Introduction to Artificial Intelligence

Data Mining & Machine Learning

Ing. Tomas Borovicka

Department of Theoretical Computer Science (KTI), Faculty of Information Technology (FIT)
Czech Technical University in Prague (CVUT)

BIE-ZUM, LS 2013/14, 7. lecture

https://edux.fit.cvut.cz/courses/BIE-ZUM/

Summary of Previous Lecture

- General problem Solving

Problem $\longrightarrow$ Model $\longrightarrow$ Language $\longrightarrow$ Solver $\longrightarrow$ Solution

- Planning vs Search

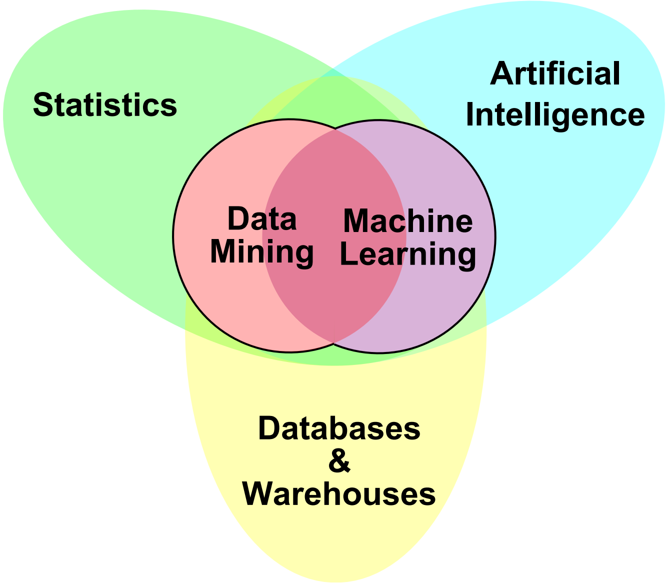
	Search	**Planning**
States	data structures	logical sentences
Actions	code	preconditions / effects
Goal	code	logical sentences
Plan	sequence of actions	constrains on actions

- Partial Order Planning

Data Mining & Machine Learning Applications

- Business
  - market basket analysis
  - direct marketing
  - recommender systems
  - predict credit rating
  - fraud detection
- Text and Web Mining
  - categorize documents / Web pages
  - classify E-mail (spam/ham)
  - identify Web usage patterns
- Intrusion detection
  - monitoring and analyzing user / system activities
  - analysis of abnormal activity patterns
- Risk analyses and quality control
  - forecasting
  - customer retention
  - quality control
  - competitive analysis
- Health care
  - disease and pathological patterns detection
  - decision support for effective treatments and best practices
  - pattern recognition from RTG, CT, MR . . .
- and many others . . .

Data Mining & Machine Learning



Data Mining & Machine Learning

- **Machine learning** is a study of design and development of generic algorithms that give computers the capability to learn.
- **Data mining** is defined as automatic or semiautomatic process of discovering patterns in data.
- Data mining utilizes machine learning algorithms to mine knowledge present in databases.
- It includes many fields:
  - database technologies
  - machine learning
  - artificial intelligence
  - statistics
  - information retrieval
  - knowledge-based systems
  - pattern recognition
  - neural networks
  - high-performance computing
  - data visualization

Taxonomy

- Unsupervised learning
  - Association rules
  - Clustering
- Supervised learning
  - Classification
  - Regression
- Semi-supervised learning
- Reinforcement learning

Unsupervised Learning

- Searching for interesting patterns or meaningful structures.
- Annotations or class labels of the data are unknown.
- Unsupervised learning helps to understand data.

Unsupervised Learning

Given a set of n examples $X = (x_1, \ldots, x_n)$ where $x_i \in \mathcal{X} \ \ \forall i \in [n] := \{1, \ldots, n\}$ are independently and identically distributed from unknown distribution $\mathcal{X}$ with density $P(x)$, the goal of unsupervised learning is to find interesting or meaningful structures in the data.

Supervised Learning

- The data (observations, measurements) are annotated.
- Learns from a set of labeled examples so that it can predict a label for any valid unseen examples.

Supervised Learning

Let $Y = (y_1, \ldots, y_n)$ be a set of targets where $y_i \in \mathcal{Y}$. Given a training set composed of pairs (x_i, y_i) independently and identically distributed from unknown distribution $\mathcal{X} \times \mathcal{Y}$ with density $P(x, y)$, the goal of supervised learning is to approximate an unknown function $f : \mathcal{X} \to \mathcal{Y}$.

