Machine Learning Notes

by Mattia Orlandi

1. The Data

Issues:

- Never perfect (missing, inconsistent, duplicated, wrong).
- Outliers (small amount of data which are different from the rest due to anomalies).

Solution:

- Improve data quality using pre-processing activities to ease mining activities.
- Use mining techniques robust w.r.t. errors.
- Better data quality $\Rightarrow$ better results.

1.1. Data Types

Data type	Description	Descriptive statistics allowed	Transformation	Numerosity
Categorical - Nominal	Values are a set of labels, it's possible to distinguish one label from another ('$=$', '$\neq$' operators)	Mode, entropy, contingency, correlation, χ^2 test	One-to-one correspondence	Discrete, possibly binary
Categorical - Ordinal	As above, plus total ordering ('$<$', '$>$', '$\leq$', '$\geq$' operators)	As above, plus median, percentiles, rank correlation	Order-preserving transformation: $new \leftarrow f(old)$, f monotonic	Discrete, possibly binary

Data type	Description	Descriptive statistics allowed	Transformation	Numerosity
Numerical - Interval	As above, plus difference is meaningful ('$+$', '$-$' operators)	As above, plus average, standard deviation, Pearson's correlation, F and t tests	Linear functions: $new \leftarrow a + b * old$	Continuous, possibly approximated
Numerical - Ratio	As above, plus all mathematic operations on numbers, univocal definition of 0	As above, plus geometric mean, harmonic mean, percentage variation	Any mathematical function, *standardization*, variation in percentage	Continuous, possibly approximated

Obs.: - Asymmetric attributes are attributes in which only the presence is considered important (non-null values).
- Binary asymmetric attributes are relevant in the discovery of association rules.

General characteristics of datasets:

- **Dimensionality**: the size of the dataset (also qualitative).
- **Sparsity**: the number of zeroes or nulls.
- **Resolution**: the degree of detail in which the analysis is performed:
  - if it's too detailed it may be affected by noise;
  - if it's too general it can hide interesting patterns.

Obs.: when a piece of information is missing, storing zero of some special values is a bad habit.

Data representation:

- Relational tables: same set of attributes for all the records.
- Data matrix:
  - each row is a point in a vector space:
  - numeric attributes;
  - N rows × D dimensions (attributes, columns, properties).
- Document:
  - each row represents a document;
  - each column represents a term;

    - each cell contains the absolute frequency of the term in the document (sequence is lost).
- Transactional: each record contains a set of objects.
- Graph data: set of nodes and (oriented) arcs.
- Ordered data: sequence of objects.

1.2. Data Quality

Noise: modification of original values.
Outliers: data whose characteristics are considerably different from most of the data in the dataset; can be generated by noise or by rare events.
Missing values: due to data not collected or inapplicable information; their management varies according to the context:
  - do not consider objects with missing values;
  - estimate missing values;
  - provide for default values;
  - insert all possible values weighted with their probabilities.
Duplicate data: data objects that are duplicated or almost duplicated, for instance due to merging data from different sources; data must be cleaned.

1.3. Data pre-processing

Aggregation

Combining two or more attributes/objects into a single one.

Purposes:

- data reduction;
- change of scale;
- more stable data.

Sampling

- Preliminary investigation and final data analysis.
- From the statistician perspective, obtaining the entire dataset could be impossible/too expensive.
- From the data processing perspective, processing the entire dataset could be too expensive/time consuming.
- Thus, using a sample will work almost as well as using the entire dataset, *if the sample is representative*.

Types:

1. Simple random:
    - a single random choice of an object with given probability distribution.
2. With replacement:

    - repetition of independent extractions of type 1
3. Without replacement:
    - repetition of extractions, in which the extracted element is removed from the population.
4. Stratified:
    - split data into several partitions according to some criteria, then draw the random samples from each partition;
    - used when the dataset is split into subsets with homogeneous characteristics;
    - representativity is guaranteed inside each subset.

Sample size:

- Statistics provides techniques to assess the *optimal sample size*, and the *sample significativity*.
- Sampling is a trade-off between data reduction and precision.
- If it’s too small ⇒ loss of information.

Sampling with/without replacement:

