Wine analysis

Łukasz Berwid

2/2/2020

1. Download and load data

Load data

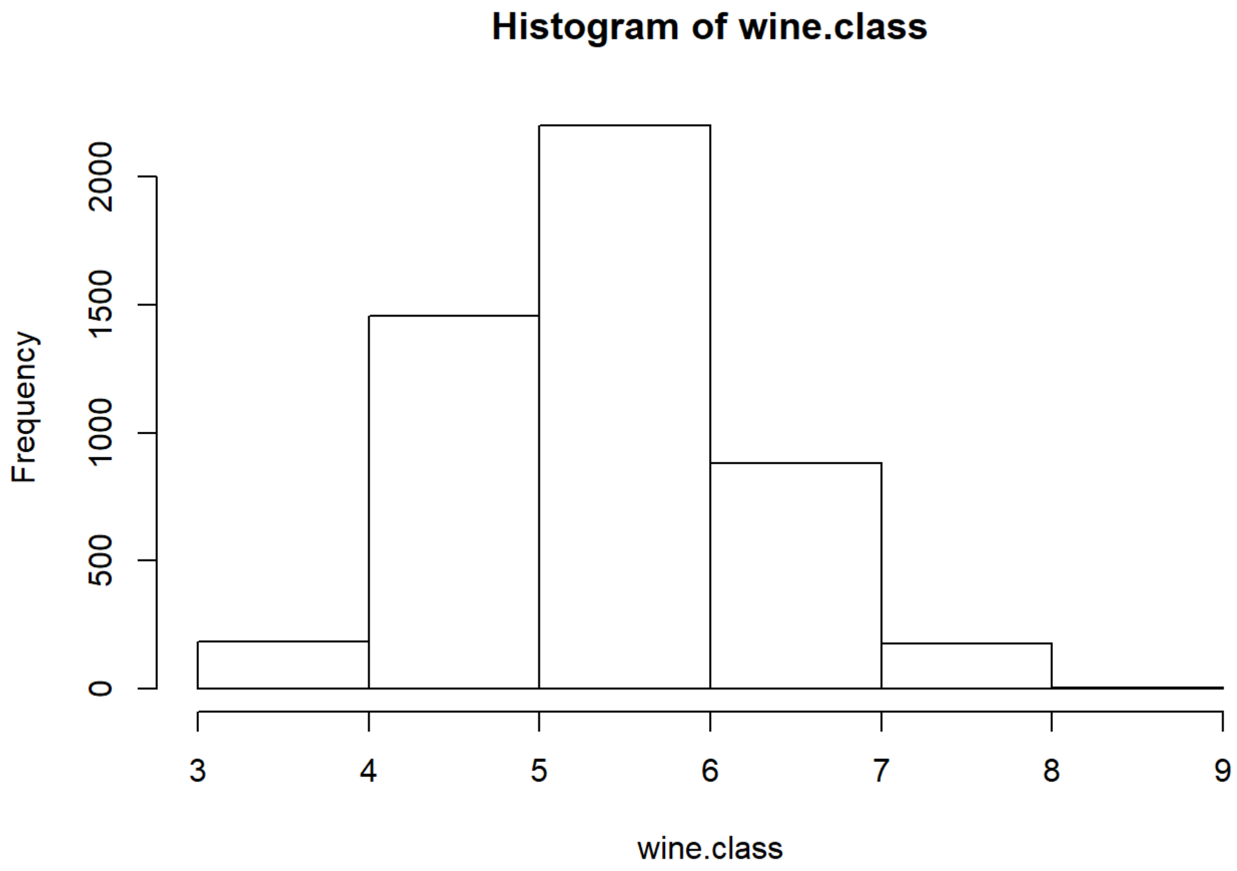
```
data('wine')
```

2. Rename columns

```
wine.class <- wine[,1]
wine.data <- wine[,-1]
```

3. Histogram

```
hist(wine.class, breaks = 5)
```



4. Scale the matrix