Semi-Supervised Learning

- Labeling is expensive, annotation requires human interaction, time, may require experts, special devices etc.
- Unlabeled data is cheap and often can be obtained abundantly.
- Semi-supervised learning makes use of both labeled and unlabeled data.
- Utilizes unlabeled data and the underlying information.
- Yields greater performance than standard supervised techniques.

Semi-Supervised Learning

Let X_l be a set of examples for which we know the label y_i and let X_u be far bigger set of examples without known label y_i. Semi-supervised learning attempts to utilize unlabeled data in order to yield greater performance than standard supervised method if only labeled data are used.

Data Mining Process

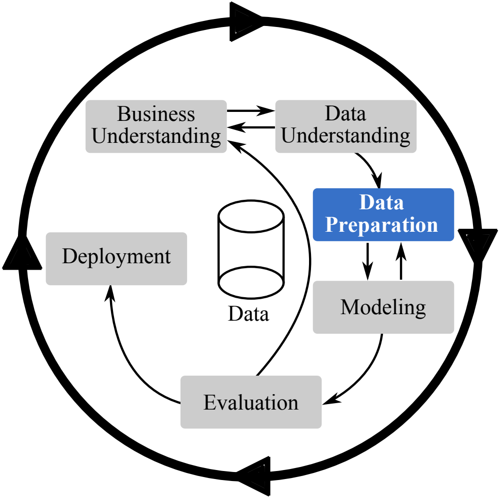


Figure : Data Mining Process by CRISP-DM

Business & Data Understanding

- For appropriate data preparation and successful modeling you need to understand the data.
- Descriptive statistics (mean, median, mode, quartiles, variance. . .).
- Visualization (histogram, box-plots, Q-Q plot, scatter).

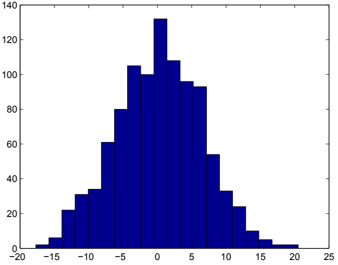


Figure : Histogram

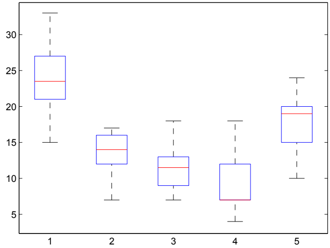


Figure : Box-plot

Data Preparation

- Preparing an input for machine learning algorithms.
- Critical part of data mining process.
- The preparation process may include:
  - Data Cleaning
    - Missing values
    - Noise filtering
  - Data Transformation
    - Normalization
    - Discretization
    - Aggregation
  - Data Reduction
    - Feature set extraction/selection
    - Numerosity reduction
    - Skew data balancing

Data Cleaning

- First step in data preparation.
- Real world data are typically
    - incomplete,
    - noisy,
    - inconsistent.
- Process of detecting and correcting (or removing) missing, corrupted or invalid values.

Missing values

- For many reasons we do not always have all the values in a feature vector.
- Some algorithms are unable to deal with missing values.
- Several options:
  - ignore values,
  - manually replace,
  - use global constant (average, median, zero),
  - use the most probable value.
  - ...
- Depends on meaning of the attribute and the value (data understanding).
- Do not bias meaning of the data!

Noise

- Noise is a random error or variance in a measured variable.
- Noise filters focus on instances that decrease prediction performance.
- Typical noise filter techniques are based on
  - Neighborhood smoothing - consider neighborhood of values.
  - Regression - fitting data with a function.
  - Clustering (prototyping)- values that fall out of cluster are considered as noise or outliers.

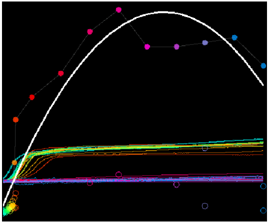


Figure : Function fitting

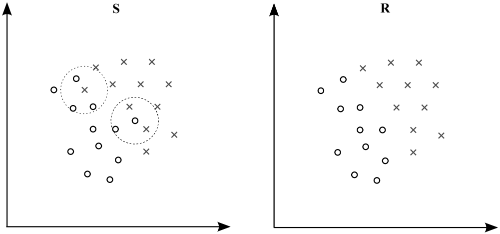


Figure : ENN (neighborhood based)

Data Transformation

- Transforms the data into a form appropriate for data mining algorithms.
- Many transformations are related to data reduction (aggregation, feature extraction etc.).

