Fraud Detection
Rita P. Ribeiro

2022/2023
DCC - FCUP

4 Hands On: Classification with k-NN and Naive Bayes

Load (install, if necessary) the following packages: tidyverse, tidymodels, kknn, discrim, klaR.

```
library(tidyverse)
library(tidymodels)
```

4.1 Data set with numeric attributes only

1. Use the PimaIndiansDiabetes dataset from the mlbench package to answer the following questions. Inspect the data set.

```
data("PimaIndiansDiabetes",package="mlbench")
str(PimaIndiansDiabetes)
summary(PimaIndiansDiabetes)
```

(a) Use the function initial_split() to separate the data set into 70% / 30% for training and test set, respectively, stratified by the diabetes variable.

```
set.seed(1234)
#pima_split <- PimaIndiansDiabetes %>% initial_split(prop=.7)
pima_split <- PimaIndiansDiabetes %>% initial_split(prop=.7,strata=diabetes)
pima_split
```

(b) With the split above, obtain two separate data sets: train and test. Inspect them.

```
train <- training(pima_split)
test <- testing(pima_split)
summary(train$diabetes)
summary(test$diabetes)
```

(c) Define the recipe by establishing the task of building a classification model using the eight predictor variables to predict the target variable diabetes.

```
pima_rec <-  recipe(diabetes ~.,train)
pima_rec
```

(d) The numeric attributes have different ranges. Some algorithms can handle the scaling by themselves. Thus, it is a safe option to ensure that the scaling is performed in the training dataset as a pre-processing step and then applied to the test data set.

```
pima_rec <- pima_rec %>% step_normalize(all_numeric_predictors()) %>% prep()
pima_train <- pima_rec %>% bake(new_data=NULL)
pima_test <- pima_rec %>% bake(new_data=test)
```

(e) Models in parsnip (part of tidymodels meta-package) are specified by the **type** of model (e.g. nearest_neighbor), the **mode** of the model (e.g. classification), and the computational **engine**, i.e. the name of the R package that has the implementation to be used. In https://www.tidymodels.org/find/parsnip/, you have detailed information on the available algorithms that tidymodels do interface with.

Resorting to parsnip, choose the k-nn algorithm to build a classification model. In the particular case of k-nn, there is one single engine. Thus only mode has to be set.

```
library(kknn)
model_knn <- nearest_neighbor(mode="classification")
```

(f) Fit the k-nn algorithm to the train data and inspect the obtained model.

```
knn_fit <- model_knn %>%
  fit(diabetes ~ ., data = pima_train)
knn_fit
```

(g) Make predictions on the test set.

```
knn_preds <- predict(knn_fit,new_data = pima_test)
```

(h) Build up a tibble containing the true and predicted values. Obtain the confusion matrix and accuracy.

```
knn_preds <-
  pima_test %>% dplyr::select(diabetes) %>%
  bind_cols(predict(knn_fit, pima_test))


knn_preds %>% conf_mat(diabetes,.pred_class) %>% autoplot(type="heatmap")

knn_preds %>% accuracy(truth=diabetes,estimate=.pred_class)
```

(i) Add to your tibble the probability output of the classifier for each class. Now calculate ROC-AUC. Be sure to define properly what the ”relevant” class is.

```
knn_preds <-
  pima_test %>% dplyr::select(diabetes) %>%
  bind_cols(predict(knn_fit, pima_test)) %>%
  bind_cols(predict(knn_fit, pima_test,type="prob"))


knn_preds %>% roc_auc(truth=fct_relevel(diabetes,"pos"),estimate=.pred_pos)
```

(j) Plot the ROC Curve.

```
roc_curve(knn_preds,fct_relevel(diabetes,"pos"),.pred_pos) %>% autoplot()
```

(k) Repeat the process above, but now change the neighbors parameter value. Be critical regarding the results.

(l) Using the same experimental setting, run the Naive Bayes algorithm. Be critical regarding the results.

```
library(discrim)
library(klaR)

# to complete
```

Data with both numeric and categorical attributes

2. Use the tae data set to perform exercise 1 again.
This data set consists of evaluations of teaching assistants' performance.
Further details on this data set can be found here

Data with categorical attributes only

3. Use the nursery data set to perform exercise 1 again.
This data set consists of a collection of nursery school applications.
Further details on this data set can be found here.