B529: Homework 2

Nathan Byers

Friday, February 24, 2015

Question 1

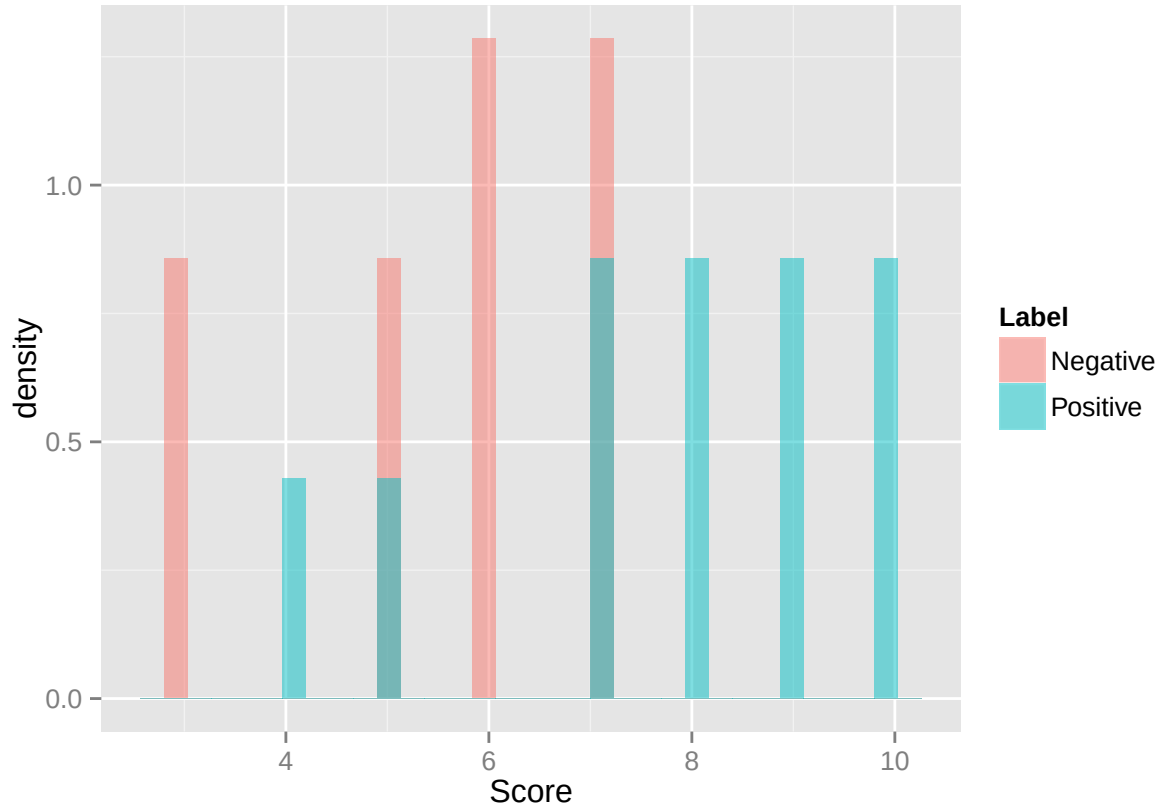
A scoring function (classifier) assigned scores for 10 positive data points and 10 negative ones as follows:

No.	Label	Score
1	Positive	10
2	Positive	10
3	Positive	9
4	Positive	9
5	Positive	8
6	Positive	8
7	Negative	7
8	Positive	7
9	Negative	7
10	Negative	7
11	Positive	7
12	Negative	6
13	Negative	6
14	Negative	6
15	Positive	5
16	Negative	5
17	Negative	5
18	Positive	4
19	Negative	3
20	Negative	3

Compute FPR and TPR for each score threshold in 3, 4, 5, 6, 7, 8, 9, 10, and plot the ROC curve of the scoring function (25 points).

Answer 1

Below is a histogram for the assigned scores.



Looking at the densities, it appears that Negative has lower scores than Positive. If we choose 3 as the cutoff (≤ 3 is Negative, > 3 is Positive), then we get the table below (the column headers indicate the target function and the first column indicates the outcome function).

	+1	-1
+1	2	0
-1	8	10

The false positive rate is $\frac{0}{0+10} = 0$ and the true positive rate is $\frac{2}{2+10} = 0.2$.

