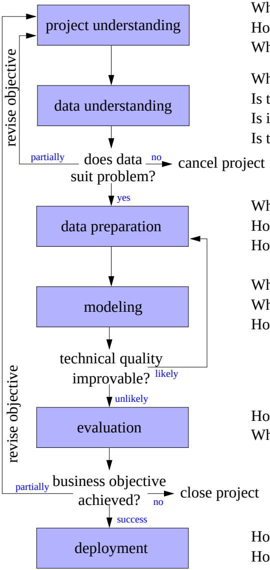
Data Preparation



What exactly is the problem, the expected benefit?
How would a solution look like?
What is known about the domain?

What data do we have available?
Is the data relevant to the problem?
Is it valid? Does it reflect our expectations?
Is the data quality, quantity, recency sufficient?

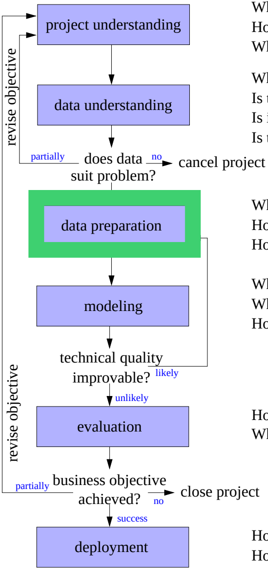
Which data should we concentrate on?
How is the data best transformed for modeling?
How may we increase the data quality?

What kind of model architecture suits the problem best?
What is the best technique/method to get the model?
How good does the model perform technically?

How good is the model in terms of project requirements?
What have we learned from the project?

How is the model best deployed?
How do we know that the model is still valid?

Data Preparation



What exactly is the problem, the expected benefit?
How would a solution look like?
What is known about the domain?

What data do we have available?
Is the data relevant to the problem?
Is it valid? Does it reflect our expectations?
Is the data quality, quantity, recency sufficient?

Which data should we concentrate on?
How is the data best transformed for modeling?
How may we increase the data quality?

What kind of model architecture suits the problem best?
What is the best technique/method to get the model?
How good does the model perform technically?

How good is the model in terms of project requirements?
What have we learned from the project?

How is the model best deployed?
How do we know that the model is still valid?

Data understanding vs Data preparation

Data understanding provides general information about the data like

- the existence and partly also about the character of missing values,
- outliers,
- the character of attributes and
- dependencies between attribute.

Data understanding vs Data preparation

Data understanding provides general information about the data like

- the existence and partly also about the character of missing values,
- outliers,
- the character of attributes and
- dependencies between attribute.

Data preparation uses this information to

- select attributes,
- reduce the dimension of the data set,
- select records,
- treat missing values,
- treat outliers,
- integrate, unify and transform data and
- improve data quality.

Feature extraction

refers to construct (new) features from the given attributes.

Example

Find the best workers in a company.

- Attributes :
  - the tasks, a worker has finished within each month,
  - the number of hours he has worked each month,
  - the number of hours that are normally needed to finish each task.
- These attributes *contain* information about the efficiency of the worker.
- But instead using these three “raw” attributes, it might be more useful to define a new attribute *efficiency*.
- $\text{efficiency} = \frac{\text{hours actually spent to finish the tasks}}{\text{hours normally needed to finish the tasks}}$

Feature selection

Feature selection refers to techniques to choose a subset of the features (attributes) that is as small as possible and sufficient for the data analysis.

Feature selection includes

- removing (more or less) irrelevant features and
- removing redundant features.

Feature selection techniques

- **Selecting the top-ranked features.**
  Choose the features with the best evaluation when single features are evaluated.

Feature selection techniques

- **Selecting the top-ranked features.**
  Choose the features with the best evaluation when single features are evaluated.
- **Selecting the top-ranked subset.**
  Choose the subset of features with the best performance. This requires exhaustive search and is impossible for larger numbers of features. (For 20 features there are already more than one million possible subsets.)

Feature selection techniques

- **Selecting the top-ranked features.**
  Choose the features with the best evaluation when single features are evaluated.
- **Selecting the top-ranked subset.**
  Choose the subset of features with the best performance. This requires exhaustive search and is impossible for larger numbers of features. (For 20 features there are already more than one million possible subsets.)
- **Forward selection.**
  Start with the empty set of features and add features one by one. In each step, add the feature that yields the best improvement of the performance.

