Machine Learning and Data Mining

Data Preprocessing

Albert Bifet(@abifet)

TELECOM ParisTech
université PARIS-SACLAY

Data Basics

Machine Learning/Data Mining Applications

- Business Analytics
  - Is this costumer credit-worthy?
  - Is a costumer willing to respond to an email?
  - Do costumers divide in similar groups?
  - How much a costumer is going to spend next semester?
- World Wide Web
- Financial Analytics
- Internet of Things
- Image Recognition, Speech
- ..

The Data Mining Process

- Data collection
- Data Preprocesing
  - Feature extraction
  - Data cleaning
  - Feature selection and transformation
- Analytical processing and algorithms
- Data Postprocesing

Multidimensional Data

- Example:

Competitor Name	Swim	Cycle	Run	Total
John T	13:04	24:15	18:34	55:53
Norman P	8:00	22:45	23:02	53:47
Alex K	14:00	28:00	n/a	n/a
Sarah H	9:22	21:10	24:03	54:35

Table: Triathlon results

- Example or Instance
  - data point, transaction, entity, tuple, object, or feature-vector
- Attribute or Feature
  - field, dimension

Instance Types

- Dense
  - red, white, Barcelona, 3, up
  - red, red, Barcelona, 4, down
  - black, white, Paris, 2, up
  - red, green, Paris, 3, down
- Sparse
  - 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
  - 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
  - 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
  - 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
  - 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
  - 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0

Attribute Type

- Numerical
  - 0, 1, 3.43, 2.34, 4.23
- Categorical or Discrete
  - +, -
  - red, green, black
  - yes, no
  - up, down
  - Barcelona, Paris, London, New York
- Text Data: vector-space representation
  - The cat is black
- Binary: Categorical or Numerical

Analytical processing and algorithms

- Attribute/Column Relationships
  - **Classification** : predict value of a discrete attribute
  - **Regression**: predict value of a numeric attribute
- Instance/Row Relationships
  - **Clustering**: determine subsets of rows, in which the values in the corresponding columns are similar
  - **Outlier Detection**: determine the rows that are very different from the other rows

Big Data Scalability

- Distributed Systems:
  - Hardware: Hadoop cluster
  - Software: MapReduce, Spark, Flink, Storm
- Streaming Algorithms
  - Single pass over the data
  - Concept Drift

Data Preparation

The Data Mining Process

- Data collection
- Data Preprocesing
  - Feature extraction
  - Data cleaning
  - Feature selection and transformation
- Analytical processing and algorithms
- Data Postprocesing

Feature Extraction

- Sensor data: wavelets or Fourier Transforms
- Image Data: histograms or visual words
- Web logs: multidimensional data
- Network traffic: specific features as network protocol, bytes transferred
- Text Data: remove stop words, stem data, multidimensional data

Feature Conversion

- Numeric to Discrete
  - Equi-width ranges
  - Equi-log ranges
  - Equi-depth ranges
- Discrete to Numeric
  - Binarization: one numeric attribute for each value
- Text to Numeric
  - remove stop words, stem data, tf-idf, multidimensional data
- Time Series to Discrete Sequence Data
  - SAX: equi-depth discretization after window-based averaging
- Time Series to Numeric Data
  - Discrete Wavelet Transform
  - Discrete Fourier Transform

Term Frequency-Inverse Document Frequency

- Term frequency
  - Boolean ”frequencies”
    - $tf(t,d) = 1$ if t occurs in d and 0 otherwise;
  - Logarithmically scaled frequency
    - $tf(t,d) = 1 + log f_{t,d}$, or zero if $f_{t,d}$ is zero;
  - Augmented frequency,

$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

- Inverse document frequency

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

- Term frequency-inverse document frequency

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

Data Cleaning

- Handling missing entries
  - Eliminate entries with a missing value
  - Estimate missing values
  - Algorithms can handle missing values
- Handling incorrect entries
  - Duplicate detection and inconsistency detection
  - Domain knowledge
  - Data-centric methods
- Scaling and normalization
  - Standardization: for instance *i*, attribute *j*:

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

  - Normalization:

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

Feature selection and transformation

- Sampling for Static Data
  - Sampling with Replacement
  - Sampling without Replacement: no duplicates
  - Biased Sampling
  - Stratified Sampling
- Reservoir Sampling for Data Streams
  - Given a data stream, choose k items with the same probability, storing only k elements in memory.

Reservoir Sampling

Reservoir Sampling

for every item i in the first k items of the stream
    do store item i in the reservoir
$n = k$
for every item i in the stream after the first k items of the stream
    do select a random number r between 1 and n
        if $r < k$
            then replace item r in the reservoir with item i
        $n = n + 1$

Figure: Algorithm Reservoir Sampling

Feature selection and transformation

- Feature Subset Selection
  - Supervised feature selection
  - Unsupervised feature selection
  - Biased Sampling
  - Stratified Sampling
- Dimensionality reduction with axis rotation
  - Principal Component Analysis
  - Singular Value Decomposition
  - Latent Semantic Analysis

Principal Component Analysis

- Goal: **Principal component analysis** computes the most meaningful basis to re-express a noisy, garbled data set. The hope is that this new basis will filter out the noise and reveal hidden dynamics

- Normalize Input Data
- Compute k orthonormal vectors to have a basis for the normalized data
- Sort these *principal components*
- Eliminate components with low variance

Principal Component Analysis

- Organize the data set X as an $m \times n$ matrix, where m is the number of features and n is the number of instances.
- Normalize Input Data: subtract off the mean for each instance x_i
- Calculate the SVD or the eigenvectors of the covariance
  - Find some orthonormal matrix P where $Y = PX$ such that

$$S_Y = \frac{1}{n-1} YY^T$$

    is diagonalized.
  - The rows of P are the principal components of X.
- Sort these *principal components*
- Eliminate components with low variance