Water Solubility

Paul J. Kowalczyk

2019-10-30

The ODSC Logo[1]

. . . and a link to ODSC West

Markdown citation[2].

[1]

[2] Allaire et al. (2019)

Read Data

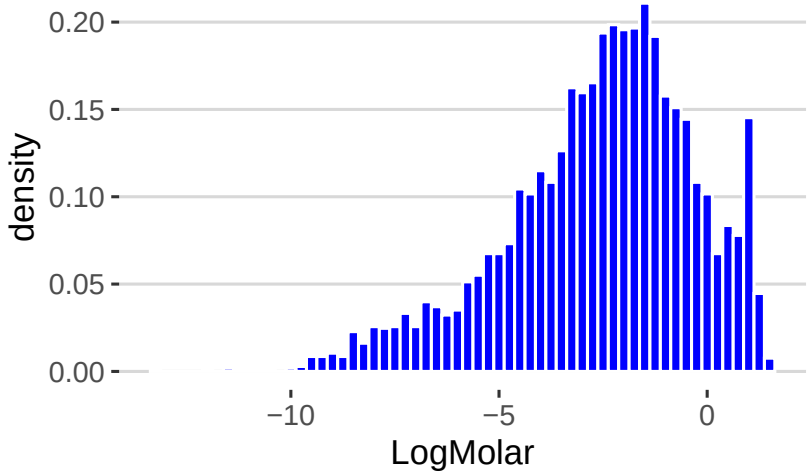
```
df <-
  read.csv('data/water_solubility.csv',
           header = TRUE,
           stringsAsFactors = FALSE) %>%
  na.omit()

kable(head(df[sample(nrow(df), 10), 1:8]), caption = 'A sampling of the input data.')
```

Table 1: A sampling of the input data.

	CAS	LogMolar	BalabanJ	BertzCT	Chi0	Chi0n	Chi0v	Chi1
3652	619-58-9	-3.9631	3.003401	240.2879	7.560478	5.042827	7.200326	4.698377
1481	616-44-4	-2.3899	3.049648	107.4989	4.405777	3.640299	4.456796	2.893847
895	122-20-3	0.6374	3.874273	102.2538	10.430721	8.642226	8.642226	5.913591
1941	2583-24-6	-1.2045	3.359860	159.1308	7.983128	5.279838	5.279838	4.625898
2088	4775-82-0	0.5240	3.265907	94.2039	7.276021	5.332624	5.332624	4.180739
1395	589-29-7	-0.3861	2.797251	167.9175	7.397341	5.618042	5.618042	4.863703

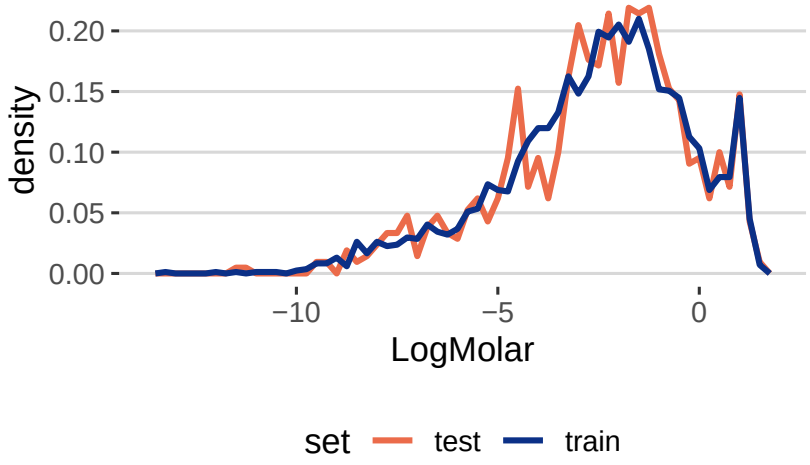
Distribution of Endpoint Values

```
LogMolar <-
  ggplot(df, aes(LogMolar, stat(density))) +
  geom_histogram(binwidth = 0.25, color = 'white', fill = 'blue') +
  theme(legend.position = "none") +
  ggthemes::theme_hc()
LogMolar
```



Build training and test sets

- stratified data partition: LogMolar
- 80% train / 20% test



Data Curation: near-zero variance

Initial number of variables in the dataset: 115.
The variables with near-zero variance are:

```
nzv <- caret::nearZeroVar(X_train, freqCut = 100/0)
names(df[ , nzv])
```

```
## [1] "NumHDonors"   "SMR_VSA6"
## [3] "SlogP_VSA7"   "SlogP_VSA9"
```

```
## [5] "VSA_EState10" "VSA_EState3"
## [7] "VSA_EState4"  "VSA_EState5"
```

Remove the near-zero variance variables

```
X_train <- X_train[ , -nzv]
X_test <- X_test[ , -nzv]
```

Number of variables in the dataset, following removal of those with near zero variance: 107

Data Curation: highly correlated variables

For all pairs of variables whose pairwise correlation exceeds 0.85, remove that variable whose mean correlation to all other variables is the greater.