Feature selection techniques

- **Selecting the top-ranked features.**
  Choose the features with the best evaluation when single features are evaluated.
- **Selecting the top-ranked subset.**
  Choose the subset of features with the best performance. This requires exhaustive search and is impossible for larger numbers of features. (For 20 features there are already more than one million possible subsets.)
- **Forward selection.**
  Start with the empty set of features and add features one by one. In each step, add the feature that yields the best improvement of the performance.
- **Backward elimination.**
  Start with the full set of features and remove features one by one. In each step, remove the feature that yields to the least decrease in performance.

Reasons for using only a subsample

Faster computation

Cross-Validation with training and test set

Timeliness. Data which is outdated can be removed.

Representativeness. Is the given sample matching the whole population? If not and we do have information about the true distribution, select a representative subsample. (e.g. there are more women than men in a questionnaire for computer scientists)

Rare events. Select well-directed more rare events to model them better.

Data cleansing

Data cleansing or data scrubbing refers to detecting / correcting / removing

- inaccurate,
- incorrect or
- incomplete

records from a data set.

Improve data quality

- Turn all characters into capital letters to level case sensitivity.
- Remove spaces and nonprinting characters.
- Fix the format of numbers, date and time (including decimal point).
- Split fields that carry mixed information into two separate attributes, e.g. "Chocolate, 100g" into "Chocolate" and "100.0". This is known as field overloading.
- Use spell-checker or stemming to normalize spelling in free text entries.
- Replace abbreviations by their long form (with the help of a dictionary).

Improve data quality

- Normalize the writing of adresses and names, possibly ignoring the order of title, surname, forename, etc. to ease their re-identification
- Convert numerical values into standard units, especially if data from different sources (and different countries) are used.
- Use dictionaries containing all possible values of an attribute, if available, to assure that all values comply with the domain knowledge.

Missing value

- **Ignorance/Deletion. Delete the whole record.**
- Imputation. The missing values may be replaced by some estimate.(The mean, the median or the mode of the attribute.)
- Explicit value. Use a specific value as missing for the model. (e.g. -1 when only positive numbers are in the domain)

Missing value

- Ignorance/Deletion. Delete the whole record.
- **Imputation. The missing values may be replaced by some estimate.(The mean, the median or the mode of the attribute.)**
- Explicit value. Use a specific value as missing for the model. (e.g. -1 when only positive numbers are in the domain)

Missing value

- Ignorance/Deletion. Delete the whole record.
- Imputation. The missing values may be replaced by some estimate.(The mean, the median or the mode of the attribute.)
- **Explicit value. Use a specific value as missing for the model. (e.g. -1 when only positive numbers are in the domain)**

Transformation of data

Some models can only handle numerical attributes, other models only categorical attributes.

Categorical $\Longrightarrow$ Numerical.

- **Binary attribute : numerical attribute with the values 0 and 1.**
- Ordinal attribute ("sortable"): enumerate in the correct order $1, \ldots, k$
- Categorical attribute(not ordinal) with more than two values, say $a_1, \ldots, a_k$,
  should not be turned into a single numerical attribute
  should be turned into k attributes $A_1, \ldots, A_k$ with values 0 and 1. a_i is represented by $A_i = 1$ and $A_j = 0$ for $i \neq j$.

Transformation of data

Some models can only handle numerical attributes, other models only categorical attributes.

Categorical $\Longrightarrow$ Numerical.

- Binary attribute : numerical attribute with the values 0 and 1.
- **Ordinal attribute ("sortable"): enumerate in the correct order** $1, \ldots, k$
- Categorical attribute(not ordinal) with more than two values, say $a_1, \ldots, a_k$,
  should not be turned into a single numerical attribute
  should be turned into k attributes $A_1, \ldots, A_k$ with values 0 and 1. a_i is represented by $A_i = 1$ and $A_j = 0$ for $i \neq j$.

Transformation of data

Some models can only handle numerical attributes, other models only categorical attributes.

Categorical $\Longrightarrow$ Numerical.

- Binary attribute : numerical attribute with the values 0 and 1.
- Ordinal attribute ("sortable"): enumerate in the correct order $1, \ldots, k$
- **Categorical attribute(not ordinal) with more than two values, say $a_1, \ldots, a_k$,**
  should not be turned into a single numerical attribute
  should be turned into k attributes $A_1, \ldots, A_k$ with values 0 and 1. a_i is represented by $A_i = 1$ and $A_j = 0$ for $i \neq j$.

Transformation of data: Discretization techniques

Splitting a numerical range into a number of bins.
Numerical $\Longrightarrow$ Categorical.

- **Equi-width discretization.** Splits the range into intervals (bins) of the same length.
- **Equi-frequency discretization.** Splits the range into intervals such that each interval (bin) contains (roughly) the same number of records.
- **V-optimal discretization.** Minimizes $\sum_i n_i V_i$ where n_i is the number of data objects in the ith interval and V_i is the sample variance of the data in this interval.
- **Minimal entropy discretization.** Minimizes the entropy. (Only applicable in the case of classification problems.)

