机器学习练习

逍遥子晴

2017 年 8 月 9 日

概述：

我所理解的机器学习简单来说，就是用现有数据和统计方法把数据转化为智能的行动，通俗来讲就是用计算机解决实际问题的行为和过程。所以学好机器学习就要从数据、算法、计算机语言三个方面去着手。

1、数据方面，包括数据获取、数据理解、数据处理、数据探索等步骤。

- 数据获取：很多时候我们的数据是现成的，有时也需要用各种方法去收集，对于网络数据一般采用爬虫的方法。总之，这一步不难实现。
- 数据理解：数据理解顾名思义就是理解数据的含义，包括变量的个数、每个变量的数据结构、数据大小等，另外还需要结合业务问题去看待。
- 数据处理：大多数时候我们拿到的数据都不是直接就能用的，它里面有很多的缺失值、重复值、错误值需要我们去做处理。有时我们也需要把原始数据转化一下再拿来使用。
- 数据探索：这一步会用到简单的描述性统计方法，在进行数据建模之前，简单的寻找各变量的特征及变量间的相关关系是很有必要的。

2、算法，下面列出了几大常用算法，每个算法还包含很多的小算法，目前这些算法在 R 语言中都能找到相应的包去实现。虽然短时间内无法对这些算法一一掌握，但是要了解每个算法的用途和实现逻辑。

- 有监督学习：决策树、朴素贝叶斯分类器、最小二乘法、逻辑回归、支持向量机、集成学习
- 无监督学习：聚类分析、主成分分析、SVD 矩阵分解、独立成分分析

3、计算机语言，暂以 R 语言为主。

实例：

机器学习目的：R 语言解决一个具有确定性类别的分类问题。

一、数据问题

1、数据获取：R 中自带 iris 鸢尾花数据集。

2、数据理解：

```
> data(iris)
> str(iris)
```

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
```

```
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
> library(caret)
```

```
Loading required package: lattice
```

```
Loading required package: ggplot2
```

```
> library(ellipse)
> iris.data <- iris
> str(iris)
```

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

由输出结果及所查资料可知该数据集包含了 150 个观测值，5 个属性：

- Sepal.Length（花萼长度），单位是 cm；
- Sepal.Width（花萼宽度），单位是 cm；
- Petal.Length（花瓣长度），单位是 cm；
- Petal.Width（花瓣宽度），单位是 cm；
- species（种类）：Iris Setosa（山鸢尾）、Iris Versicolour（杂色鸢尾）、Iris Virginica（维吉尼亚鸢尾）。
- 初步理解：数据集中的 5 个属性有四个是数值型变量，一个分类变量。由此可以想到鸢尾花的分类是否与四个数值型变量中的一个或几个有关系？如果存在一定的关系我们可否在已知鸢尾花花萼和花瓣长宽度的条件下对它进行分类？

3、数据探索

- summary 函数

