

Fundamentals of Machine Learning

Data Transformation

Jan Platos

December 8, 2021

Data Transformation

Data Transformation

- Data transformation techniques include methods, that may help to transform data before any machine learning algorithm is applied.
- The goal is to remove the unnecessary data and highlight the most important aspects of the data to process.
- The techniques includes:
 - Attribute selection
 - Attribute discretization
 - Data projections
 - Sampling
 - Data cleansing
 - Converting multi-class problems to two-class ones.

Data Transformation- Attribute selection

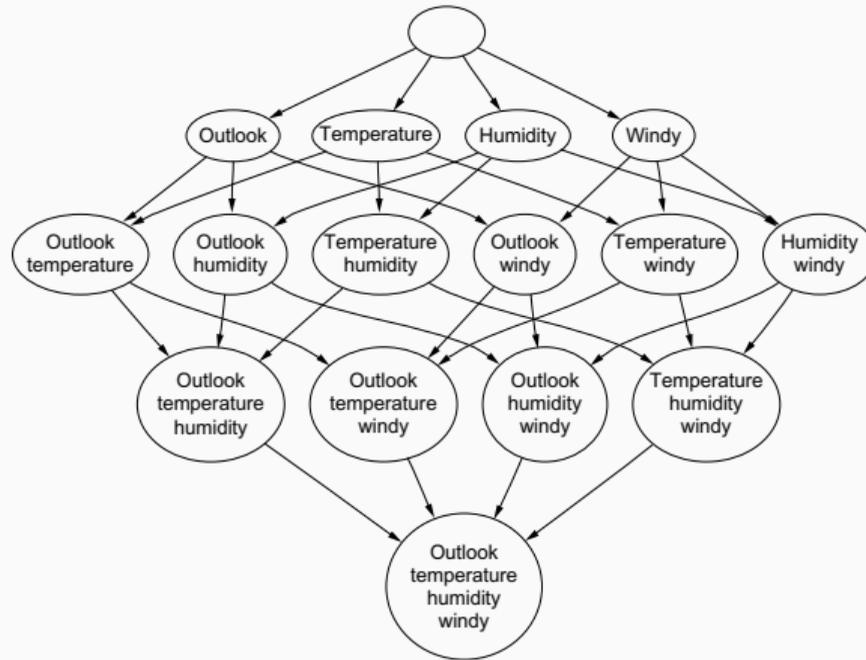
- Most machine learning algorithms are designed to learn which are the most appropriate attributes to use for making their decisions.
- In practice, even a random variable may decrease the performance of the classifier (chaotic behavior in deeper levels).
- Moreover, even a relevant attribute may decrease the performance (fragmentation due to splitting near the root).

- Scheme-independent selection, also called a filter method prepares the dataset before a machine learning algorithm is applied.
- Selection of the *Relevance* measure universally is impossible.
- The first possibility is to select just enough attribute to optimally cover the required classes.
- Another possibility is to use model that select the useful attributes, e.g. Decision Tree, before a target algorithm is used.

- Described algorithm may fail when two almost identical attributes exists - both are rejected or accepted.
- Another way of eliminating redundant attributes as well as irrelevant ones is to select a subset of attributes that individually correlate well with the class but have little inter-correlation
- One possibility is *symmetric uncertainty*.
- Subset may be generated using exhaustive search.

- Greedy search for optimal subset of features may be used with evaluation using the mentioned metric.
- There are two possible direction of the greedy algorithm:
 - Forward selection
 - Backward elimination
- Each approach have its start and stopping criterion.

Data Transformation - Attribute selection- Scheme-independent



- The selected set of attributes is evaluated using the concrete Machine Learning model.
- The cross-validation, holdout or bootstrap estimation may be used to precisely predict the efficiency.
- Greedy algorithm is again used.
- Backward elimination produces larger set of attributes.

Data Transformation- Discretizing of Numeric Attributes

- Numeric attributes are problematic for many algorithms.
- Even the algorithms that accept numeric attributes works faster and/or more effective.
- Main reason is the assumption of the probability distribution about the numeric attribute.
- Discretization may be unsupervised or supervised (with respect to the classes).

