Feature Extraction: Discussion

Dingcheng Yue

(Document page limit: 1xA4 pages)

Reflections and Findings while doing Task 1

I used python script for this task. Firstly, I conform the dataset to a uniformed representation by hand. For the dataset itself, I removed the extra column and rows that are empty and add/modify/remove the headers to a single line. I also conformed the encoding and new line symbols of the file. I also convert all the names to lower case for consistency.

During this step, there are several files that are illegible as they do not contain data so I disregard them.

I found that the different data files all have different length, while some records contain only 10s of rows and others contain more than 1000s of rows. I also find that the data

Description of additional features you extracted from the data at hand in Task 2

1. The mean of the first moment from Euclidian distance.
2. The maximum difference from the temperature
3. The maximum difference from the EDA

Reflections on the full data set

1. The data all have different length, while some records contains only 10s of rows and others contains more than 1000s of rows.
2. The short data could be problematic if one data column did not change. For example, EDA from DY_af_03 does not change at all, thus it is not normalisable given the formula from the equation.
3. Some data are partial, I.e. (AK_Ha data), I still include the row and keep the missing data blank

Reflections on the feature set

1. The temperature does not vary too much, I found that for thoes does not containing enough data, the temperature does not vary at all.
2. The EDA is also an delicated indicator, and it could provides information
3. Even x, y, z are behaving similar, they are at different scale for some motion, thus the normalised version should be more informative when we do classification.
4. I realised that the maximum difference are not robust to outliers, thus, I should use it along with the mean and first/second difference

End of file