Chemical Sample Classification Report

Exploratory Data

Petter Lövehagen

2023-05-08

### Exploratory Data Analysis

Throughout EDA and model training, only the train dataset will be analysed. It has 1225 rows, with an overall minimum value of -0.48 (NA) and a maximum value of 16.73 (NA).

The variables have different scales and variances:

| Statistic | Minimum | Maximum |
| --- | --- | --- |
| **Mean** | 0.25 ( X8 ) | 13.79 ( X5 ) |
| **Variance** | 0.04 ( X10 ) | 1.52 ( X4 ) |
| **Range** | 1.44 ( X10 ) | 8.02 ( X14 ) |

**Table** **:** Minimum / Maximum Statistics in 'train'

**?(caption)**

At glance, we can see that there is a significant spread between the variables in terms of mean, variance and range. In order to mitigate against any one variable having undue influence and because nothing is known about the variables, they will be normalised or standardised so that they have the same scale from 0 to 1.

Observations \* many normal distributions \* some which differ across labels, eg. X7-X11

## NOTES BELOW

so there appears to be value in two different approaches \* correlation for dimension reduction \* needs domain knowledge to clarify and ratifiy, explain

* feature reduction - several variables which may be superfluous, measuring similar things or do not add to the model

To do \* normalise and scale \* check for outliers \* address outliers as appropriate \* correlation plot

To try:

interpretation - say what you see \* sparse - lots of variables which are not correlated with each other \* x17-20, and x7-10 appear to correlate, some neg, some pos

### dimension reduction:

PCA - identify principal components, linear combos of original variabes explaining the most variation in the data, plot to see in a lower dimension space t-SNE

PCA:

like a plot with PC1, PC2 the actual 5 classes (shape and label by label) pca arrows and variable labels

(see penguin example)