```
wine.data_scale <- scale(wine.data, center = TRUE, scale = TRUE)
```

5. Create training set

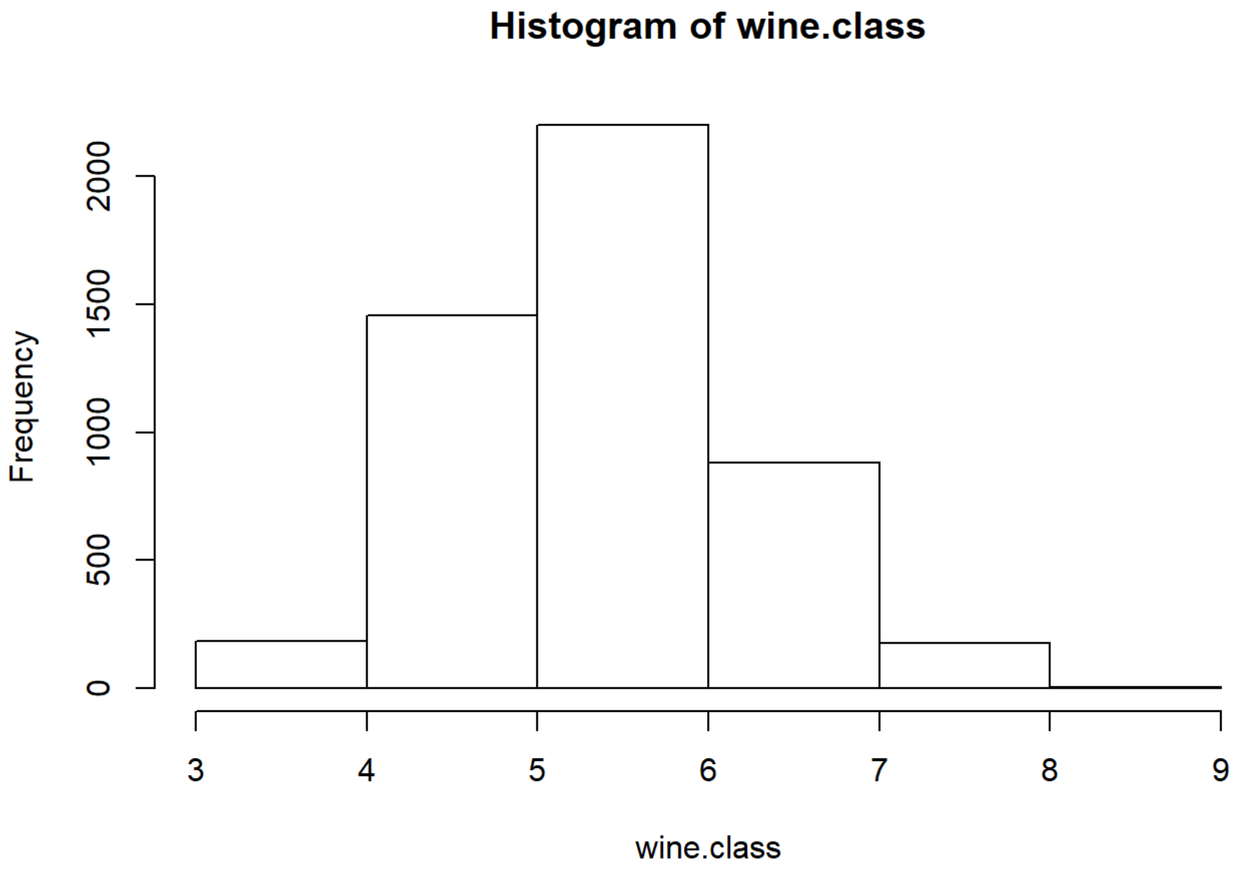
```
set.seed(123)
idx <- sample(nrow(wine.data_scale), 4000)

wine.train.data <- wine.data_scale[idx, ]
wine.train.class <- wine.class[idx]
```

