UNIVERSIDADE CATÓLICA PORTUGUESA

BRAGA

Machine Learning

Session 3 - T

Data Scaling and Feature Selection

Ciência de Dados Aplicada

2023/2024

Data Scaling

- Data scaling refers to the procedure of adjusting the range of features within a dataset to a **comparable scale;**

- Real-world datasets often contain features with **different orders of magnitude, ranges, and measurement units**.

Car	Model	Volume	Weight	CO2
Toyota	Aygo	1.0	790	99
Skoda	Citigo	1.0	929	95
Fiat	500	0.9	865	90
Mini	Cooper	1.5	1140	105
Skoda	Fabia	1.4	1109	90
...	...	...	...	...

Why do we need scaling?

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Some machine learning models are **sensitive to feature scale;**
- Features with larger scales may **dominate the learning process**;
- Models may **converge faster**;
- Model **performance may improve** (specially for models that rely on **distance metrics**);

Data Scaling Methods

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Standardization** (Z-score normalization): centers the data around mean 0 and standard deviation 1:

$$z = \frac{x_i - \mu}{\sigma}$$

- **Normalization**: scales the data between 0 and 1;

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

- **Min-Max Scaling**: scales data betwenn a maximum and minimum value;

$$X_{scaled} = \frac{X - \min(X)}{\max(X) - \min(X)}(NewMax - NewMin) + NewMin$$

Data Scaling Methods

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

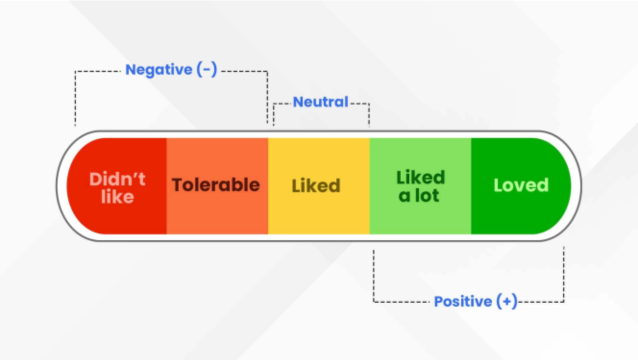
- **Robust Scaling**: Scales data based on the interquartile range, making it robust to outliers;

$$X = \frac{X - Q_1(X)}{Q_3(X) - Q_1(X)}$$

- **Log transformation**: scales data by applying the natural logarithm function;

$$X = log(X)$$

- **Ordinal scaling**: assign integer values to categories with a meaningful order.



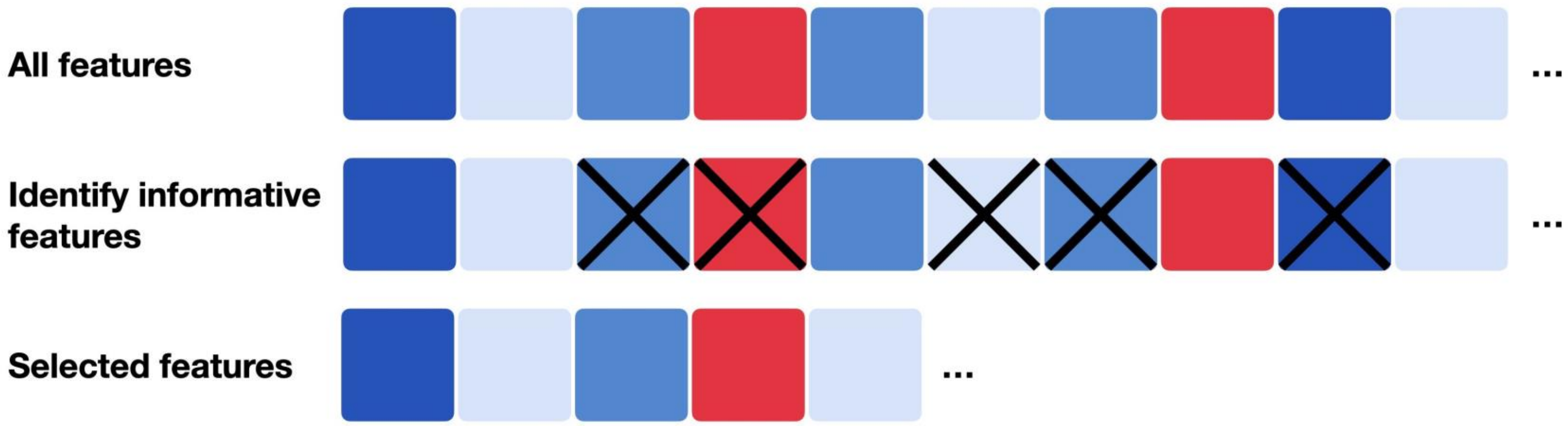
Which Scaling Method to Choose?