Identify highly correlated variables

```
allCorrelations <- cor(X_train)
highCorr <- findCorrelation(allCorrelations, cutoff = 0.85)
```

Remove highly correlated variables

```
X_train <- X_train[ , -highCorr]
X_test <- X_test[ , -highCorr]
```

Having removed the highly correlated variables, there are 72 variables remaining.

Data Curation: names of removed variables (due to high correlation)

```
##  [1] "Chi0"
##  [2] "Chi1"
##  [3] "ExactMolWt"
##  [4] "HeavyAtomCount"
##  [5] "HeavyAtomMolWt"
##  [6] "Kappa1"
##  [7] "LabuteASA"
##  [8] "MinAbsPartialCharge"
##  [9] "MolMR"
## [10] "MolWt"
## [11] "NumAromaticRings"
## [12] "NumHAcceptors"
## [13] "NumHDonors"
## [14] "NumValenceElectrons"
## [15] "SMR_VSA7"
## [16] "VSA_EState10"
```

```
## [17] "Chi0n"
## [18] "Chi0v"
## [19] "Chi1n"
## [20] "Chi1v"
## [21] "Chi2n"
## [22] "Chi2v"
## [23] "Chi3n"
## [24] "Chi3v"
## [25] "FpDensityMorgan1"
## [26] "FpDensityMorgan2"
## [27] "MaxAbsEStateIndex"
## [28] "MaxAbsPartialCharge"
## [29] "Kappa2"
## [30] "NumAliphaticCarbocycles"
## [31] "NumAliphaticHeterocycles"
## [32] "NumAliphaticRings"
## [33] "SMR_VSA10"
## [34] "SMR_VSA5"
## [35] "NOCount"
```

Data Curation: Linear combinations

Identify variables that are a linear combination

```
comboInfo <- findLinearCombos(X_train)
names(X_train[ , comboInfo$remove])
```

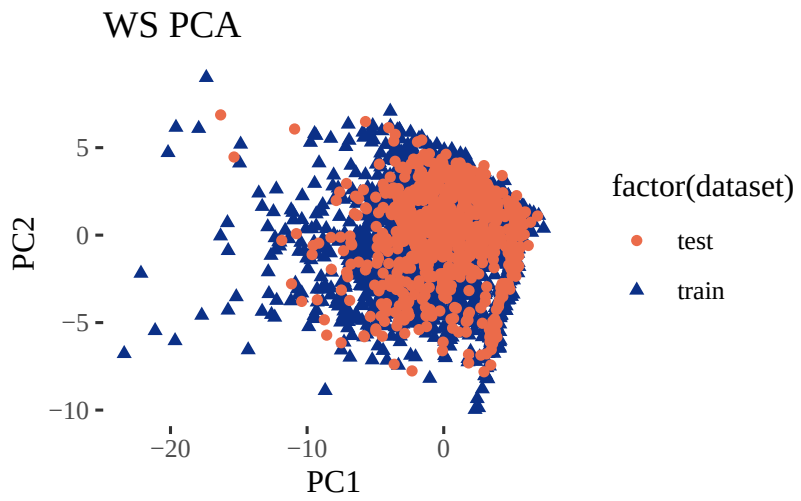
```
## [1] "NumSaturatedRings" "PEOE_VSA9"
## [3] "SlogP_VSA8"
```

Remove those variables that are a linear combination

```
X_train <- X_train[ , -comboInfo$remove]
X_test <- X_test[ , -comboInfo$remove]
```

Having removed variables that are a linear combination, there are 69 variables in the dataset.

Principal Components Analysis




```
file.edit(
  tint:::template_resources(
    'tint', '..', 'skeleton', 'skeleton.Rmd'
  )
)
```

References

JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *rmarkdown: Dynamic Documents for R*, 2019. URL https://CRAN.R-project.org/package=rmarkdown. R package version 1.16.