Data Preparation

Irene Siragusa, PhD

CHILAB
Università degli Studi di Palermo

Outline

- Feature extraction
- Data type portability
- Data cleaning
- Data reduction

Feature Extraction

Raw data is often in a form that is not suitable for processing:

- Derive meaningful features from the data.
  Features with good semantic interpretability are more desirable.

- Data integration from multiple sources and data type portability, where low-level features of one type may be transformed to higher-level features of another type.

Feature Extraction

Domain	Raw Data	Features
Sensor	Low-level signals	Wavelet or Fourier transforms
Image	Pixels	Color histograms, Visual words
Web logs	Text strings	IP address, Action
Network traffic	Characteristics of the network packets	Number of bytes transferred, Network protocol
Document data	Text strings	Bag-of-words, Entity extraction

Data type portability

- Data is often heterogeneous
  - A demographic data set may contain both numeric and mixed attributes
- Possible solutions
  - Designing an algorithm with an arbitrary combination of data types
    - Time-consuming and sometimes impractical
  - Converting between various data types
    - Utilize off-the-shelf tools for processing

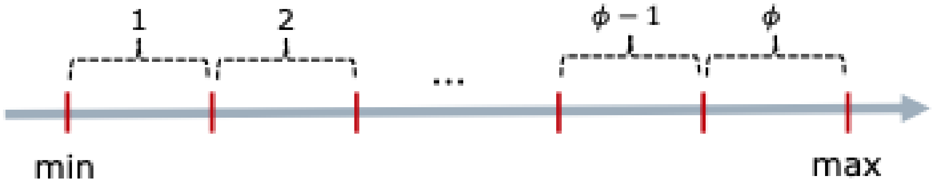
Data type portability

Source data type	Destination data type	Methods
Numeric	Categorical	Discretization
Categorical	Numeric	Binarization
Text	Numeric	Latent semantic analysis (*LSA*)
Time series	Discrete sequence	*SAX*
Time series	Numeric multidimensional	*DWT*, *DFT*
Discrete sequence	Numeric multidimensional	*DWT*, *DFT*
Spatial	Numeric multidimensional	2-d *DWT*
Graphs	Numeric multidimensional	*MDS*, spectral
Any type	Graphs	Similarity graph (Restricted applicability)

Numeric -> Categorical

- Discretization

  - Divides the ranges of the numeric attribute into ϕ ranges



  - Age attribute
    - ✓ [0,10], [11,20], [21,30], …
  - Salary
    - ✗ [0, 10000], [10001, 20000], [20001, 30000], ...

Numeric -> Categorical

- Discretization
  - Equi-width Ranges
    - Each range $[a, b]$ is chosen such that $b - a$ is a constant
  - Equi-log Ranges
    - Each range $[a, b]$ is chosen such that $\log b - \log a$ is a constant
    - For example, $[1, a], [a, a^2], [a^2, a^3], \ldots$
    - In general, $[a, b] \rightarrow f(b) - f(a)$ for a chosen $f(\cdot)$
  - Equi-depth Ranges
    - Each range has an equal number of records

Categorical -> Numeric

- Binarization
  - Two categories
    - $[0,1]$ or $[-1,1]$ as possible values
  - ϕ categories
    - ϕ-dimensional indicator vector
    - The position 1 of indicates the category
    - One-hot encoding

$$\begin{array}{cl} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} & \begin{array}{l} \longrightarrow 1^{st}\ \text{Category} \\ \longrightarrow 2^{nd}\ \text{Category} \\ \longrightarrow 3^{rd}\ \text{Category} \end{array} \end{array}$$

$$\phi = 3$$

Text -> numeric

- Tokenization
- Stop word Removal
- Stemming
- Term frequency - Inverse Document Frequency (TF-IDF)
- Dimensionality reduction via Latent Semantic Analysis (LSA)
- Normalization

$$\begin{array}{rc} & \begin{matrix} the & an & \end{matrix} \\ \begin{matrix} The\ cat\ on\ the\ table \\ An\ apple \\ \dots \end{matrix} & \begin{bmatrix} 2 & 0 & \dots \\ 0 & 1 & \dots \\ \dots & \dots & \dots \end{bmatrix} \end{array}$$

Text -> numeric

- Term frequency - Inverse Document Frequency (TF-IDF)
  - Numerical statistic that reflects the significance of a word within a document relative to a collection of documents (corpus)
  - Quantify the importance of a term in a document with respect to its frequency in the document and its rarity across multiple documents.

