dummy variable

Jan Alexander Jensen

2020/4/26

Handling nominal variable with dummy variable

Since we have several nominal (categorical) variables in our data, we need to convert them into a numeric format, to be able to use it for training. Here we will try the dummy variable approach, where it takes only the value 1 or 0 to indicate whether the category is present or not. For each category, there will be an independent column. Here, we use the library 'fastDummies' to transform our nominal variables into dummy variables.

```r
library(fastDummies)
library(janitor)

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)
  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
}

# --------------------------------------------
# ----------------------- handle missing data
# ----------------------- OCCUPATION_TYPE
# -------------------------------------------


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)
```

```
# row, column, NA
cat("dl_dummy_data\t", dim(dl_dummy_data), any(is.na(dl_dummy_data)), "\n")
cat("mf_dummy_data\t", dim(mf_dummy_data), any(is.na(mf_dummy_data)), "\n")
cat("mice_dummy_data\t", dim(mice_dummy_data), any(is.na(mice_dummy_data)), "\n")
```

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

dummy vs one-hot

As dummy encoding is very similar to one-hot, we decided to leave out dummy encoding to reduce our training data size to speed up the tunning parameter process.