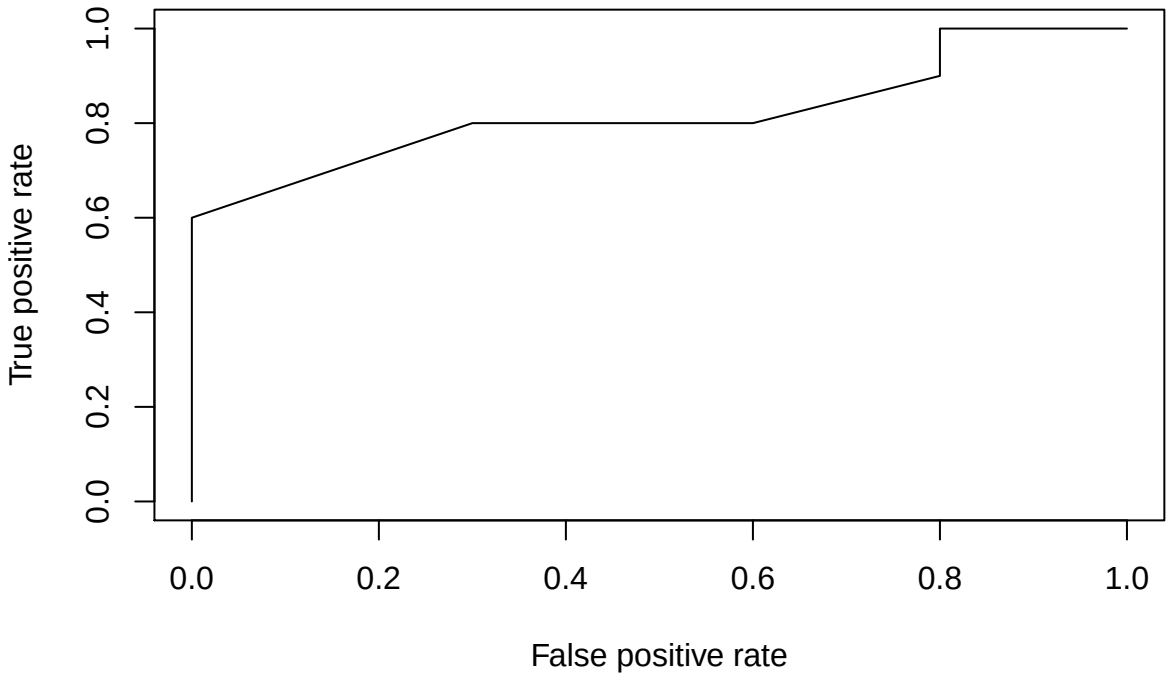
If 4 is the cutoff:

	+1	-1
+1	2	1
-1	8	9

$FPR = \frac{0}{0+10} = 0$ and $TPR = \frac{2}{2+10} = 0.2$.

For all of the cutoffs, we get the following table and ROC curve:

Score	TP	P	TPR	FP	N	FPR
≤ 10	10	10	1	10	10	1
≤ 9	10	10	1	8	10	0.8
≤ 8	10	10	1	6	10	0.6
≤ 7	10	10	1	2	10	0.2
≤ 6	7	10	0.7	2	10	0.2
≤ 5	4	10	0.4	2	10	0.2
≤ 4	2	10	0.2	1	10	0.1
≤ 3	2	10	0.2	0	10	0



Question 2

The abundance of a protein follows a normal distribution $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$. *Four measurements reported the following abundance values: 10, 13, 17, and 18. Use the maximum likelihood estimation method to estimate the value for parameter* θ. *(25 points)*

Answer 2

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2}}$$

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum(x_i-\theta)^2}$$

$$ln(L(\theta)) = -\frac{n}{2} log(2\pi) - \frac{1}{2}\sum(x_i-\theta)^2$$

$$\frac{\partial(ln(L\theta))}{\partial\theta} = 0 - \frac{1}{2}2\sum(x_i-\theta)(-1) = 0$$

$$\sum x_i - n\theta = 0 => \theta = \frac{1}{n}\sum x_i$$

The maximum likelihood estimate of theta is $\frac{10+13+17+18}{4} = 4.5$.

Question 3

In an experiment, the relationship between age, cholesterol level and signs of heart disease was studied. The data of five people was acquired (see the following table).

No.	Age	Cholesterol.level	Sign.of.heart.disease
1	50	70	0
2	70	120	1
3	60	130	0
4	60	150	1
5	70	180	1

Please use R to do logistic regression using the data set. Please give the commands, report the final weights, and plot the logistic function (25 points).