6. Create test set

```
wine.test.data <- wine.data_scale[-idx, ]
wine.test.class <- wine.class[-idx]
```

7.

```
hist(wine.class, breaks = 5)
```



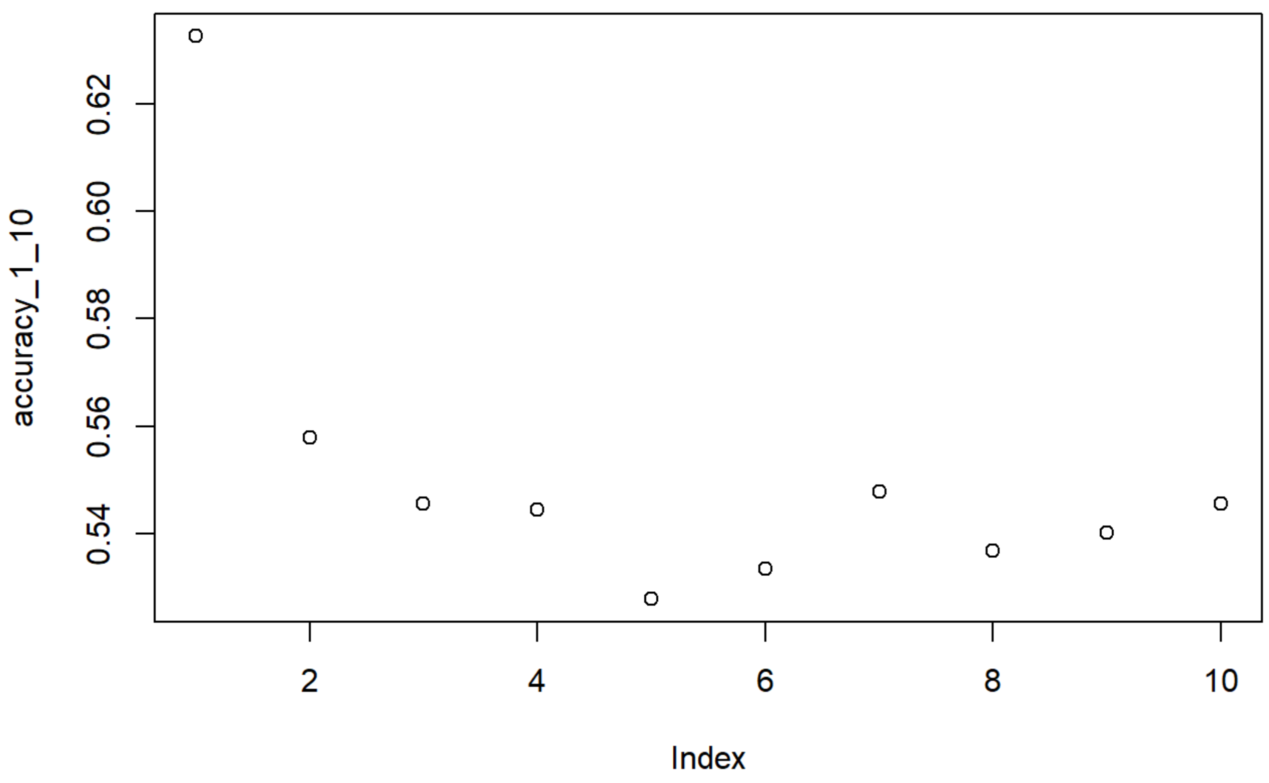
8. KNN fixed accuracy

```
wine_pred <- knn(wine.train.data, wine.test.data, wine.train.class, k = 7)
accuracy <- mean(wine_pred == wine.test.class)
accuracy
```

```
## [1] 0.5412027
```

9. KNN in range

```
knn_function <- function(k) {
    wine_pred <- knn(wine.train.data, wine.test.data,
    wine.train.class, k = k)
    accuracy <- mean(wine_pred == wine.test.class)
    return(accuracy)
}

k.values <- 1:10
accuracy_1_10 <- sapply(k.values, knn_function,simplify = TRUE, USE.NAMES = TRUE)
plot(accuracy_1_10)
```



Highest accuracy achieved for k:

```
max_val <- max(accuracy_1_10)
sprintf('highest accuracy achieved for k :%f value: %f', max_val, which(accuracy_1_10 == max_
val))
```

```
## [1] "highest accuracy achieved for k :0.632517 value: 1.000000"
```

KNN Implementation

```
euclidean <-  function(x, v) {
  return(sqrt(sum(x-v)^2))
}

custom_knn <- function(x_val, x_labels, k) {
  # calculate euclidean of each
  val <-  apply(wine.test.data, 1, function(x) euclidean(x, x_val))
  distances <- data.frame(val, wine.test.class)
  # get top k rows with closest distance
  top_k <- head(distances[with(distances, order(val)), ], k)
  colnames(top_k)[2] <- 'lab'
  print(top_k)
  # labels group by count
  labels_freq <- data.frame(with(top_k, table(lab)))
  most_common <- labels_freq[labels_freq$Freq == max(labels_freq$Freq),]
  #return most frequent label
  return(as.vector(most_common$lab))
}
```

Example

```
x_val <- head(wine.test.data, 1)
x_lab <- head(wine.test.class, 1)
print(x_lab)
```

```
## [1] 6
```

```
custom_knn(x_val, x_lab, 5)
```

```
##            val lab
## 1   0.00000000   6
## 885 0.02652948   7
## 254 0.04170168   6
## 706 0.04264208   5
## 3   0.04550814   8
```

```
## [1] "6"
```