homework 6

Zhang Zhiyuan 15220162202517

2019/5/16

this exercise uses the data relevant to smoking to predict whether a person is a smoker. and the detailed prediction process is a simple network, with the input being : education, gender, color and working area smoking ban.

```
data<-read.csv("D:/smoker.csv",header=T)
smoker<-data$smoker
ban<-data$smkban
hsdrop<-data$hsdrop
hsgrad<-data$hsgrad
colsome<-data$colsome
colgrad<-data$colgrad
black<-data$black
female<-data$female
```

to import data and create all the variables.

smoker: 1 if a smoker, o otherwise ban: 1 if working area smoking ban, 0 otherwise hsdrop: 1 if drop highschool, 0 otherwise hsgrad: 1 if graduate from highschool, 0 otherwise colsome: 1 if drop college, 0 otherwise colgrad: 1 if graduate from college, 0 otherwise balck: 1 if balck, 0 otherwise female: 1 if female, 0 otherwise

```
edu<-tanh(((1*hsdrop+2*hsgrad+4*colsome+8*colgrad)/13)^(1/2))
```

use the tanh function to generate the response for education input. different level of education are assigned with different weight, going with the underlying assumption that higher education decreases the probability of smoking.

this step is the first processment of input data, with other input going to the second processment unchanged.

```
edutr<-edu[1:8000]
bltr<-black[1:8000]
fetr<-female[1:8000]
bantr<-ban[1:8000]
sotr<-smoker[1:8000]

edute<-edu[8001:10000]
blte<-black[8001:10000]
fete<-female[8001:10000]
bante<-ban[8001:10000]
sote<-smoker[8001:10000]
```

divide the first 8000 point as train set, the remaining 2000 point being validation set.

```
pre<-glm(sotr~edutr+bltr+bantr+fetr,family=binomial)
```

use the logistic function to generate the output, ie, the probability of a smoker.

```
coef<-pre$coefficients
reg<-data.frame(1,edute,blte,bante,fete)
presmo<-crossprod(t(reg),coef)
msfe<-0
```

```
for (i in 1:2000){
  if(presmo[i]<0.5){
    t<-0
  }else{
    t<-1
  }
  msfe<-msfe+(sote[i]-t)^2
}
msfe<-msfe/2000
msfe
```

```
## [1] 0.268
```

use the validation set to get the mean squared error of the prediction of the network model.

so this model on average makes 73.2% correct predictions.

in general, this network model uses the input data in a different way from regular regressions, with different weitht controlled reasonably according to their importance, and the goodness of prediction is also quiet well.