Transformation of data: Discretization techniques

Splitting a numerical range into a number of bins.
Numerical $\Longrightarrow$ Categorical.

- **Equi-width discretization.** Splits the range into intervals (bins) of the same length.
- **Equi-frequency discretization.** Splits the range into intervals such that each interval (bin) contains (roughly) the same number of records.
- **V-optimal discretization.** Minimizes $\sum_i n_i V_i$ where n_i is the number of data objects in the ith interval and V_i is the sample variance of the data in this interval.
- **Minimal entropy discretization.** Minimizes the entropy. (Only applicable in the case of classification problems.)

Transformation of data: Discretization techniques

Splitting a numerical range into a number of bins.
Numerical $\Longrightarrow$ Categorical.

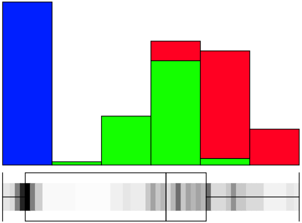
- **Equi-width discretization.** Splits the range into intervals (bins) of the same length.
- **Equi-frequency discretization.** Splits the range into intervals such that each interval (bin) contains (roughly) the same number of records.
- **V-optimal discretization.** Minimizes $\sum_i n_i V_i$ where n_i is the number of data objects in the ith interval and V_i is the sample variance of the data in this interval.
- **Minimal entropy discretization.** Minimizes the entropy. (Only applicable in the case of classification problems.)

Transformation of data: Discretization techniques

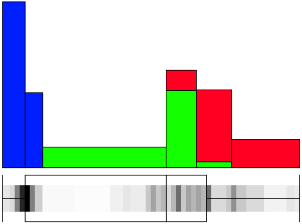
Splitting a numerical range into a number of bins.
Numerical $\Longrightarrow$ Categorical.

- **Equi-width discretization.** Splits the range into intervals (bins) of the same length.
- **Equi-frequency discretization.** Splits the range into intervals such that each interval (bin) contains (roughly) the same number of records.
- **V-optimal discretization.** Minimizes $\sum_i n_i V_i$ where n_i is the number of data objects in the ith interval and V_i is the sample variance of the data in this interval.
- **Minimal entropy discretization.** Minimizes the entropy. (Only applicable in the case of classification problems.)

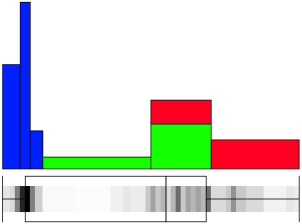
Transformation of data: Discretization



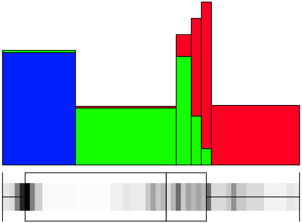
Equi-width



Equi-frequency



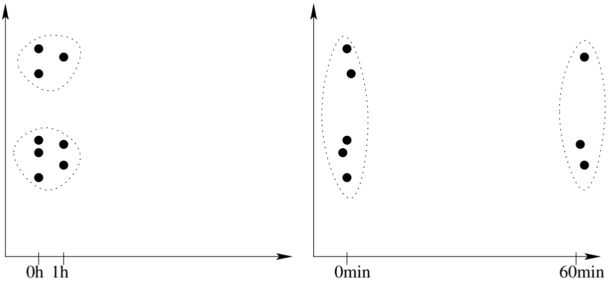
V-optimal



Minimal entropy

Normalisation/Standardisation

For some data analysis techniques (e.g. PCA, MDS; cluster analysis) the influence of an attribute depends on the scale or measurement unit.



To guarantee impartiality, some kind of standardisation or normalisation should be applied.

Normalization/Standardization

For a numerical attribute X:

min-max normalization.

$$n : \mathsf{dom}(X) \to [0,1], \qquad x \mapsto \frac{x - \min_X}{\max_X - \min_X}$$

z-score standardization. sample mean : $\hat{\mu}_X$ and empirical standard deviation: $\hat{\sigma}_X$

$$s : \mathsf{dom}(X) \to \mathbb{R}, \qquad x \mapsto \frac{x - \hat{\mu}_X}{\hat{\sigma}_X}$$

decimal scaling. s is the smallest integer value larger than $\log_{10}(\max_X)$

$$d : \mathsf{dom}(X) \to [0,1], \qquad x \mapsto \frac{x}{10^s}$$