- Choose the method that aligns with your **data type**, **distribution**, and **Machine Learning model** to use;

- Methods to use:
  - **Continuous data**:
    - Uniform distribution: Min-Max scaling / normalization;
    - Normally distributed: Z-score standardization;
    - Data with outliers: Robust scaling;
    - Skewed or exponential distribution: Log transformation.
  - **Categorical data**:
    - Ordinal: Ordinal scaling;
    - Nominal: other strattegies like one-hot encoding and frequency encoding.

Feature Selection

- Feature selection is a process of **selecting a subset of the initial features while minimizing the loss of information** related to the final task (classification, regression, etc);



Why Feature Selection?

- In many cases, there are multiple advantages in **reducing the number of input features** used in a model.

- This is especially important when:
  - Dealing with **noisy data**;
  - Handling numerous **low-frequency features**;
  - Data has too **many features compared to samples**;
  - Managing **complex models**;
  - …

https://medium.com/barnbridge/barnbridge-dao-built-for-the-future-3735fbd671c5

Why Feature Selection?

- The reduction of the number of features can:
  - Improve the **model's performance**;
  - Enhance **optimization stability** by removing **multicollinearity**;
  - Increase **computational eficiency**;
  - Reduce **cost of future data collection**;
  - Simplify the model, making it **easier to understand and interpret**;
  - ...

- However, it is **not always a necessary step**:
  - Some models have **implicit feature selection**
    - Tree-based models;
  - Models with **regularization**
    - LASSO;
    - RIDGE.

https://www.kaggle.com/code/heyspaceturtle/feature-selection-is-all-u-need

Feature Selection Algorithms

- From the **label perspective** they can be:
  - **Supervised**;
  - **Unsupervised**.

- From the **selection strategy perspective** they can be:
  - **Filter** methods;
  - **Wrapper** methods;
  - **Embbeded** methods.

Feature Selection: Filter Methods

- Filters methods evaluate feature relevance based on **intrinsic data characteristics:**
  - First, features are individually **ranked** based on specific criteria such as **distance, correlation, or entropy;**
  - Second, the **best-ranked features** are selected using a predetermined **threshold**.

Feature Selection: Filter Methods

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Unsupervised**:
  - Unsupervised filters compute a metric for each feature **based solely on its values**;
  - Metrics include those measuring variability, such as **variance** (continuous variables) and **entropy** (discrete variables);
  - Selection can occur through **ranking** (e.g., percentile) or by **absolute value**, by keeping all features with a value below/above a specified threshold.

- **Supervised**:
  - Supervised filters use a metric calculated for each input feature, **comparing its values to those of the output**;
  - Metrics include **mutual information** (discrete variables) and **correlations** (continuous variables);
  - Scores may also rely on **univariate statistical tests** applied to the paired sets of values;
  - Different tests may be applied depending on the type of features.

Feature Selection: Filter Methods

- **Supervised**:
  - For **classification problems** (discrete output):
    - **T-test** may be used for continuous and binary output; for multiclass output, **one-way ANOVA** may be used;
    - **Chi-square** is used for counts/frequencies or binary variables.
  - For **regression problems** (continuous output):
    - **Correlation** is an option for continuous inputs (Pearson or Spearman);
    - **Kendall's rank correlation** for discrete inputs (ordinal).
  - In both cases, **mutual information** can be used as a **non-parametric** alternative.

Feature Selection: Filter Methods

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Advantages**:
  - Independent of a learning model;
  - Computationally efficient;
  - Suitable for high dimensional data;
- **Disadvantages**:
  - Interactions between features are ignored;
  - May fail to handle redundant features;
  - No interaction with the learning algorithm.

