Report on BPL data analysis

Zhou Zhihao, 2014-2-18

data summar:

use recent 5 seasons, 1900 observations

HomeTeam = Home Team
AwayTeam = Away Team
FTHG = Full Time Home Team Goals
FTAG = Full Time Away Team Goals
HS = Home Team Shots
AS = Away Team Shots
HST = Home Team Shots on Target
AST = Away Team Shots on Target
HC = Home Team Corners
AC = Away Team Corners
HF = Home Team Fouls Committed
AF = Away Team Fouls Committed
HR = Home Team Red Cards
AR = Away Team Red Cards

Missing data:
HO = Home Team Offsides
AO = Away Team Offsides
HY = Home Team Yellow Cards
AY = Away Team Yellow Cards

datat manipulation

```
#read 5 years data
d1 = read.csv("PL2012-13.csv",header = T)
d2 = read.csv("PL2011-12.csv",header = T)
d3 = read.csv("PL2010-11.csv",header = T)
d4 = read.csv("PL2009-10.csv",header = T)
d5 = read.csv("PL2008-09.csv",header = T)
#get desired columns
label = c("HomeTeam","AwayTeam","FTHG","FTAG","HS","AS","HST","AST","HC","AC","HF","AF","HR","AR","Lambd
data = d1[,label]
data = rbind(data,d2[,label])
data = rbind(data,d3[,label])
data = rbind(data,d4[,label])
data = rbind(data,d5[,label])

#label 4 class
H_minus_lambda = data[,"FTHG"] - data[,"Lambda.H."]
A_minus_lambda = data[,"FTAG"] - data[,"Lambda.A."]
num_class =4
m1 = ifelse((H_minus_lambda + A_minus_lambda>=0),1,0)
m2 = ifelse((H_minus_lambda - A_minus_lambda>=0),1,0)
m3 = as.matrix(cbind(m1,m2))
```

```
Res_class  =apply(m3,1,function(x){ifelse(sum(x)==0,1,
                            ifelse(x[1]==1 & x[2]==0,2,
                            ifelse(x[1]==0 & x[2]==1,3,4)))})
#level 1,2,3,4

#combine predictor and response
d = cbind(data[,3:14],Res_class)
```

Can perform any statistical test on attributes under different class?

Apply Feed-forward Neural Networks with logistic output

Several Problems on ANN:

1. treat all match as same or need to classify them based on different team? I treat all game same here.
2. Model ends with wired result, predicts all case into class 4.
3. Is normalization necessary?
4. how to set proper parameter when training ANN model?
5. Suppose we get a well-trained ANN model, how to explain it? Can we get any useful information?

```
#construct target table for logistic output in nueral network
target = matrix(c(rep(0,nrow(d)*num_class)),nrow=nrow(d),ncol=num_class)
for( i in 1:nrow(d)){
  target[i,d[i,13]]=1
}

library(nnet)
#apply Feed-forward Neural Networks
#seems wired why fixed result?
#is normalization necessary?
ann = nnet(d[,1:12],target,size =10,decay=0,rang = 0,maxit =20000)
```

```
## # weights:  174
## initial  value 1900.000000
## iter  10 value 1618.978922
## iter  20 value 1498.650460
## iter  30 value 1418.229274
## final  value 1417.929474
## converged
```

```
#why all predict class 4?
p = max.col(ann$fit)
table(p,Res_class)
```

```
##    Res_class
## p     1   2   3   4
##   4 506 455 392 547
```

Try Multinomial Logistic Regression

1. Is it useful?
2. how to preperly explain coefficients?

```r
test<-multinom(Res_class~.,data=d[,3:13])
```

```
## # weights:  48 (33 variable)
## initial  value 2633.959286
## iter  10 value 2494.568568
## iter  20 value 2470.609045
## iter  30 value 2415.924443
## iter  40 value 2410.417412
## iter  40 value 2410.417412
## iter  40 value 2410.417412
## final  value 2410.417412
## converged
```

```r
summary(test)
```

```
## Call:
## multinom(formula = Res_class ~ ., data = d[, 3:13])
##
## Coefficients:
##   (Intercept)        HS        AS     HST      AST       HC      AC
## 2     -1.2259 -0.006296 -0.095523 0.07925  0.33470 -0.03574 -0.0336
## 3      0.1057 -0.078064  0.028858 0.09515 -0.03811 -0.10889  0.1040
## 4     -0.2738 -0.119186 -0.009147 0.33102  0.05416 -0.12502  0.0524
##          HF       AF      HR      AR
## 2  0.012617 -0.01792  0.2277 0.04652
## 3 -0.006772  0.01203 -1.2897 0.71301
## 4 -0.004152 -0.02076 -1.2023 0.78508
##
## Std. Errors:
##   (Intercept)      HS      AS     HST     AST      HC      AC      HF
## 2      0.4707 0.02561 0.03047 0.03610 0.04254 0.02420 0.02887 0.01981
## 3      0.4783 0.02735 0.02989 0.03850 0.04267 0.02609 0.02838 0.02028
## 4      0.4492 0.02541 0.02886 0.03566 0.04069 0.02376 0.02769 0.01917
##        AF     HR     AR
## 2 0.01834 0.2241 0.2582
## 3 0.01865 0.3150 0.2353
## 4 0.01753 0.3044 0.2216
##
## Residual Deviance: 4821
## AIC: 4887
```

```r
#how to performing model diagnostics?
z<-summary(test)$coefficients/summary(test)$standard.errors
#The multinom package does not include p-value calculation for the regression coefficients,
#so calculate p-values using Wald tests (here z-tests).
(p<-(1-pnorm(abs(z),0,1))*2)  # 2-tailed z test
```

```
##   (Intercept)        HS       AS     HST       AST        HC        AC
## 2    0.009199 8.058e-01 0.001721 0.02817 3.553e-15 1.398e-01 0.2444840
## 3    0.825060 4.316e-03 0.334395 0.01345 3.719e-01 2.998e-05 0.0002492
## 4    0.542190 2.727e-06 0.751283 0.00000 1.832e-01 1.424e-07 0.0584453
##       HF     AF        HR        AR
```

```
## 2 0.5241 0.3287 3.096e-01 0.8569932
## 3 0.7385 0.5190 4.230e-05 0.0024434
## 4 0.8285 0.2362 7.809e-05 0.0003968
```

```
#exp(coef(test)) #how to explain coefficients?
pre = max.col(fitted(test)) #predicted probabilities for each of our outcome levels
table(pre,Res_class) #training table
```

```
##    Res_class
## pre   1   2   3   4
##   1 207 121 118 100
##   2  95 170  62  92
##   3  59  36  95  54
##   4 145 128 117 301
```

Some General Problems

1. not fully undertand the meaning of respones? What information do they contain?
2. several key attributes are missing in data set(yellow cards, number of offsides)
3. data before year 2004-2005 didn't have attributes BdAvH, BdAvD and BdAvA. Can't estimated implied lambda.