KNN & Regression imputation MSE

2024-10-16

```r
library(dplyr)
```

```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```r
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```r
library(DMwR2)
```

```
## Registered S3 method overwritten by ’quantmod’:
##   method            from
##   as.zoo.data.frame zoo
```

```r
library(Metrics)
```

```
##
## Attaching package: ’Metrics’

## The following objects are masked from ’package:caret’:
##
##     precision, recall
```

```r
data <- read.csv("AirQualityUCI.csv", sep = ";")
set.seed(123)

missing_indices <- sample(1:nrow(data), size = 0.2 * nrow(data))
data$RH_missing <- data$RH
data$RH_missing[missing_indices] <- NA
```

```r
train_data <- data %>% filter(!is.na(RH_missing))

train_data$AH <- as.numeric(gsub(",", ".", train_data$AH))
train_data$T <- as.numeric(gsub(",", ".", train_data$T))
train_data$RH_missing <- as.numeric(gsub(",", ".", train_data$RH_missing))
train_data$RH <- as.numeric(gsub(",", ".", train_data$RH))

lm_model <- lm(RH_missing ~ AH + T, data = train_data)

data$AH <- as.numeric(gsub(",", ".", data$AH))
data$T <- as.numeric(gsub(",", ".", data$T))
data$RH_missing <- as.numeric(gsub(",", ".", data$RH_missing))
data$RH <- as.numeric(gsub(",", ".", data$RH))
data$RH_predicted <- predict(lm_model, newdata = data)

data$RH_filled_regression <- data$RH_missing
data$RH_filled_regression[is.na(data$RH_missing)] <- data$RH_predicted[is.na(data$RH_missing)]


missing_rows_indices <- which(is.na(data$RH_missing))
mse_regression <- mse(data$RH[missing_rows_indices], data$RH_filled_regression[missing_rows_indices])

a = data$RH[missing_rows_indices]
b = data$RH_filled_regression[missing_rows_indices]

mse_regression <- mean((a - b)^2, na.rm = TRUE)
mse_regression
```

```
## [1] 168.95
```

```r
# MSE:168.95



#KNN

data_subset <- data %>% select(RH_missing, AH, T)
data_subset_imputed <- knnImputation(data_subset, k = 5)

mse_knn <- mean((data_subset_imputed$RH_missing[missing_rows_indices] - data$RH[missing_rows_indices])^2
mse_knn
```

```
## [1] 4.080879
```

```r
# MSE: 4.080879
```