Answer 3

The table above was read into R as a `data.frame` named `hdisease`. Below I fit a model for each explanatory variable.

```
fit.chol <- glm(Sign.of.heart.disease ~ Cholesterol.level, hdisease, family=binomial("logit"))
fit.age <- glm(Sign.of.heart.disease ~ Age, hdisease, family=binomial("logit"))
```

Below are the summaries.

```
fit.chol
```

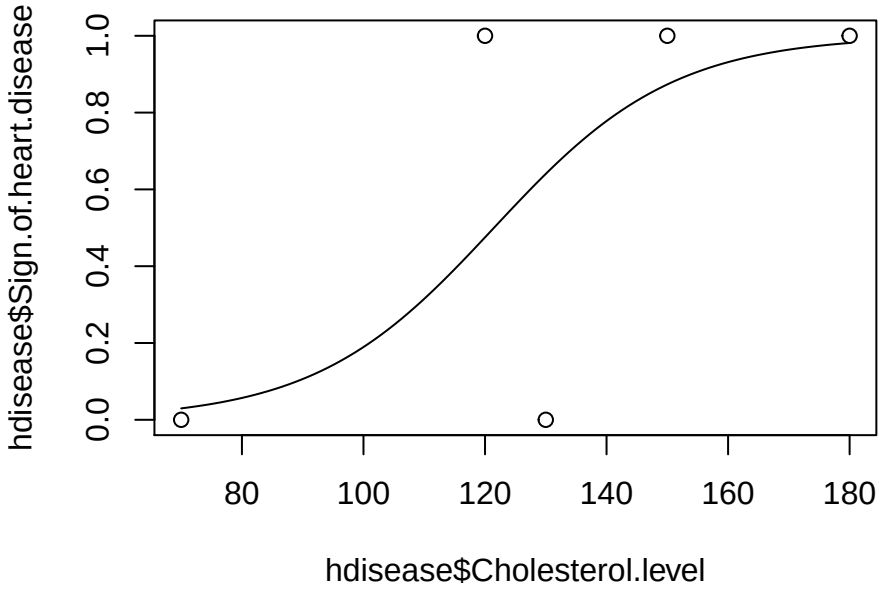
```
##
## Call:  glm(formula = Sign.of.heart.disease ~ Cholesterol.level, family = binomial("logit"),
##     data = hdisease)
##
## Coefficients:
##       (Intercept)  Cholesterol.level
##          -8.22451            0.06771
##
## Degrees of Freedom: 4 Total (i.e. Null);  3 Residual
## Null Deviance:       6.73
## Residual Deviance: 3.903     AIC: 7.903
```

```
fit.age
```

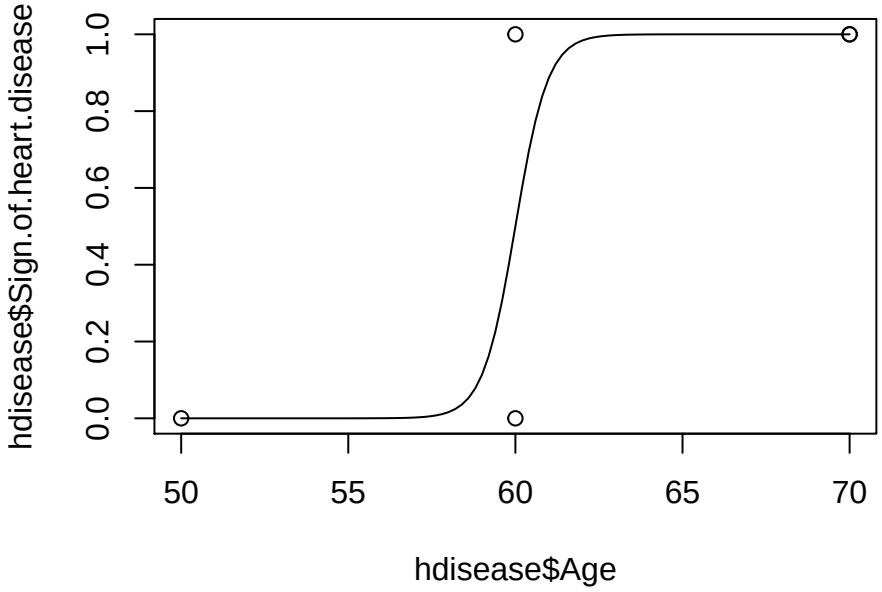
```
##
## Call:  glm(formula = Sign.of.heart.disease ~ Age, family = binomial("logit"),
##     data = hdisease)
##
## Coefficients:
## (Intercept)          Age
##    -123.396        2.057
##
## Degrees of Freedom: 4 Total (i.e. Null);  3 Residual
## Null Deviance:       6.73
## Residual Deviance: 2.773     AIC: 6.773
```