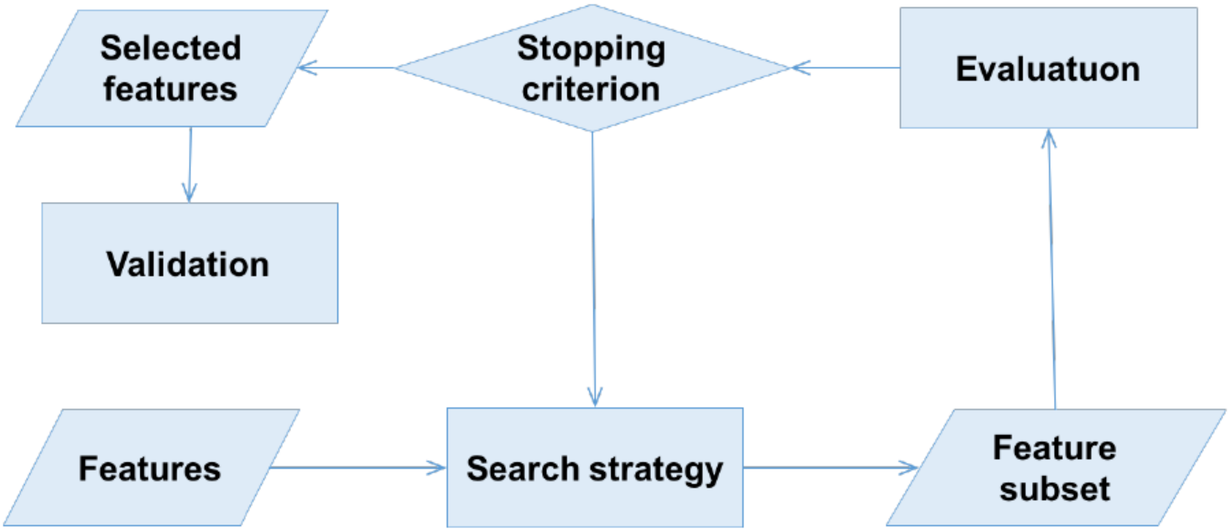
- Nearly equivalent if sample size is a small fraction of dataset size.
- Sampling with replacement, in a small population, could lead to an underestimate of small subsets.
- Sampling with replacement is easier to implement and to be interpreted (extractions are independent).
- Missing class: the probability of sampling at least one element for each class (with replacement) is independent from the dataset size.

Dimensionality reduction

- *Curse of dimensionality*:
    - when dimensionality is very high the occupation of space becomes very sparse;
    - thus, discrimination on the basis of the distance becomes uneffective.
- Purposes:
    - avoid the *curse of dimensionality*;
    - noise reduction;
    - reduce time and memory complexity of mining algorithms;
    - visualization.
- Techniques:
    - principal component analysis (PCA);
    - singular values decomposition (SVD);
    - supervised techniques;
    - non-linear techniques.
- PCA:
    - Find the projections that capture most of the data variation, by computing the eigenvectors of the covariance matrix: those vectors define the new space.
    - The new dataset will have *only the attributes* which capture most of the data variation.

Feature subset selection

- A *local* way to reduce dimensionality:
  - Redundant attributes.
  - Irrelevant attributes.
- Approaches:
  1. **Brute force**: try all possible feature subsets as input to data mining algorithm and measure its effectiveness with reduced dataset.
  2. **Embedded approach**: feature selection occurs naturally as part of data mining algorithm (e.g. Decision Tree).
  3. **Filter approach**: features are selected before data mining algorithm is run.
  4. **Wrapper approach**: data mining algorithm chooses the best set of attributes.



Feature creation

New features can capture more efficiently data characteristics:

- Feature extraction (e.g. from pixels of a picture of a face to eye distance).
- Mapping to a new space (e.g. signal to frequencies using Fourier).
- New features (e.g. volume and weight to density).

Discretization and binarization

Sometimes it is better to work with distinct values, therefore discretization is applied:

- some algorithms work better with categorical data;
- a small number of distinct values can let pattern emerge more clearly.
- a small number of distinct values let the algorithms be less influenced by noise.

Discretization:

- Continuous ⇒ Discrete
  - thresholds
  - binarization (single threshold)
- Discrete ⇒ Discrete with less values
  - domain knowledge.

Attribute transformation

- The entire set of values is mapped to a new one, according to a function (in general, the distribution changes).
- Standardization: $x \to \frac{x-\mu}{\sigma}$
  - if the original values have a *gaussian* distribution, the transformed ones will have a *standard gaussian* distribution ($\mu = 0$, $\sigma = 1$);
  - translation and shrinking/stretching, no change in distribution.
- Normalization: the domains are mapped to standard ranges
  - e.g. $x \to \frac{x-x_{min}}{x_{max}-x_{min}}$ (0 to 1), $x \to \frac{x-\frac{x_{max}+x_{min}}{2}}{\frac{x_{max}-x_{min}}{2}}$ (-1 to 1)
  - translation and shrinking/stretching, no change in distribution.

1.4. Similarity and dissimilarity

- Similarity:
  - Numerical measure of how alike two data objects are.
  - Higher when objects are more alike.
  - Often falls in range [0,1].
- Dissimilarity:
  - Numerical measure of how different two data objects are.
  - Lower when objects are more alike.
  - Minimum is often 0, upper limit varies.
- Proximity refers to similarity or dissimilarity.

