics to test which of the assumptions of this model are reasonable. What changes would you suggest making to the model to better model the data?

3. A scientist is reviewing data about the factors affecting health of captive animals, in the file `HW4Q3.txt`.

Variable	Meaning
social.type	The type of social group that the animal usually lives in in the wild.
diet	The animal's diet — herbivore, carnivore, etc.
born	Whether the animal was born in captivity.
enclosure.size	The size of the enclosure in which the animal is kept.
body.weight	The animal's body weight.
enclosure.shared	The number of other animals sharing the enclosure.
health.index	An overall assessment of the animal's health.

He has fitted a generalised additive model, a random forest model and a generalised linear model including a number of interaction terms and polynomial terms, to predict the health index, using the code in the file `HW4Q3_models.R`. Assess which of these models is better at predicting the data. [You may need to modify the code provided to do this.]

4. The file `HW4Q4.txt` contains data from about the probability that an individual will be injured during a sports match. The data set contains the following variables:

Variable	Meaning
age	The age of the participant.
sex	The sex of the participant.
contact	Whether the sport is a contact sport.
match.length	The length of the match.
fitness	An overall assessment of the fitness level of the individual.
strength	A measure of the strength of the individual.
previous.injury	Whether the individual has been injured in the previous six months.
injured	Whether the individual is injured.

A data analyst uses the following code to fit a decision tree to the data:

```
HW4Q4<-read.table("HW4Q4.txt")

library(rpart)

HW4Q4_dt<-rpart(formula=injured~.,
                data=HW4Q4,
                control=rpart.control(minbucket=1, # smallest size of node
                                      maxdepth=10, # largest depth of tree.
                                      cp=0.000001)) # complexity

### Find the minimum cross-validated error.
### Using 1-s.e. chooses a very simple tree.
HW4Q4_min<-min(HW4Q4_dt$cptable[,4])
HW4Q4_which_min<-min(which(HW4Q4_dt$cptable[,4]==HW4Q4_min))
HW4Q4_cp_min<-HW4Q4_dt$cptable[HW4Q4_which_min,1]
HW4Q4_dt_min<-prune(HW4Q4_dt,cp=HW4Q4_cp_min)
```

and uses the following code to select variables using stepwise regression with AIC:

```
HW4Q4_Null_model<-glm(injured~1,data=HW4Q4,family=binomial(link=logit))

HW4Q4_Full_model<-glm(injured~.,data=HW4Q4,family=binomial(link=logit))


library(MASS)
HW4Q4_Stepwise<-stepAIC(HW4Q4_Null_model,
                        direction="both",
                        scope=list(lower=HW4Q4_Null_model,
                                   upper=HW4Q4_Full_model))
```

The code is in the files `HW4Q4_Decision_tree.R` and `HW4Q4_Stepwise_AIC.R` respectively.

Based on the results of these analyses, should she try to adjust the models to better fit the data, and if so, how might she do so?