Text -> numeric

- Term frequency - Inverse Document Frequency (TF-IDF)

$$TF(t,d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d}$$

$$IDF(t,D) = \log \frac{Number\ documents\ in\ the\ corpus}{Number\ of\ documents\ containing\ term\ t}$$

$$TF-IDF(t,d,D) = TF(t,d)\ \times\ IDF(t,D)$$

Time series -> Discrete sequences

- Symbolic Aggregate Approximation (SAX)
  - Is an equi-depth discretization approach after window-based averaging.

1. Window-based averaging
   - Evaluate the average value in each windows
2. Value-based discretization
   - Discretize the average value by equi-depth intervals constructed by assuming that the time-series values are distributed with a Gaussian assumption.
   - The idea is to ensure that each symbol has an approximately equal frequency in the time series.

Time series -> Numeric

- Once a time series is converted as numeric data:
  - Can be processed with algorithms for multidimensional data
  - Have a reduced dimensionality

- Discrete Wavelet Transform (DWT)

- Discrete Fourier transform (DFT)

Discrete sequences -> Numeric

1. Convert the discrete sequence to a set of binary time series.

ACACACTGTGACTG	(4 Symbols)
10101000001000	(A)
01010100000100	(C)
00000001010001	(T)
00000010100010	(G)

2. Wavelet transformation over each of these time series.
3. Features from the different series are combined to create a single multidimensional record.

Spatial -> Numeric

- Similar to time series, but with 2-dimentional contextual attributes
- Obtained data can be processed with multi-dimensional algorithms
- 2d-Discrete Wavelet Transform (DWT)

Graph -> Numeric

- For graph whose edges are weighted and represent similarity or distance relationships between nodes.

- Multi-Dimensional Scaling (MDS)
  - Edge represents distances

- Spectral transformations
  - Edge represents similarity

Any type -> Graph

- Useful for applications based on the notion of similarity

- Building a neighborhood graph
  - Each object in the dataset is considered as a node O_i
  - If $d(O_i, O_j) < \varepsilon$, an edge is built with weight obtained via kernel functions (e.g. heat kernel) application expressing similarity

$$w_{ij} = e^{\frac{d(o_i, o_j)^2}{t^2}}$$

Data Cleaning

- Needed process due to errors associated with the data collection process

- Why is needed?
  - Troubles in data collection technologies (sensor, scan)
  - Privacy reasons
  - Manual errors
  - Costly data collection