```
> summary(iris.data)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
```

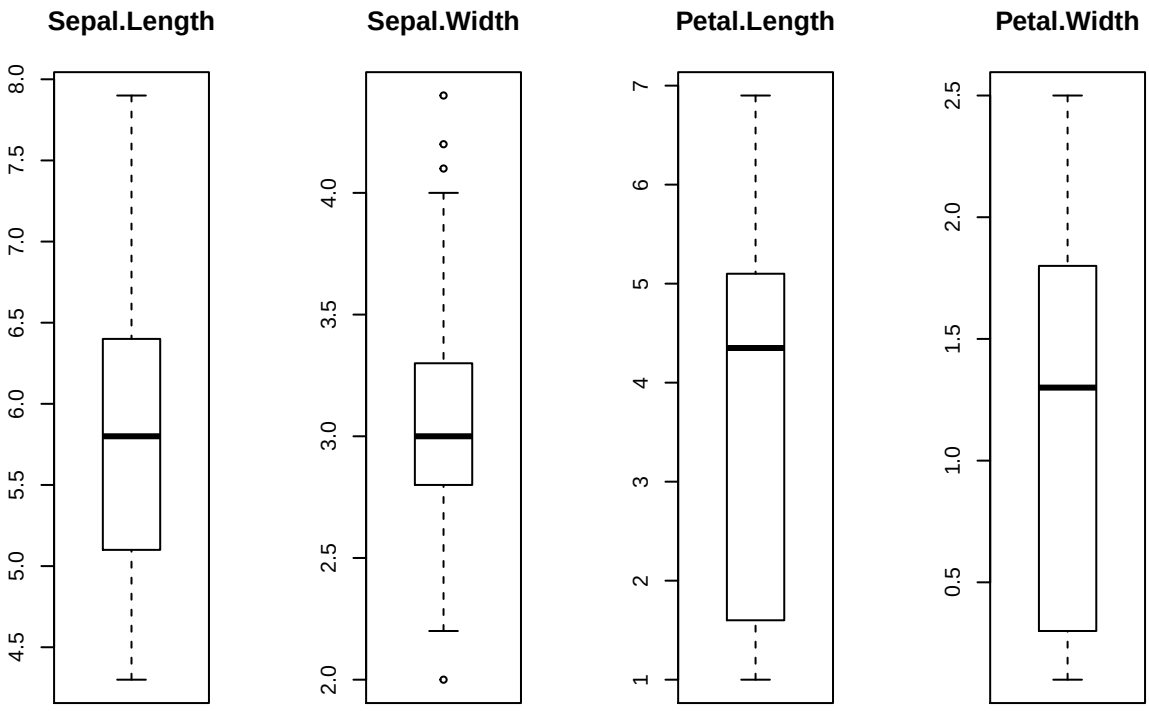
```
 virginica :50
```

使用 summary 函数获取数据集摘要，对于数值型变量返回 5 个数字化特征：最小值，第一分位数，中位数，均值，第三分位数和最大值，对于因子型变量，返回每个类别的频数。

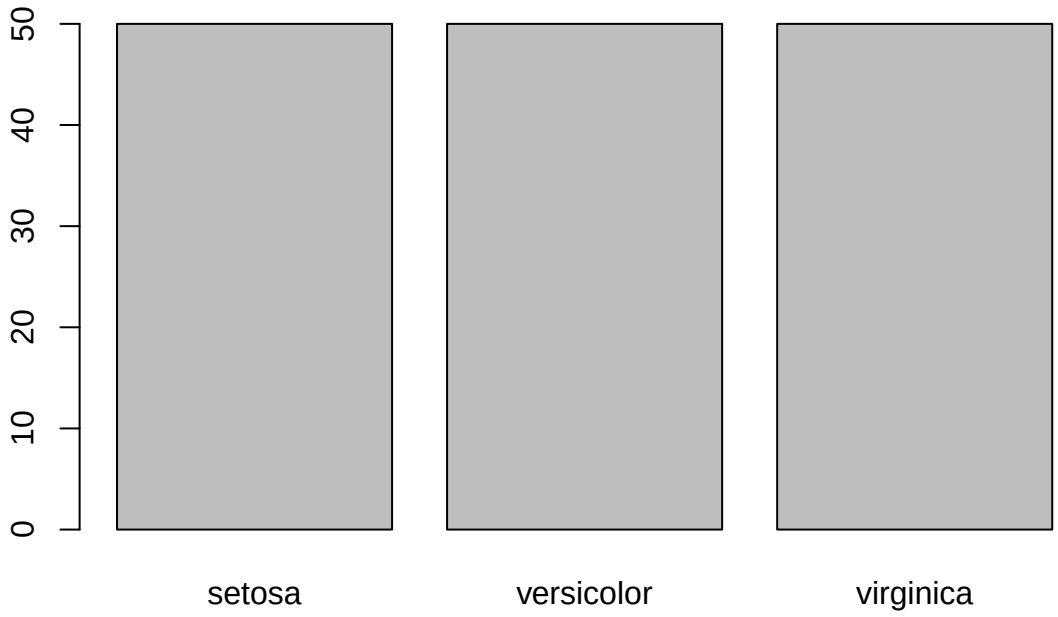
- 单变量可视化

单变量可视化，即针对数据集中的单个变量画图，通过图形观察和理解变量的分布情况。以下通过箱线图和直方图分别对四个数值变量和一个因子变量做出初步分析。

