Classification K-Nearest-Neighbors

Chris Bentz

13/01/2023

Load Packages

If the libraries are not installed yet, you need to install them using, for example, the command: install.packages("ggplot2"). For the Hrate package this is different, since it comes from github. The devtools library needs to be installed, and then the install_github() function is used.

```
library(ggplot2)
library(dplyr)
library(class)
library(gridExtra)
library(gmodels)
library(caret)
```

Load Data

Load data table with values per text file.

```
# load estimations from stringBase corpus
estimations.df <- read.csv("~/Github/NaLaFi/results/estimation10chars.csv")
# alternatively: "~/Github/NaLaFi/results/estimation100chars.csv"
# "~/Github/NaLaFi/results/estimation1000chars.csv"
#head(estimations10.df.)
```

Select relevant columns of the data frame, i.e. the measures to be included in classification and the ''corpus" or ''subcorpus" column.

```
estimations.subset <- estimations.df[c("corpus", "subcorpus", "huni.chars", "hrate.chars", "ttr.chars",
```

Remove NAs (whole row)

```
estimations.subset <- na.omit(estimations.subset)
```

Center and scale the data

```
estimations.scaled <- cbind(estimations.subset[1:2], scale(estimations.subset[3:ncol(estimations.subset)
```

Create Training and Test Sets

```
# Generating seed
set.seed(1234)
# Randomly generating our training and test samples with a respective ratio of 2/3 and 1/3
datasample <- sample(2, nrow(estimations.scaled), replace = TRUE, prob = c(0.67, 0.33))
# Generate training set
estimations.training <- estimations.scaled[datasample == 1, 3:ncol(estimations.scaled)]
# Generate test set
estimations.test <- estimations.scaled[datasample == 2, 3:ncol(estimations.scaled)]
```

Get training and test labels

```
# Generate training labels
training.labels <- estimations.scaled[datasample == 1, 1]
# Generate test labels
test.labels <- estimations.scaled[datasample == 2, 1]
```

Initialize data frame

```
knn.results <- data.frame(k = numeric(0), precision = numeric(0),
                          recall = numeric(0), f1 = numeric(0))
```

Building knn classifier

```
# choose maximum number of neighbors n
n = 10
# run a loop over different numbers of neighbors up to n
for (k in 1:n){
  # knn estimation of labels
  estimations.knn <- knn(train = estimations.training, test = estimations.test, cl = training.labels, k

  # model evaluation
  # creating a dataframe from known (true) test labels
  test.labels <- data.frame(test.labels)
  # combining predicted and known classes
  class.comparison <- data.frame(estimations.knn, test.labels)
  # giving appropriate column names
  names(class.comparison) <- c("predicted", "observed")
  # inspecting our results table
  head(class.comparison)
  # get confusion matrix
  cm <- confusionMatrix(class.comparison$predicted,
                        reference = class.comparison$observed)
  # print(cm)
  # get precision, recall, and f1 from the output list of confusionMatrix()
```

```
  f1 <- cm[["byClass"]]["F1"]
  recall <- cm[["byClass"]]["Recall"]
  precision <- cm[["byClass"]]["Precision"]

  # prepare data frame with results
  local.results <- data.frame(k, precision, recall, f1, row.names = NULL)
  local.results.rounded <- round(local.results, 2)
  # print(local.results.rounded)
  knn.results <- rbind(knn.results, local.results.rounded)
}
print(knn.results)
```

```
##     k precision recall   f1
## 1   1      0.53   0.44 0.48
## 2   2      0.61   0.41 0.49
## 3   3      0.65   0.41 0.51
## 4   4      0.67   0.39 0.49
## 5   5      0.64   0.39 0.48
## 6   6      0.64   0.39 0.48
## 7   7      0.64   0.39 0.48
## 8   8      0.63   0.41 0.50
## 9   9      0.58   0.37 0.45
## 10 10      0.64   0.39 0.48
```

Write to file.

```
write.csv(knn.results, file = "~/Github/NaLaFi/results/knn/knn_results_10chars.csv", row.names = F)
```