Chemical Sample Classification Report

Petter Lövehagen

2023-05-08

## Brief

The brief is to develop a classification tool to classify chemicals into classes.

In addition, the client is interested in knowing whether all variables are *necessary* for achieving optimal out-of-sample classification performance, or if a subset of measurements will suffice.

This report will aim to:

* explore, analyse and feedback on the provided dataset
* develop an optimised classification model
* determine whether all variables are required
* provide recommendations and options, as applicable

|  |
| --- |
| Note |
| Additional technical details are available in the code comments. |

## Data

The supplied dataset contains 2500 measurements of 20 numerical variables.

The analysis team and report writer were not supplied with any additional information about the data, e.g.:

* what has been measured
* variable importance, interactions
* measurement scale, units, reliability

Therefore, variables are treated equally and without bias in the analysis and modelling. Decisions are made on statistical grounds and stated assumptions.

|  |
| --- |
| Tip |
| Additional contextual information about the data could result in alternative modelling decisions and outcomes and should be considered. |

### Missing data

Before proceeding with [Exploratory Data Analysis](#sec-exploratory-data-analysis), it is important to explore the extent of missing data and whether there are patterns to the missingness. Some classification models assume no missing data, and depending on the amount, prevalence and patterning of missing data, different assumptions and techniques can be applied.

We investigate missingness to see whether the assumption that missing data is ‘missing at random’ (MAR) holds - that is, is the probability of missingness only dependent on **observed** variables? If so and there is a significant amount of missing data (>5%), multiple imputation approaches may be explored to *impute* missing values by estimation through statistical inference.

There are 49 rows with at least one missing value which is 0.09% of the entire dataset. There are significantly fewer missing values than the 5% threshold, so imputation is not required.

|  |
| --- |
| Figure 1: Missing Data |

If the missing data is not MAR - that is, it appears to depend on **unobserved** variables, there may be other, more appropriate imputation methods such as maximum likelihood imputation. Unobserved influences in this case may include the sample purity/quality, sample handling/measuring differences, measurement/equipment discrepancies or chemical compositions of the samples.

*Imputation* can introduce bias into the analysis if the assumptions of the imputation method are violated, or the imputed values differ significantly from true missing values.

|  |
| --- |
| Caution |
| The client should consider missingness in their data:   * Are there any concerns regarding missing data? * Is the equipment, staff, technique, facilities, chemicals, etc. consistent? * What are the reasons for missing data? Can they be mitigated? * Can data be collected where it has been identified as missing? * Is there likely to be (more, less, similar amount of) missing data in the future? |

On a case basis, no single observation has more than 1 missing value. This means that, at most, a sample is missing 4.76% of its data. If we group by ‘label’ we see that Class E has the most missing data at 0.52%.

| Label | Count | Percentage |
| --- | --- | --- |
| A | 10 | 0.40 |
| B | 6 | 0.24 |
| C | 10 | 0.40 |
| D | 10 | 0.40 |
| E | 13 | 0.52 |

Table 1: Missing Data by ‘label’

Variable X9 has the most missing values with 6, accounting for 0.24% of this variable’s data.

| Variable | Count | Percentage |
| --- | --- | --- |
| X1 | 4 | 0.16 |
| X2 | 3 | 0.12 |
| X3 | 2 | 0.08 |
| X4 | 1 | 0.04 |
| X5 | 0 | 0.00 |
| X6 | 4 | 0.16 |
| X7 | 0 | 0.00 |
| X8 | 2 | 0.08 |
| X9 | 6 | 0.24 |
| X10 | 1 | 0.04 |
| X11 | 2 | 0.08 |
| X12 | 1 | 0.04 |
| X13 | 2 | 0.08 |
| X14 | 2 | 0.08 |
| X15 | 4 | 0.16 |
| X16 | 2 | 0.08 |
| X17 | 4 | 0.16 |
| X18 | 5 | 0.20 |
| X19 | 2 | 0.08 |
| X20 | 2 | 0.08 |
| label | 0 | 0.00 |

Table 2: Missing Data by Variable

The upset plot below shows the five variables with the most missing values and confirms that there are no instances where there are missing values in two variables in the same observations.

|  |
| --- |
| Figure 2: Upset plot of missing data |

The target variable (‘label’) has no missing values and when we group data by the classes,  
patterns of missingness begin to emerge. It appears that missing values for sequential variables (measurements) may be correlated with the chemical class. There is a clear stepped pattern in the heatmap below and the ‘Missing Data’ plot above.

|  |
| --- |
| Warning |
| Is there anything systemic, in terms of missingness:   * X1-X4 is *only* missing for chemical A * X6, X8 are *only* missing for chemical B * X9-X12 is *only* missing for chemical C * X13-X16 are *only* missing for chemical D * X17-X19 are *only* missing for chemical E   Is there a relationship between the variables, in terms of what they measure? |

|  |
| --- |
| Figure 3: Missing data per variable, grouped by Chemical Class (label) |

[include citation to missing data references]

|  |
| --- |
| Decision |
| Given the minimal amount of missing data - it will be removed before exploratory data analysis, data splitting and modelling.  If this dataset is representative, the missing data may not be an issue - assuming the pattern identified above is understood. |

### Splitting data

Before proceeding with additional exploratory analysis, we split the dataset into three random subsets:

* train - used to *train* the classification model to *learn* the relationship between the variables and the label
* validation - used to *validate* the performance of the trained model and to tune any hyperparameters; determines how well the model generalises to ‘unseen’ data
* test - used to *evaluate* the final performance of the model post training and tuning; this data is ony used once and kept separate.

As EDA will explore and visualise relationships between variables, potentially creating new features, it is essential to split the data so that we do not introduce bias into the process. Any insights or observations during the EDA phase will therefore emerge solely from the training dataset.

The only thing we will check in advance is the target variable balance. Given that it is not perfectly balanced, the data subsets will be split with ‘stratification’ which ensures each partition has a representative proportion of each class.

| label | Freq |
| --- | --- |
| A | 508 |
| B | 492 |
| C | 492 |
| D | 487 |
| E | 472 |

Table 3: Frequency Table

|  |
| --- |
| Decision |
| The clean dataset is sufficiently large to split into:   * train - 50% * validation - 25% * test - 25% |

### Exploratory Data Analysis

Throughout EDA and model training, we will only use the train dataset which has a total of 1225 rows, with a minimum value of -0.5177491 (X7) and a maximum value of 16.7575557 (X5).

The variables have different scales and variances:

Table 4: Statistics across Variables

| Statistic | Min | Max |
| --- | --- | --- |
| Mean | 0.2527747 (X8) | 13.7976095 (X5) |
| Variance | 0.0390197 (X10) | 1.5727625 (X4) |
| Range | 1.4207844 (X18) | 8.371569 (X14) |

[histograms] [outliers] [pair plots] [correlation]