Stats216v: Statistical Learning

Stanford University
Summer 2017
Gyu-Ho Lee (gyuhox@gmail.com (mailto:gyuhox@gmail.com))

4. Classification

4.1.R

Which of the following is the best example of a Qualitative Variable?

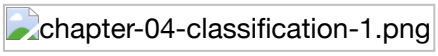
1. Height
2. Age
3. Speed
4. Color

Gyu-Ho's Answer: 4.

Colors are discrete values with no clear ordering. Height, Age, and Speed are all continuous.

4.1.R2

Judging from the plots on page 2 of the notes, which should be the better predictor of Default: Income or Balance?

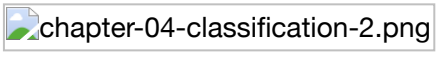


Gyu-Ho's Answer: Balance.

Default is clearly associated with higher balances. On the other hand, the rate of default seems fairly constant across income levels.

4.2.R1

Using the model on page 8 of the notes, what value of Balance will give a predicted Default rate of 50%? (within 3 units of accuracy)



Gyu-Ho's Answer: 1937.

$math.pow(math.e, -10.6513 + 0.0055 * 1937)/(1 + math.pow(math.e, -10.6513 + 0.0055 * 1937)) * 1($

We know that $logit(.5) = \beta_0 + \beta_1 * Balance$. Thus,
$Balance = (logit(.5) - \beta_0)/\beta_1 = (log(.5/(1 - .5)) + 10.6513)/.0055 = 1936.6.$

4.3.R1

Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficients $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

Estimate the probability that a student who studies for 40h and has an undergrad GPA of 3.5 gets an A in the class (within 0.01 accuracy):

Gyu-Ho's Answer: 0.37.

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2}} = \frac{e^{-6+0.05 X_1+X_2}}{1+e^{-6+0.05 X_1+X_2}} = \frac{e^{-6+0.05*40+3.5}}{1+e^{-6+0.05*40+3.5}} = 0.37754$$

4.3.R2

How many hours would that student need to study to have a 50% chance of getting an A in the class?:

Gyu-Ho's Answer: 50.

$$P((h, 3.5)) = \frac{e^{-6+0.05*h+3.5}}{1+e^{-6+0.05*h+3.5}} = 0.5$$

4.4.R

In which of the following problems is Case/Control Sampling LEAST likely to make a positive impact?

1. Predicting a shopper's gender based on the products they buy
2. Finding predictors for a certain type of cancer
3. Predicting if an email is Spam or Not Spam

Gyu-Ho's Answer: 1.

Case/Control sampling is most effective when the prior probabilities of the classes are very unequal. We expect this to be the case for the cancer and spam problems, but not the gender problem.

4.5.R1

Suppose that in Ad Clicks (a problem where you try to model if a user will click on a particular ad) it is well known that the majority of the time an ad is shown it will not be clicked. What is another way of saying that?

1. Ad Clicks have a low Prior Probability.
2. Ad Clicks have a high Prior Probability.
3. Ad Clicks have a low Density.
4. Ad Clicks have a high Density.

Gyu-Ho's Answer: 1.

Whether or not an ad gets clicked is a Qualitative Variable. Thus, it does not have a density. The Prior Probability of Ad Clicks is low because most ads are not clicked.

4.6.R1

Which of the following is NOT a linear function in x:

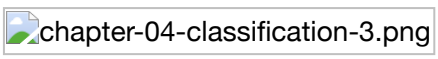
1. $f(x) = a + b^2 x$
2. The discriminant function from LDA
3. $\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)$
4. logit($P(y = 1|x)$) where $P(y = 1|x)$ is as in logistic regression
5. $P(y = 1|x)$ from logistic regression

Gyu-Ho's Answer: 5.

$P(y = 1|x)$ from logistic regression is not linear because it involves both an exponential function of x and a ratio. Notice that $f(x) = a + b^2 x$ is not a linear function of b, but is a linear function of x.

4.7.R1

Why does Total Error keep going down on the graph on page 34 of the notes, even though the False Negative Rate increases?



1. The False Negative Rate does not affect Total Error.
2. A higher False Negative Rate generally decreases Total Error.
3. Positive responses are so uncommon that the False Negatives make up only a small portion of the Total Error.
4. All of the above

Gyu-Ho's Answer: 3.

The Total Error is a weighted average of the False Positive Rate and False Negative Rate. The weights are determined by the Prior Probabilities of Positive and Negative Responses.

4.8.R1

Which of the following statements best explains the relationship between Quadratic Discriminant Analysis and naive Bayes with Gaussian distributions in each class?

1. Quadratic Discriminant Analysis is a more flexible class of models than naive Bayes
2. Quadratic Discriminant Analysis is a less flexible class of models than naive Bayes
3. Quadratic Discriminant Analysis is an equivalently flexible class of models to naive Bayes
4. For some problems Quadratic Discriminant Analysis is more flexible than naive Bayes, for others the opposite is true.

Gyu-Ho's Answer: 1.

With Gaussian distributions, naive Bayes is equivalent to Quadratic Discriminant Analysis with the additional requirement that each class covariance matrix Σ_k be diagonal. Thus, Quadratic Discriminant Analysis is more flexible.

4.Q.1

Which of the following tools would be well suited for predicting if a student will get an A in a class based on the student's height, and parents' income? Select all that apply:

1. Linear Discriminant Analysis
2. Linear Regression
3. Logistic Regression
4. Random Guess

Gyu-Ho's Answer: 1, 3.

1, 2, 3.

Whether or not a student gets an A is a categorical variables. Thus, we should use a classification technique like LDA or Logistic Regression. For binary classification, linear regression and LDA are almost equivalent.

4.R.R

In ch4.R, line 13 is "attach(Smarket)." If that line was omitted from the script, which of the following lines would cause an error?:

In [1]:
```
LoadLibraries = function() {
    library(MASS)
    install.packages("ISLR")
    library(ISLR)
    print("Libraries have been loaded!")
}

LoadLibraries()
```

```
Updating HTML index of packages in '.Library'
Making 'packages.html' ... done
```

```
[1] "Libraries have been loaded!"
```

In [2]:
```
names(Smarket)
```

'Year' 'Lag1' 'Lag2' 'Lag3' 'Lag4' 'Lag5' 'Volume' 'Today' 'Direction'

In [3]:
```
mean(glm.pred==Direction)
```

```
Error in mean(glm.pred == Direction): object 'glm.pred' not found
Traceback:

1. mean(glm.pred == Direction)
```

In [ ]:
```
glm.fit = glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,data=Smarket,fam
ily=binomial, subset=train)
```