- **Normalization**
  - Transforming attributes by to some interval.
- **Discretization**
  - Categorizing, approximation.
- Other transformations
  - linear / nonlinear transformations, fourier transform, wavelet transform, . . .

Normalization

- For most of the learning algorithms it is necessary to normalize data in order to uniform attributes' weight.
- **min-max Normalization**

$$x' = \frac{x - min(X)}{max(X) - min(X)}$$

- **z-score Normalization**

$$x' = \frac{x - \mu(X)}{\sigma(X)}$$

- **Median Absolute Deviation Normalization**

$$x' = \frac{x - median(X)}{median(|x - median(X)|)}$$

- **Decimal Scaling**

$$x' = \frac{x}{10^n}, \ n = log_{10} max(X)$$

Discretization

- Reduces the number of values for a given continuous attribute into a small number of intervals.
- Simplifying and reducing the data by categorization.

- **Top-down discretization** - splitting
  - Binning (equal width, frequency)
  - Entropy based
  - Clustering
- **Bottom-up discretization** - merging
  - Interval merging by χ^2 analyses
  - Clustering (hierarchical)

$$a = \begin{cases} youth, & age \in \langle 0, 30) \\ middleage, & age \in \langle 30, 60) \\ senior, & age > 60 \end{cases}$$

Data Reduction

- Redundancy reduction and information extraction.
- Reduce amount of time and memory required by data mining algorithms.
- Most of data mining algorithms are not effective for high dimensional data.
  - Curse of dimensionality.

- **Aggregation**
  - Aggregating examples into a single object (average, max, deviation).
- **Feature set extraction / selection**
  - Reducing number of attributes (dimensionality).
- **Instance selection**
  - Reducing number of examples.

Feature Set Extraction/Selection

- **Redundant features**
  - Duplicate information in one or more attributes.
- **Irrelevant features**
  - Contain no useful information for the data mining task.

- **Feature extraction**
  - Principle Component Analysis - projection that captures the largest amount of variation in the data.
  - Singular Value Decomposition - transforming correlated variables into a set of uncorrelated.
  - Linear discriminant analysis finding the line that best separates two classes.
- **Feature selection**
  - Search (Greedy forward/backward selection, tabu search)
  - Correlation / mutual information
  - Performance optimization (GA)

Instance Selection (Reduction)

- **Redundant examples**
  - Redundant examples are usually useless for learning.
- Large data set may considerably slow down the learning process.
  - Time complexity is a function of data set size.

- **Wrapper methods** - The selection criterion is based on the predictive performance or the error of a model (instances that do not contribute to the predictive performance are discarded from the training set).
- **Filter methods** - The selection criterion is a function that is based on features of the instance vector (decision border, centroids, prototypes . . .).

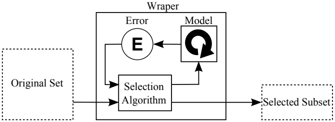


Figure : Wraper method

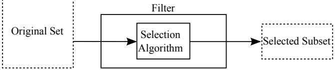
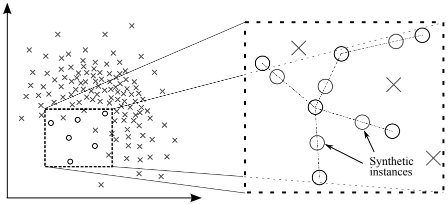


Figure : Filter method

Class Balancing

- A dataset is well-balanced, when all classes are represented with the same proportion.
- In practice many domains are characterized by a small proportion of positive instances and a large proportion of negative instances.
- **Data level methods**
  - Various methods of re-sampling, under-sampling the majority class, over-sampling the minority class, SMOTE.
- **Algorithm level methods**
  - Algorithm modification in order to handle imbalanced data (cost sensitive learning, one class learning, ensemble learning).



SMOTE
Synthetic Minority Oversampling Method

Modeling

- Learning part, we are building the model.
- Estimating parameters of the model.
- The modeling process usually consist of several steps:
  1. Model selection
     1. Model learning (Training phase)
     2. Model validation (Validation phase)
  2. Model assessment (Testing phase)

We focus on:

- Frequent pattern analyses, association, correlations
- Classification & regression
- Clustering

Frequent patterns, Associations, Correlations

- Unsupervised techniques.
- Frequent patterns (itemsets, subsequences, substructures)
  - Shopping basket: $\{milk, diapers, beer\}$
- Associations
  - Shopping basket: $diapers \Rightarrow beer$
- Correlation analyses
  - Excluding correlated itemsets (not interesting associations).