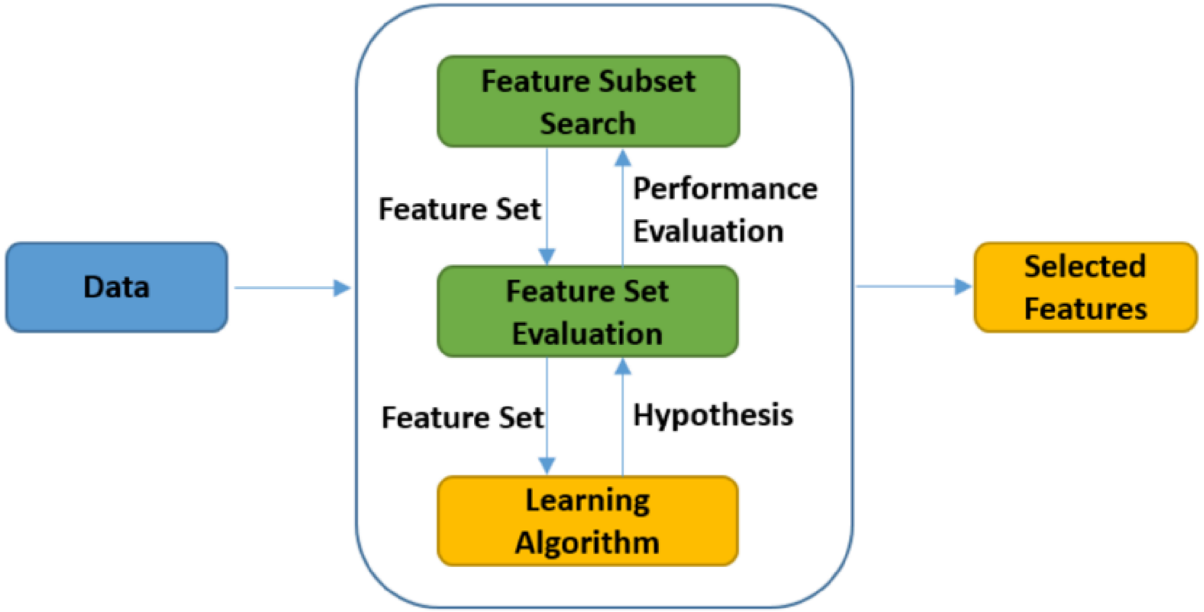
Feature Selection: Filter Methods

- **Examples**:
  - **VarianceThreshold**: keeps only features whose variance exceeds a specified threshold;
  - **SelectKBest**: selects the top k (user-defined parameter) features based on a scoring function (mutual information, chi2, ANOVA, etc);
  - **SelectPercentile**: selects the top percentage (percentile) of features based on a scoring function (mutual information, chi2, ANOVA, etc).
  - **SelectFpr**, **SelectFdr**, and **SelectFwe**: sseatures are selected based on their significance according to statistical tests. SelectFpr controls the false positive rate, SelectFdr controls the false discovery rate, and SelectFwe controls the family-wise error rate.

Feature Selection: Wrapper Methods

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Wrapper methods directly involve a **learning algorithm** in the feature selection process.
- **Subsets of features** are used to train and test a model iteratively, selecting the subset that optimizes a performance metric.



Feature Selection: Wrapper Methods

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Subset selection method:
  - **Forward search**:
    - Start with no features;
    - Greedily include the most relevant feature;
    - Stop when the desired number of features is selected.
  - **Backward search**:
    - Start with all the features;
    - Greedily remove the least relevant feature;
    - Stop when the desired number of features is reached.

Feature Selection: Wrapper Methods

- **Advantages:**
  - Better performance attainability;
  - Take into account interaction between features;
  - Identify feature interactions of higher order.

- **Disadvantages:**
  - Computationally expensive;
  - Prone to overfitting;
  - The learning algorithm is built from scratch for each subset.

Feature Selection: Wrapper Methods

- **Examples**:
  - **RFE:** recursively selects features by training a model, removing the least important features, and repeating until the desired number of features is reached;
  - **SequentialFeatureSelector**: evaluates different combinations of features by adding or removing features iteratively based on model performance;
  - **SelectFromModel**: selects features based on importance weights provided by a pre-trained model.
  - **Boruta**: all-relevant feature selection method that identifies important features by comparing their importance with that of random shadow features.

Feature Selection: Embedded Methods

- Feature selection is **integrated directly into the model training** process;
- Features are selected or discarded based on their **importance to the model's performance** during training;
- Some examples include **tree-based models** and models with **regularization** (L1/L2).

Feature Selection: Embedded Methods

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Advantages:**
  - Faster than wrapper methods;
  - Take into account interactions between features;
  - Identify feature dependencies;

- **Disadvantages:**
  - Specific to the learning algorithm;
  - Selection dependent on the learning algorithm.

Feature Selection: Embedded Methods

- **Examples**:
  - **Lasso Regression**: performs feature selection by penalizing the absolute size of the regression coefficients, effectively shrinking some coefficients to zero and eliminating corresponding features.
  - **Ridge Regression:** performs feature selection by penalizing the square of the regression coefficients, which encourages smaller coefficients and effectively shrinks the impact of less important features.
  - **Random Forests**: ensemble learning method that naturally performs feature selection by assessing the importance of features based on how much they decrease node impurity across multiple decision trees.
  - **Gradient Boosting Machines:** ensemble learning method that builds decision trees sequentially, each focusing on the residuals of the previous trees, effectively performing feature selection by giving more importance to relevant features.

Resources:

- Zheng, A. (2018). Feature Engineering for Machine Learning. Sebastopol, CA: O’Reilly Media.

- Kuhn, M., & Johnson, K. (2019). Feature engineering and selection. Philadelphia, PA: Chapman & Hall/CRC.