Similarity and Dissimilarity by attribute type

Given p, q values of an attribute for two data objects:

Attribute type	Similarity	Dissimilarity
Nominal	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$
Ordinal (integers $\in [0, V-1]$)	$s = 1 - \frac{\lvert p-q \rvert}{V-1}$	$d = \frac{\lvert p-q \rvert}{V-1}$
Interval or Ratio	$s = \frac{1}{1+d}$ or $s = 1 - \frac{d-\min(d)}{\max(d)-\min(d)}$	$d = \lvert p-q \rvert$

Euclidean distance - L_2

$$dist = \sqrt{\sum_{d=1}^{D} (p_d - q_d)^2}$$

- D is the number of dimensions (attributes) and p_d, q_d are the d-th attributes (components) of data objects p and q, respectively.
- Standardization/normalization is necessary if scales differ.

Minkowsky distance - L_r

$$dist = \left(\sum_{d=1}^{D} |p_d - q_d|^r \right)^{\frac{1}{r}}$$

- Same properties of Euclidean distance.
- The parameter r is chosen depending on the dataset and on the application:
  - $r = 1 \Rightarrow$ *city block* / *Manhattan* / L_1 norm:
    - best way to discriminate between *zero* distance and *near-zero* distance;
    - an ϵ change on any coordinate causes an ϵ change in the distance;
    - works better than the Euclidean norm in very high dimensional spaces.
  - $r = 2 \Rightarrow$ *Euclidean* / L_2 norm
  - $r = \infty \Rightarrow$ *Chebyshev* / *supremum* / L_{max} / L_∞ norm:
    - considers only the dimension where the difference is maximum;
    - provides a simplified evaluation, disregarding the dimensions with lower differences:

$$dist_\infty = \lim_{r \to \infty} \left(\sum_{d=1}^{D} |p_d - q_d|^r \right)^{\frac{1}{r}} = \max_d |p_d - q_d|$$

Mahalanobis distance

Given the covariance matrix of the data set:

$$\Sigma_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} (e_{ki} - \bar{e_i})(e_{kj} - \bar{e_j})$$

the Mahalanobis distance is defined as:

$$dist_m = \sqrt{(p-q)\Sigma^{-1}(p-q)^T}$$

It takes into account the direction of greater variation of data.

Properites of a distance

- **Positive definiteness**: $dist(p,q) \geq 0 \; \forall p,q$ and $dist(p,q) = 0$ iff $p = q$
- **Symmetry**: $dist(p,q) = dist(q,p)$
- **Triangle inequality**: $dist(p,q) \leq dist(p,r) + dist(r,q) \; \forall p,q,r$

A distance satisfying all the properties above is called **metric**.

Properites of a Similarity

- $sim(p, q) = 1$ iff $p = q$
- $sim(p, q) = sim(q, p)$

Between **binary vectors**:

- consider
  M_{00} : number of attributes where $p = 0$ and $q = 0$,
  M_{01} : number of attributes where $p = 0$ and $q = 1$,
  M_{10} : number of attributes where $p = 1$ and $q = 0$,
  M_{11} : number of attributes where $p = 1$ and $q = 1$.
- Simple Matching Coefficient

$$SMC = \frac{\text{number of matches}}{\text{number of attributes}} = \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}}$$

- Jaccard Coefficient

$$JC = \frac{\text{number of 11 matches}}{\text{number of non-both-zero attributes}} = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

Cosine similarity: cosine of the angle between two vectors

$$cos(p, q) = \frac{p \cdot q}{||p|| \cdot ||q||}$$

Tanimoto (extended Jaccard Coefficient): variation of Jaccard for continuous or count attributes

$$T(p, q) = \frac{p \cdot q}{||p||^2 + ||q||^2 - p \cdot q}$$

The right proximity measure depends on data:

- Dense, continuous ⇒ **metric** measure, i.e. Euclidean distance.
- Sparse, asymmetric ⇒ cosine, Jaccard, Tanimoto.

Correlation

Measure of the linear relationship between a pair of attributes:

- Standardize the values.
- For two given attributes p, q, consider as vectors the ordered lists of the values over all the data records.
- Compute their dot product.

$$\mathbf{p} = [\, p_1, \ldots, p_N \,] \Rightarrow^{\text{standardize}} \mathbf{p}'$$
$$\mathbf{q} = [\, q_1, \ldots, q_N \,] \Rightarrow^{\text{standardize}} \mathbf{q}'$$
$$corr(p, q) = \mathbf{p}' \cdot \mathbf{q}'$$

- Independent variables ⇒ correlation is zero.
- Correlation is zero ⇒ absence of *linear relationship* between variables.
- Positive values ⇒ positive linear relationships.

Between nominal attributes: **Symmetric Uncertainty**

$$U(p, q) = 2\frac{H(p) + H(q) - H(p, q)}{H(p) + H(q)} \in [0, 1]$$

where $H(\cdot)$ is the entropy of a single attribute, while $H(\cdot, \cdot)$ is the joint entropy (computed from the joint probabilities).