dummy variable

Jan Alexander Jensen

2020/4/26

Handling nominal variable with dummy variable

Since we have several nominal and binary variables in our data, we need to convert them into a numeric format, to be able to use it for training. We will try using dummy encoding, where it takes only the value 1 or 0 to indicate whether the category is present or not. For example, the variable **occupations** in our dataset contains various values. For each type of value (occupation), we want to create a single column, specifying whether the data point belongs to this category.

Here, we use the function **dummy_cols** from package **fastDummies** to transform our nominal variables into dummy variables.

```r
library(fastDummies)

# **dummy_var** is used to
dummy_var <- function(data){

  # Since the library 'fastDummies' tansforms all factor variable into dummy variables,
  # we will convert our target "y" (factor) into a character variable
  # to avoid it being transformed to dummy variable.

  data$y <- as.numeric(as.character(data$y))

  # transform all factor variables to dummy variables,
  # and removes the original variables that were used to generate the dummy variables.
  data_dummy <- fastDummies::dummy_cols(data, remove_selected_columns=TRUE)

  # column name convention fix (mlr3 name convention - space to underscore)
  # this is needed because the function **dummy_cols** doesn't remove white spaces
  library(janitor)
  data_dummy <- clean_names(data_dummy)


  data_dummy <- as.data.frame(sapply(data_dummy, as.numeric))
  data_dummy$y <- as.factor(data_dummy$y)

  dummy_var <- data_dummy
}


## read data, and perform dummy encoding
dl_dummy_data <- read.csv2("../../credit_card_prediction/dl_na_data.csv", header = TRUE)
dl_dummy_data <- dummy_var(dl_dummy_data)

mf_dummy_data <- read.csv2("../../credit_card_prediction/mf_na_data.csv", header = TRUE)
mf_dummy_data <- dummy_var(mf_dummy_data)

mice_dummy_data <- read.csv2("../../credit_card_prediction/mice_na_data.csv", header = TRUE)
mice_dummy_data <- dummy_var(mice_dummy_data)
```

```r
# row, column
# The output shows that transformed all non-numeric variables into dummy variables,
# and now have 55 columns.
cat("dl_dummy_data\t", dim(dl_dummy_data), "\n")
cat("mf_dummy_data\t", dim(mf_dummy_data), "\n")
cat("mice_dummy_data\t", dim(mice_dummy_data), "\n")
```

```
## dl_dummy_data     25134 55
## mf_dummy_data     36457 55
## mice_dummy_data   36457 55
```