```
> input.val <- iris.data[, 1:4]
> par(mfrow = c(1, 4))
> for (i in 1:4) {
+     boxplot(input.val[, i], main = names(iris.data)[i])
+ }
```

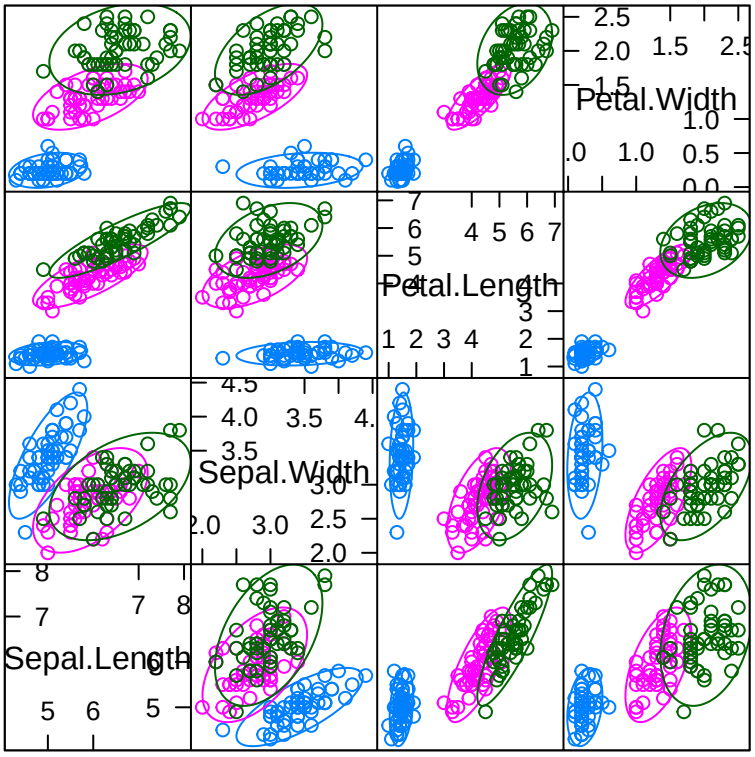



```
> par(mfrow = c(1, 1))
> output.val <- iris.data[, 5]
> plot(output.val)
```



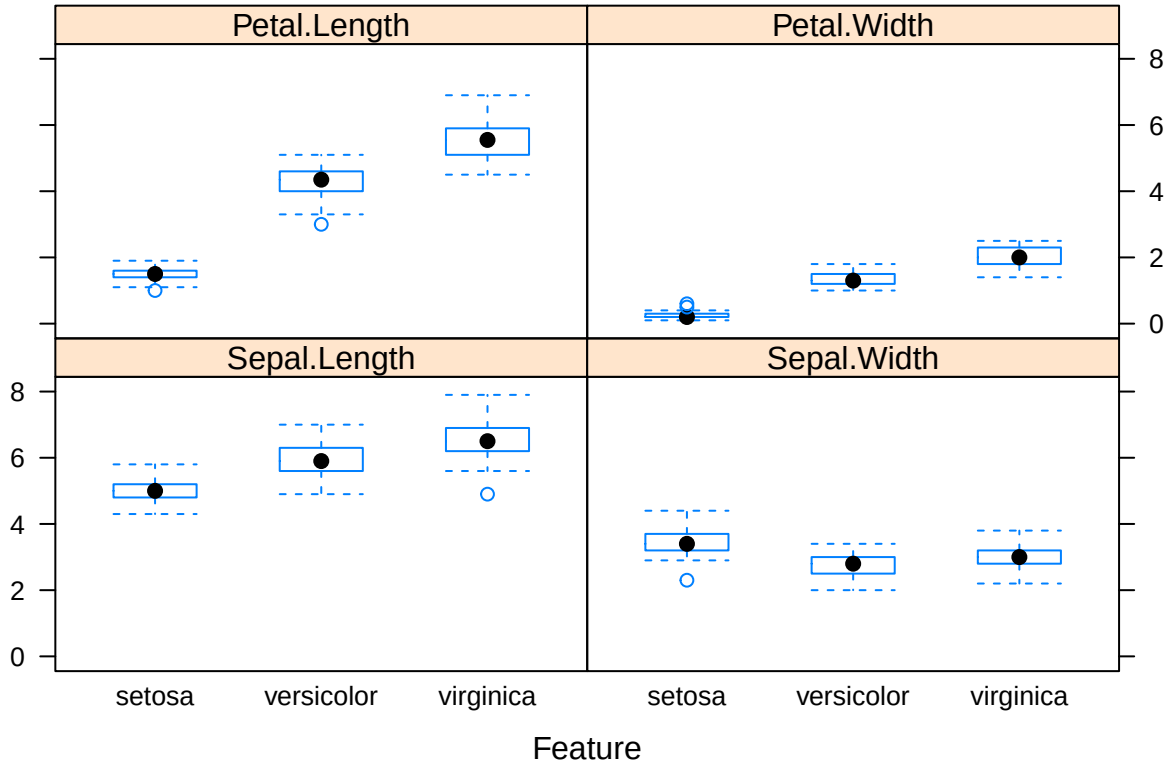
- 多变量可视化