Handling missing entries

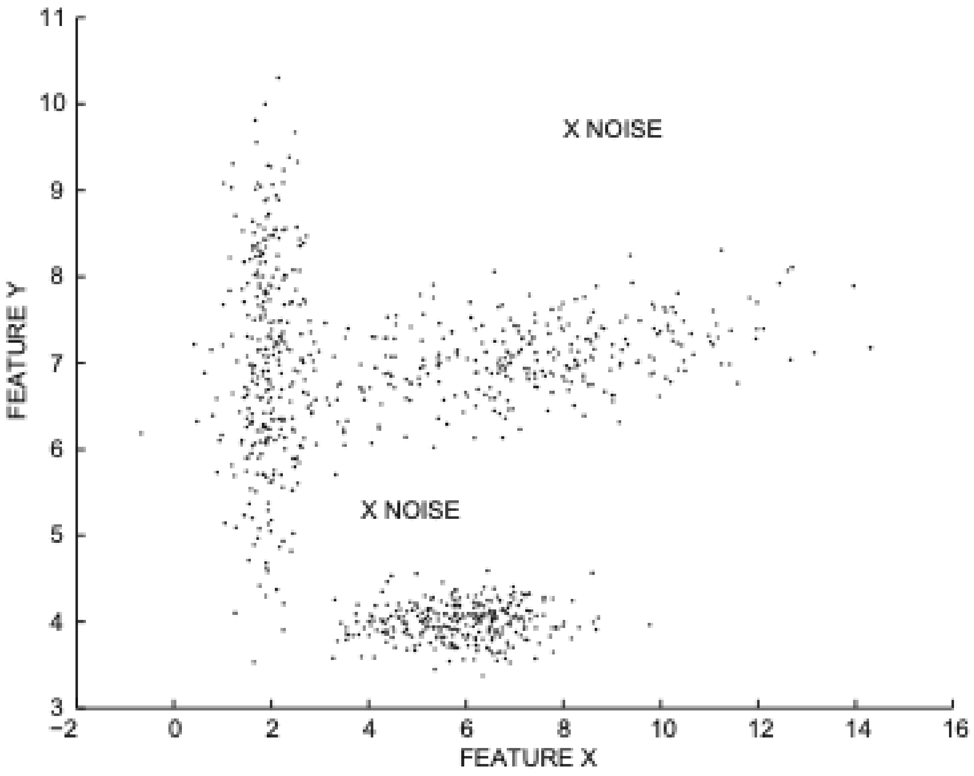
- Drop records with missing entries

- Estimate or impute missing values
  - Collaborative filtering to estimate missing values relying on similar records according to some similarity function

- Data mining methods are inherently designed to work robustly with missing values, thus it is possible to work with missing data.

Handling Incorrect and Inconsistent Entries

- Inconsistency detection
  - Same data stored in different formats
  - e.g. I. Siragusa, Irene Siragusa, Siragusa Irene
- Domain knowledge
  - Inconsistencies may be detected with domain-base knowledge
- Data-centric methods
  - Data from statistical perspective
  - Noise data vs Outliers



Scaling and normalization

- Related to the scale and ranging of the data
  - Data-field with larger magnitude biases low-magnitude data
- Standardization

$$z_i^j = \frac{x_i^j - \mu_j}{\sigma_j}$$

- Normalization/min-max scaling

$$z_i^j = \frac{x_i^j - min_j}{max_j - min_j}$$

Data reduction

- ☑ Reduce space complexity
- ☑ Reduce time complexity
- ☑ Reduce noise
- ☑ Reveal hidden structures
- ✗ Information loss

Sampling

- Simple, intuitive, and relatively easy to implement
- Type of sampling used may vary with the application at hand
- Sampling over
  - Static data
  - Data Streams

Sampling for static data

- Static data
  - Data for which we have the entire dataset available
  - Two available techniques
    - Unbiased sampling
    - Biased Sampling

- Unbiased sampling
  - Technique of uniform sampling of f data points from a static data set $\mathcal{D}$ with n records

Sampling for static data

- Sampling w/o replacement
  - $n \cdot f$ records are randomly picked from $\mathcal{D}$
- Sampling w/ replacement
  - records are sampled sequentially and independently from the entire data set $\mathcal{D}$, $n \cdot f$ times

Sampling for static data

- Biased sampling
  - Some partes of the data are intentionally emphasized since they have a greater importance according to a probability distribution (the best the actual one)
  - e.g. temporal-decay bias, more recent records are preferred

$$p(\bar{X}) \propto e^{-\lambda \cdot \delta t}$$

- Stratified sampling
  - Needed when important parts of the data may not be sufficiently represented by sampling because of their rarity
    - 1. partition the data into desired sets
    - 2. independent sampling

Rensevoir Sampling for Data Streams

- In data streams, data changes dynamically since new data arrive sequentially
- We want to keep a sample of k points from a data stream
  1. first k points are maintained
  2. For the k+1 point and subsequent ones
     - 2.1 Pick the new point with a probability k/n (n is increasing)
     - 2.2 If a new sample is picked, drop one of the existing data points
- After n stream points have arrived, the probability of any stream point being included in the reservoir is the same and equal to k/n.

Sampling

- We will see that a cured data set of dimensionality $\mathcal{D}$ can be further under-sampled or over-sampled
  - Under-sampled
    - Samples in numerous classes are reduced
  - Over-sampled
    - Samples in scarce classes are artificially enriched (data augmentation)
- This will be needed for classification purposes in which highly unbalanced classes occur

Feature subset selection

- Features that are known to be irrelevant can be discarded
- Unsupervised feature selection
  - removal of noisy and redundant attributes from the data
  - best defined in terms of its impact on clustering applications
- Supervised feature selection
  - relevant to the problem of data classification
  - only the features that can predict the class attribute effectively are the most relevant