Below are the regression curves.

```
plot(hdisease$Cholesterol.level,hdisease$Sign.of.heart.disease)
curve(plogis(coef(fit.chol)[1] + coef(fit.chol)[2]*x),add=TRUE)
```




```
plot(hdisease$Age, hdisease$Sign.of.heart.disease)
curve(plogis(coef(fit.age)[1] + coef(fit.age)[2]*x),add=TRUE)
```



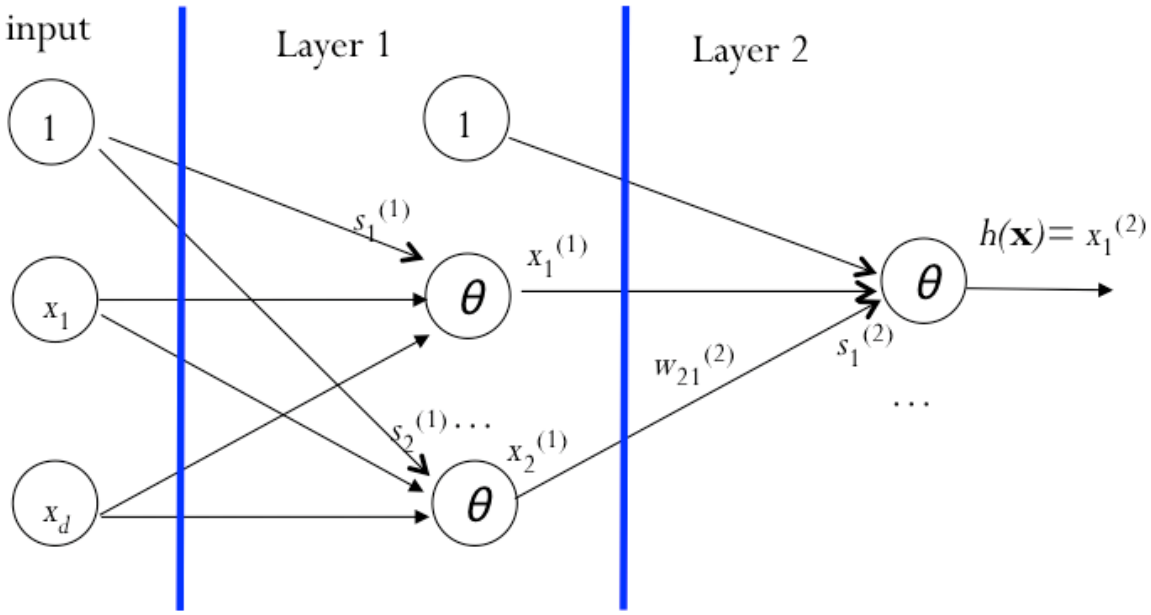
Age appears to be a better discriminator. The weights are

```
coef(fit.age)
```

```
## (Intercept)         Age
## -123.396411    2.056607
```

Quesion 4

The following neural network is trained using the back propagation algorithm.



The data points in the training data set are

x_1	x_2	y
-3	1	1
3	1	-1
3	-1	1
-3	-1	-1

Please use R to train the neural network. Please give the commands, report the final weights, and plot the neural network (25 points).

Answer 4

First I train the network using the `neuralnet()` function.

```
library(neuralnet)
network <- read.csv("C:\\Repositories\\MachineLearning\\Homework2\\network.csv",
```

```
                    stringsAsFactors = FALSE)
nn <- neuralnet(y ~ x1 + x2, data=network, hidden=2, err.fct="sse")
```

Below are the results.

```
nn$result.matrix
```

```
##                                    1
## error                   1.002374631060
## reached.threshold       0.009418480949
## steps                  68.000000000000
## Intercept.to.1layhid1 -4.570309812984
## x1.to.1layhid1        -1.352669806962
## x2.to.1layhid1        -2.031327587247
## Intercept.to.1layhid2  5.300246265830
## x1.to.1layhid2         1.907280318631
## x2.to.1layhid2        -3.341918768574
## Intercept.to.y         1.103692918681
## 1layhid.1.to.y        -1.275977793977
## 1layhid.2.to.y        -1.102899862282
```

Below is the plot of the network.

```
plot(nn)
```