```
> library(caret)
> library(ellipse)
> featurePlot(x = input.val, y = output.val, plot = "ellipse")
```



Scatter Plot Matrix

```
> featurePlot(x = input.val, y = output.val, plot = "box")
```



- 数据分解

把数据集分解为训练数据集和验证数据集，按着 8:2 比例基于各个样本类别的情况进行分解。

```
> # 获取原数据集的 80% 的行索引号
> validation.index <- createDataPartition(iris.data$Species, p = 0.8, list = FALSE)
> # 选择 20% 的数据用来验证模型
> validation.data <- iris.data[-validation.index, ]
> # 选择 80% 的数据用来训练和测试模型
> train.data <- iris.data[validation.index, ]
> # 10-折交叉验证
> control <- trainControl(method = "cv", number = 10)
> metric <- "Accuracy"
```

二、数据建模

1、交叉验证

为了选择最佳或者最优模型，采用一种典型的处理方法，交叉验证方法。常用 10-折交叉验证。将数据集分成十份，轮流将其中 9 份作为训练数据，1 份作为测试数据，进行试验。

```
> control <- trainControl(method = "cv", number = 10)
> metric <- "Accuracy"
```

2、构建模型

（1）LDA 线性判别分析算法

```
> # a) 线性算法
> library(e1071)
> # set.seed(7) 是为了保证每次生成的随机数都是一样的
> set.seed(7)
> lda.model <- train(Species ~ ., data = train.data, method = "lda", metric = metric,
+     trControl = control)
```

```
Loading required package: MASS
```

（2）rpart 回归树算法

```
> # b) 非线性算法
> library(rpart)
> set.seed(7)
> cart.model <- train(Species ~ ., data = train.data, method = "rpart", metric = metric,
+     trControl = control)
```

（3）RF 随机森林算法

```
> library(randomForest)
```

```
randomForest 4.6-12
Type rfNews() to see new features/changes/bug fixes.

Attaching package: 'randomForest'
The following object is masked from 'package:ggplot2':

    margin
```

```
> rf.model <- train(Species ~ ., data = train.data, method = "rf", metric = metric,
+     trControl = control)
```

三、模型评价

训练完模型之后，将它们添加到一个 list 中，然后调用 resamples() 函数。此函数可以检查模型是可比较的，并且模型都使用同样的训练方案（训练控制配置）。这个对象包含每个待评估算法每次折叠和重复的评估指标。

