Data Analysis Using R: Chapter12

罗智超 (ROKIA.ORG)

1 通过本章你将学会

- Logistic 回归
- 朴素贝叶斯分类
- 练习使用朴素贝叶斯分类

2 身高体重数据性别分类

```
#Data:heights_weights_genders.csv


# Start visualizing data using the ggplot2 package.
library('ggplot2')

# Load the data from scratch for purity.

heights.weights <- read.csv("data/heights_weights_genders.csv", header = TRUE, sep = ',
# Experiment with histograms.

ggplot(heights.weights, aes(x = Height)) +
  geom_histogram(binwidth = 0.01)

# Experiment with kernel density estimates.
```

```r
ggplot(heights.weights, aes(x = Height)) +
  geom_density()


# Separate out heights and weights based on gender.
ggplot(heights.weights, aes(x = Height, fill = Gender)) +
  geom_density()


ggplot(heights.weights, aes(x = Weight, fill = Gender)) +
  geom_density()


# Produce two facets in a single plot to make it easier to see the hidden structure.
ggplot(heights.weights, aes(x = Weight, fill = Gender)) +
  geom_density() +
  facet_grid(Gender ~ .)


# Experiment with random numbers from the normal distribution.
m <- 0
s <- 1
ggplot(data.frame(X = rnorm(100000, m, s)), aes(x = X)) +
  geom_density()


# Compare the normal distribution with the Cauchy distribution.
set.seed(1)
normal.values <- rnorm(250, 0, 1)
cauchy.values <- rcauchy(250, 0, 1)
range(normal.values)
range(cauchy.values)
```

```
ggplot(data.frame(X = normal.values), aes(x = X)) +
  geom_density()
ggplot(data.frame(X = cauchy.values), aes(x = X)) +
  geom_density()


# Experiment with random numbers from the gamma distribution.
gamma.values <- rgamma(100000, 1, 0.001)
ggplot(data.frame(X = gamma.values), aes(x = X)) +
  geom_density()


# Generate scatterplots of the heights and weights to see their relationship.
ggplot(heights.weights, aes(x = Height, y = Weight)) +
  geom_point()


# Add a smooth shape that relates the two explicitly.
ggplot(heights.weights, aes(x = Height, y = Weight)) +
  geom_point() +
  geom_smooth()



# See how the smooth shape gets better with more data.
ggplot(heights.weights[1:20, ], aes(x = Height, y = Weight)) +
  geom_point() +
  geom_smooth()
ggplot(heights.weights[1:200, ], aes(x = Height, y = Weight)) +
  geom_point() +
  geom_smooth()
ggplot(heights.weights[1:2000, ], aes(x = Height, y = Weight)) +
```

```
  geom_point() +
  geom_smooth()


# Visualize how gender depends on height and weight.
ggplot(heights.weights, aes(x = Height, y = Weight)) +
  geom_point(aes(color = Gender, alpha = 0.25)) +
  scale_alpha(guide = "none") +
  scale_color_manual(values = c("Male" = "black", "Female" = "gray")) +
  theme_bw()

# An alternative using bright colors.
ggplot(heights.weights, aes(x = Height, y = Weight, color = Gender)) +
  geom_point()



heights.weights <- transform(heights.weights,
                             Male = ifelse(Gender == 'Male', 1, 0))

logit.model <- glm(Male ~ Weight + Height,
                   data = heights.weights,
                   family = binomial(link = 'logit'))

ggplot(heights.weights, aes(x = Height, y = Weight)) +
  geom_point(aes(color = Gender, alpha = 0.25)) +
  scale_alpha(guide = "none") +
  scale_color_manual(values = c("Male" = "black", "Female" = "gray")) +
  theme_bw() +
  stat_abline(intercept = -coef(logit.model)[1] / coef(logit.model)[2],
              slope = - coef(logit.model)[3] / coef(logit.model)[2],
              geom = 'abline',
              color = 'black')
```

3 朴素贝叶斯分类

```
#e1071::naiveBayes()
```

4 练习

```
#DATA:survey2014_student.xls
```