Data Transformation- Discretizing of Numeric Attributes

- Unsupervised discretization distribute instances according the valud into predefined bins.
- Equal-interval binning simply divide the full range into equal sized bins.
- Equal-frequency binning distribute instances to the bins with equal frequency based on the histogram.

Data Transformation- Discretizing of Numeric Attributes

- Supervised discretization distribute instances according the value and classes.
- Entropy-based discretization follow the approach used in Decision Tree.
- The bins are generated dynamically based on the split criterion until intervals with the same class remains.
- The split point is always between different class labels not the same, which may be used in optimization.
- The stopping criterion may be confused by multiple class values.

Data Transformation- Projection

- Projection is a mapping that transform data in a some way.
- Data often need application of some mathematical algorithms, e.g. difference between two dates (age),
- Other transformation may involve general knowledge, e.g. holidays, day in a week, chemical atomic numbers, ...
- Clustering may be another type of transformation that produces a new attribute.
- Special kind of transformation is able to map data into lower dimension.

Data Transformation- Projection- Principal Component Analysis

- Principal Component Analysis (PCA) is a linear algebra transformation that is able to map data into a space with lower dimension.
- Data with k numeric attributes may be visualized as a points in k -dimension space.
- The axis used in visualization are based on the used projection.
- The axis may be computed in completely different way, but we prefer orthogonal axis (i.e., each axis is at right angles to the others).
- The variance (distribution of points) along the axis may be easily computed.

Data Transformation- Projection- Principal Component Analysis

- Principal Component Analysis places the first axis in the direction of the largest variance.
- The second axis is perpendicular to the first and maximized the remaining variance.
- The computation procedure requires a covariance matrix and decomposition into so called *eigen-vectors* and *eigen-values*.
- This process is called eigen-decomposition and it is possible in general for any square matrix.

$$A = P\lambda P^{-1}$$

- Other approaches includes:
 - Random Projection - generate a random mapping into lower dimension.
 - Singular Value Decomposition - More general version of PCA.
 - Independent Component Analysis - decomposed data into a statistically independent parts.
 - Fisher's Linear Discriminant analysis - includes a class labels into decomposition.

Data Transformation- Sampling

- Sampling enables processing of large amount of data by a standard algorithms.
- Sampling may choose the same instance several times (with replacement) or only once (without replacement).
- Reservoir sampling the sampling solution for data stream where data are not stored locally.
- The reservoir of size r is used as a cache for sampled instances.
- The i -th instance has the chance r/i to be placed into a reservoir.

Data Transformation- Cleansing

- Data are in general noisy and corrupted.
- The manual check of the data is hard and expensive, sometimes even impossible.
- Removing misclassified instances from the dataset may improve the precision and reduce the size of the decision tree model. (these instances are probably corrupted or a noise)
- Moreover, training model on clean data leads to worse performance on noisy test data.

Data Transformation- Cleansing - Anomaly detection

- Anomalies are not exactly the same as noise data, it represents an error or incorrect data.
- Extreme data may be identified using exploratory analysis.
- Some methods combines several classifiers to decide about anomaly of the data.
- Erroneous data are usually misclassified by some of the classifier.

Data Transformation- Multiple Classes into Binary Classification

- Some algorithm does not support multiple class classification, neural networks, logistic regression, etc.
- Multi-class variants may be developed but complicate the procedure.
- Clever selected strategies may be used instead.
- Decomposition of class into binary problems simplifies the solution.
- Basic variants are *one-vs-rest* and *one-vs-one*.

Data Transformation- Multiple Classes into Binary Classification

- *one-vs-rest*
 - A new dataset is generated for each class with two classes only.
 - On each dataset a classifier is built.
 - Final decision is made based on the partial results and/or confidence of models.
- *one-vs-one*
 - Each pair of classes create a new dataset.
 - On each dataset a classifier is built.
 - Majority voting is used for final decision.
 - $k(k - 1)/2$ classifiers is constructed.

Questions