Machine Learning in R

Chenshu Liu

April 2022

Packages

```
# for data splitting into training and testing
library(caTools)
```

Linear Regression

Linear regression is a way to find the best fit linear expression that can show the observed trend(s). The parameters of the best fit line $y = mx + c$ can be calculated by:

$$m = \frac{(n \times \Sigma(x \times y)) - (\Sigma(x) \times \Sigma(y))}{(n \times \Sigma(x^2)) - (\Sigma(x)^2)}$$

$$c = \frac{(\Sigma(y) \times \Sigma(x^2)) - (\Sigma(x) \times \Sigma(x \times y))}{(n \times \Sigma(x^2)) - (\Sigma(x)^2)}$$

Data

```
sales <- read.csv("/Users/chenshu/Documents/Programming/R/Machine Learning in R/datasets/revenue.csv")

# split into training and testing data
set.seed(2)
# SplitRatio means the percentage of data for training
split <- caTools::sample.split(sales$Profit, SplitRatio = 0.7)
train <- sales[split, ]
test <- sales[!split, ]
```

Modeling

```
Model <- lm(Profit ~., data = train)
summary(Model)
```

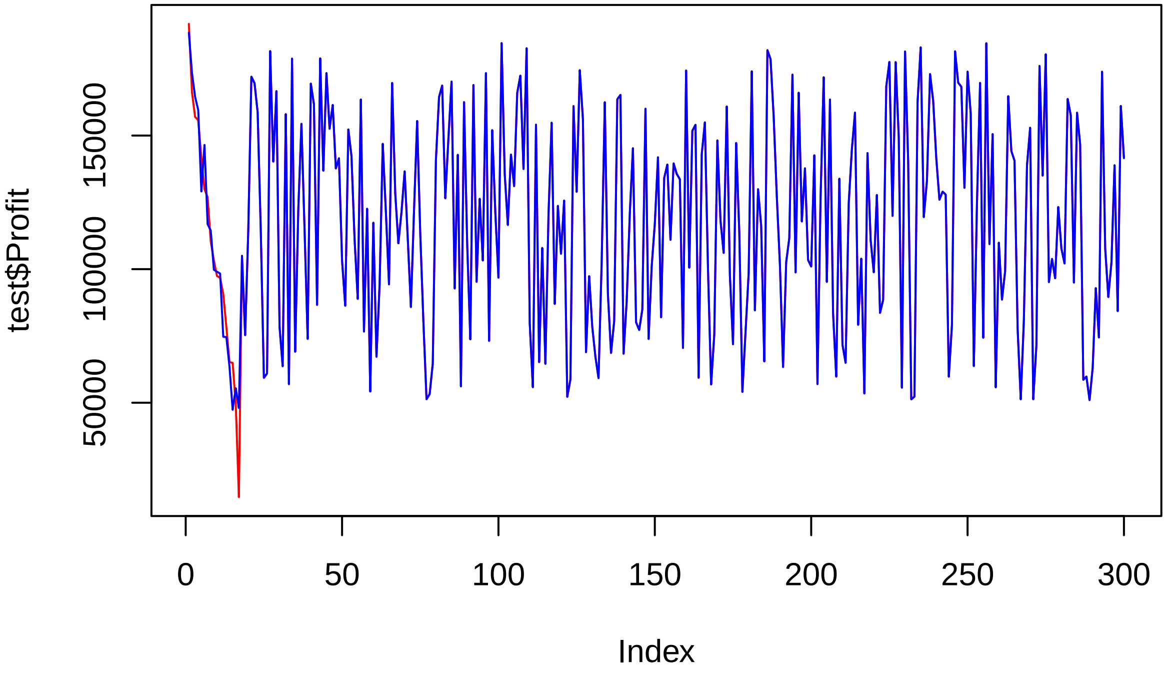
```
##
## Call:
## lm(formula = Profit ~ ., data = train)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16134.9    -32.2    -17.7     -6.5  10133.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.395e+04  1.145e+03  47.126  < 2e-16 ***
## Paid         8.155e-01  6.982e-03 116.802  < 2e-16 ***
## Organic     -6.007e-02  9.547e-03  -6.292 5.53e-10 ***
## Social       2.488e-02  3.247e-03   7.661 6.21e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1480 on 696 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9986
## F-statistic: 1.693e+05 on 3 and 696 DF,  p-value: < 2.2e-16
```

Predict

```
pred <- predict(Model, test)

# comparing predicted vs. actual values
plot(test$Profit, type = 'l', lty = 1.8, col = "red")
lines(pred, type = 'l', lty = 1.8, col = "blue")
```



```
# determining prediction accuracy
rmse <- sqrt(mean(pred - test$Profit)^2)
rmse
```

```
## [1] 61.56887
```

Logistic Regression

Logistic regression is a **classification algorithm**, not a linear prediction algorithm

Data