Week 17 - Feature Selection

Data Pre-Processing

Transforming raw data into clean data that we can apply models on.

- Remove or fill with missing values.
- Transform categorical / ordinal into numerical values.
- Deal with outliers.

Feature Extraction and Selection

- Different techniques for manipulating features
- Creating new features / modifying existing ones
- Goal: achieve better results in ML algorithm
  - Variables with large variance are a problem (read: large outliers).
  - Skewed data will lead to skewed coefficients.

Scaling

- Range and interval over which the values are distributed can vary greatly.

- Some algorithms, particularly those which are based on distance calculations, will put more weight on those features that display large changes in value, interpreting these features as artificially more important.

- For these algorithms, it is important that we scale our features, or put features with naturally different scales on the same scale.

Standardization

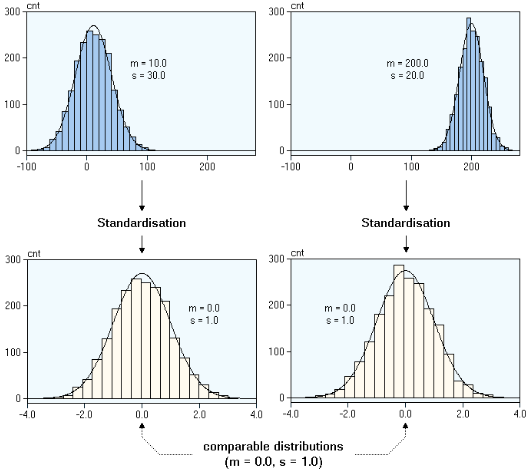
- Impose several statistical properties on the variable: the mean value is set to 0, and the standard deviation is set to 1.

- So-called z-score normalization:

$$x_{stand} = \frac{x - \mu}{\sigma}$$

- Standardization reduces the effects of outliers in the feature. Additionally, it allows two features with dissimilar scales or units to be compared.

Standardization



Normalization

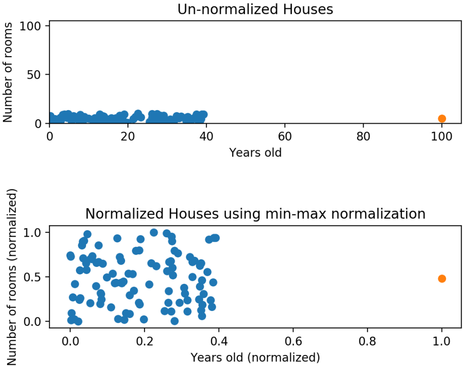
- The feature is rescaled to a range between 0 and 1, without any changes in its original distribution within that range.

- Subtracting the minimum value of the feature from each value of the feature, and dividing by the difference between the largest value and the minimum value

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Also called Min-Max Normalization.

Normalization



Only the range of the data has changed, any outliers in the feature are still included. Normalization is more useful in cases where your data has few outliers but highly variable ranges.

Correlation Thresholds

- Focuses on removing repetitive data from the dataset.
- Features that are highly correlated to each other are effectively communicating the same information twice, and overweighting the importance of this information in determining the outcome.
- A threshold for maximal correlation between features is determined, and one of those features whose correlation is higher than the threshold is removed, leaving only one of the features.
- Reduce the redundant information in the dataset, as well as multicollinearity between the features. Problem: analyst must determine threshold.

Backwards Elimination vs. Forwards Selection

- **Forward Selection**: multiple models are created, each using just one feature to as the independent variable being used to predict the outcome. Then, features are added to each model one by one, starting with the most significant first. When the accuracy of the model does not improve with the addition of features, the process is stopped and the dataset with just those features is used.

- **Backward Selection**: All features are included as independent variables used to determine the outcome. Then, starting with the least significant features (those with the highest p-values), features are removed from the model one by one until the accuracy does not improve when more features are removed.

Process Example

Does the data make sense? - THINK about what you are looking at, compute summary statistics, plot histograms.

1. Are there any mistakes in my data?
2. How is the data encoded?
3. Are the ranges sensible? Are there any variables for which I should change the scale?
4. How are the variables distributed? Are there any variables that barely fluctuate?

Process Example

4. Do the features have outliers? (box plots!)
5. Do the variables have null values? Is the number of null values acceptable?
6. Do we have duplicate rows? (unintentionally!)
7. Are variables in our dataset correlated with one another or with the outcome? How strong is this correlation? (heatmaps!)