```
> results.model <- resamples(list(lda = lda.model, cart = cart.model, rf = rf.model))
> summary(results.model)
```

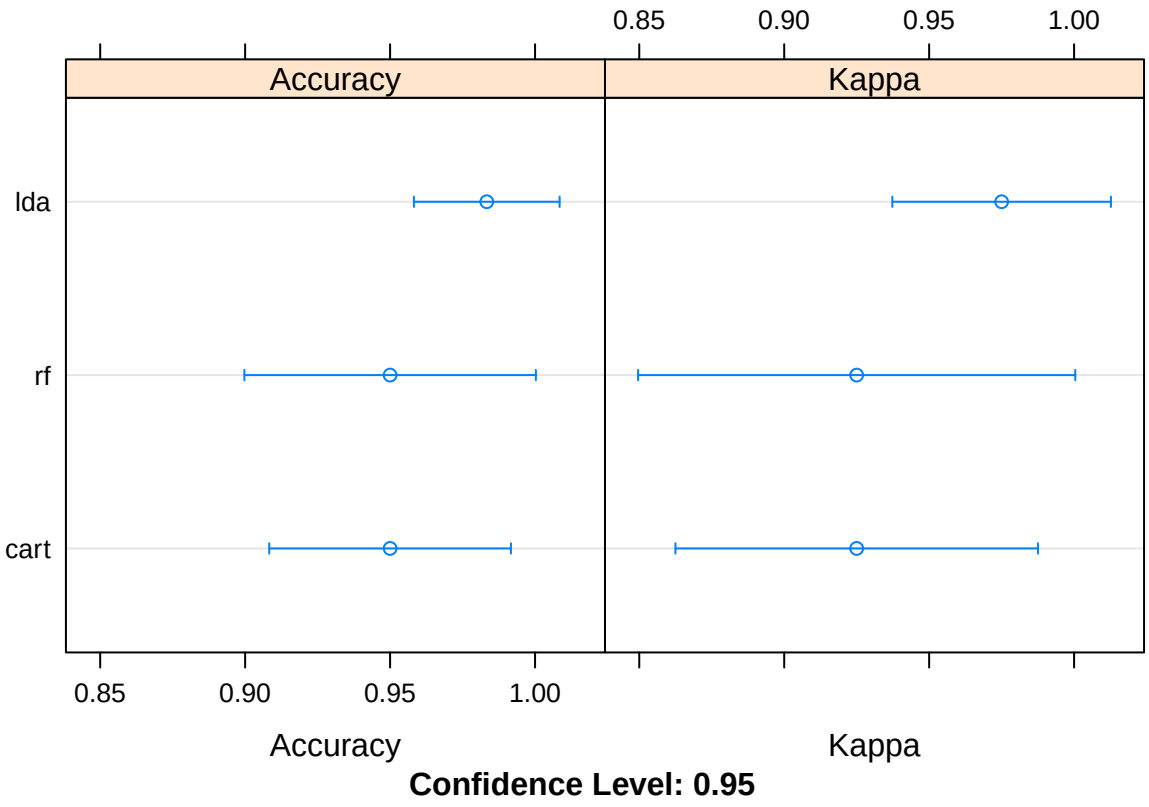
```
Call:
summary.resamples(object = results.model)
```

```
Models: lda, cart, rf
Number of resamples: 10

Accuracy
          Min.   1st Qu.    Median      Mean 3rd Qu. Max. NA's
lda  0.9166667 1.0000000 1.0000000 0.9833333       1    1    0
cart 0.8333333 0.9166667 0.9583333 0.9500000       1    1    0
rf   0.8333333 0.9166667 1.0000000 0.9500000       1    1    0

Kappa
      Min. 1st Qu. Median  Mean 3rd Qu. Max. NA's
lda  0.875   1.000 1.0000 0.975       1    1    0
cart 0.750   0.875 0.9375 0.925       1    1    0
rf   0.750   0.875 1.0000 0.925       1    1    0
```

```
> dotplot(results.model)
```



kappa 值通常会落在 0~1 之间，Kappa 值越高，代表训练效果越好。

由以上分析可知，针对 iris 数据集，LDA 算法最佳。

四、模型应用

利用以上算法对验证的数据集进行分类并使用混淆矩阵对预测结果进行评估。混淆矩阵（confusionmatrix），又称为可能性表格或是错误矩阵。它是一种特定的矩阵用来呈现算法性能的可视化效果，通常是监督学习（非监督学习，通常用匹配矩阵：matchingmatrix）。其每一列代表预测值，每一行代表的是实际的类别。这个名字来源于它可以非常容易的表明多个类别是否有混淆（也就是一个 class 被预测成另一个 class）。

1、LDA 线性判别分析算法

（1）分类预测

```
> # LDA 算法预测测试数据分类
> pred.result <- predict(lda.model, validation.data)
> pred.result
```

```
 [1] setosa     setosa     setosa     setosa     setosa     setosa
 [7] setosa     setosa     setosa     setosa     versicolor versicolor
[13] virginica  versicolor versicolor versicolor versicolor versicolor
[19] versicolor versicolor virginica  virginica  virginica  virginica
[25] virginica  virginica  versicolor virginica  virginica  virginica
Levels: setosa versicolor virginica
```

（2）预测结果评估

```
> confusionMatrix(pred.result, validation.data$Species)
```

```
Confusion Matrix and Statistics

            Reference
Prediction   setosa versicolor virginica
  setosa         10          0         0
  versicolor      0          9         1
  virginica       0          1         9

Overall Statistics

               Accuracy : 0.9333
                 95% CI : (0.7793, 0.9918)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 8.747e-12

                  Kappa : 0.9
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
```

```
Sensitivity                 1.0000           0.9000           0.9000
Specificity                 1.0000           0.9500           0.9500
Pos Pred Value              1.0000           0.9000           0.9000
Neg Pred Value              1.0000           0.9500           0.9500
Prevalence                  0.3333           0.3333           0.3333
Detection Rate              0.3333           0.3000           0.3000
Detection Prevalence        0.3333           0.3333           0.3333
Balanced Accuracy           1.0000           0.9250           0.9250
```

由混淆矩阵可知，LDA 线性判别算法模型对验证数据集分类的预测没有错误。Kappa 值为 1。

2、rpart 回归树算法

（1）分类预测

```
> pred.result <- predict(cart.model, validation.data)
> pred.result
```

```
 [1] setosa     setosa     setosa     setosa     setosa     setosa
 [7] setosa     setosa     setosa     setosa     versicolor versicolor
[13] virginica  versicolor versicolor versicolor versicolor versicolor
[19] versicolor versicolor virginica  virginica  virginica  virginica
[25] virginica  virginica  versicolor virginica  virginica  virginica
Levels: setosa versicolor virginica
```

（2）预测结果评估

```
> confusionMatrix(pred.result, validation.data$Species)
```

```
Confusion Matrix and Statistics

            Reference
Prediction   setosa versicolor virginica
  setosa         10          0         0
  versicolor      0          9         1
  virginica       0          1         9

Overall Statistics

               Accuracy : 0.9333
                 95% CI : (0.7793, 0.9918)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 8.747e-12

                  Kappa : 0.9
 Mcnemar's Test P-Value : NA
```

```
Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            0.9000           0.9000
Specificity                 1.0000            0.9500           0.9500
Pos Pred Value              1.0000            0.9000           0.9000
Neg Pred Value              1.0000            0.9500           0.9500
Prevalence                  0.3333            0.3333           0.3333
Detection Rate              0.3333            0.3000           0.3000
Detection Prevalence        0.3333            0.3333           0.3333
Balanced Accuracy           1.0000            0.9250           0.9250
```

由混淆矩阵可知，rpart 回归树算法将 2 个 versicolor 类的鸢尾花错误预测为 virginica 类，将 1 个 virginica 类错误预测为 versicolor 类。Kappa 值为 0.85。

3、RF 随机森林算法

（1）分类预测

```
> pred.result <- predict(rf.model, validation.data)
> pred.result
```

```
 [1] setosa     setosa     setosa     setosa     setosa     setosa
 [7] setosa     setosa     setosa     setosa     versicolor versicolor
[13] virginica  versicolor versicolor versicolor versicolor versicolor
[19] versicolor versicolor virginica  virginica  virginica  virginica
[25] virginica  virginica  versicolor virginica  virginica  virginica
Levels: setosa versicolor virginica
```

（2）预测结果评估

```
> confusionMatrix(pred.result, validation.data$Species)
```

```
Confusion Matrix and Statistics

            Reference
Prediction   setosa versicolor virginica
  setosa         10          0         0
  versicolor      0          9         1
  virginica       0          1         9

Overall Statistics

               Accuracy : 0.9333
                 95% CI : (0.7793, 0.9918)
```

```
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 8.747e-12

                  Kappa : 0.9
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            0.9000           0.9000
Specificity                 1.0000            0.9500           0.9500
Pos Pred Value              1.0000            0.9000           0.9000
Neg Pred Value              1.0000            0.9500           0.9500
Prevalence                  0.3333            0.3333           0.3333
Detection Rate              0.3333            0.3000           0.3000
Detection Prevalence        0.3333            0.3333           0.3333
Balanced Accuracy           1.0000            0.9250           0.9250
```

由混淆矩阵可知 RF 随机森林算法将 3 个 versicolor 类的鸢尾花错误预测为 virginica 类，将 1 个 virginica 类错误预测为 versicolor 类。Kappa 值为 0.8。

通过预测结果可知，三种算法中，针对此题目和此数据集，LDA 算法最佳。