- **FP-growth**
- **Apriori algorithm**

Classification & Regression

In supervised learning problem we approximate an unknown target function

$$f : \mathcal{X} \to \mathcal{Y}$$

or equivalently, we approximate posterior probability

$$P(Y|X).$$

Generative models

Generative methods model class-conditional density $p(x|y)$. Having class priors $p(y)$ and applying Bayes' theorem generative models estimate the posterior probability as

$$p(y|x) = \frac{p(x|y)p(y)}{\sum_i p(x|y_i)p(y_i)}.$$

Discriminative models

Discriminative methods directly model conditional distribution $p(y|x)$ or even only $p(y|x) > 0.5$. Modeling the input distribution $p(x)$ is not needed.

Classification

- Supervised learning method.
- Classification predicts categorical (discrete, unordered) labels.

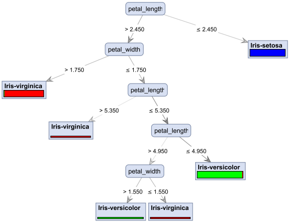
Classification problem

Classification problem is approximation of unknown function (**classifier**)

$$f : \mathcal{X} \rightarrow \mathcal{Y}$$

from the feature space, $\mathcal{X} \in \mathcal{R}^n$, to a label space, $\mathcal{Y} \in \{0, 1, \ldots, n\}$.

- **Nearest Neighbors**
- **Naive Bayes**
- **Decision Tree**
- SVM
- **Neural Networks**



Regression

- Supervised learning method.
- Estimates real value variable.

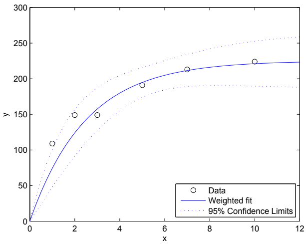
Regression problem

Regression problem is approximation of unknown function (**estimator**)

$$f : \mathcal{X} \rightarrow \mathcal{Y}$$

from the feature space, $\mathcal{X} \in \mathcal{R}^n$, to the output space, $\mathcal{Y} \in \mathcal{R}$.

- **Linear regression**
- Non-linear regression



Clustering

- Unsupervised learning method.
- Grouping of objects into meaningful categories.

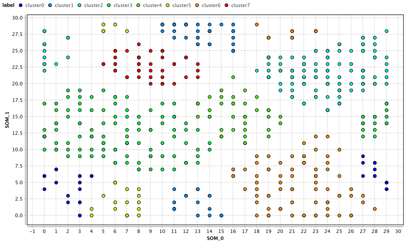
Clustering problem

Given a set of N unlabeled examples, find partitioning π based on a measure of similarity ϕ such that

$$\pi^* = \arg\min_{\pi} f(\pi),$$

where $f(\cdot)$ is formulated according to ϕ.

- **k-means**
- Hierarchical Clustering
- Self Organizing Maps (SOM)



Evaluation

- Evaluation is an important but often underestimated part of model building and assessment.
- Model that perfectly fits training data does not guarantee accurate future prediction - **overfitting**.
- We want reliable model after deployment in the real use.
- Appropriate **evaluation measure**.

Strategies of evaluation:

- Comparison of the model with physical theory.
- Comparison of model with theoretical or empirical model.
- Collect new data for evaluation.
- Use the same data as for model building.
- **Reserve part of the learning data for evaluation.**

Confusion Matrix

		Predicted	
		Positive	Negative
True	Positive	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

Figure : Confusion matrix

- Terms **Positive** and **Negative** refer to the classes.
- **True Positives** are correctly classified instances of positive class.
- **True Negatives** are correctly classified instances of negative class.
- **False Positives** are incorrectly classified positive instances.
- **False Negatives** are incorrectly classified negative instances.

Evaluation Measures for Classification

- **Accuracy**
  - percentage of correctly classified instances

$$Acc(X) = \frac{correctly\ classified\ instances}{all\ instances},$$

in a two-classes case

$$Acc(X) = \frac{TP + TN}{TP + TN + FP + FN}.$$

- **Error rate**

$$Err(X) = 1 - Acc(X).$$

Evaluation Measures for Classification

- **Sensitivity** (*True Positive Rate* or *Recall*)
  - the percentage of truly positive instances that were classified as positive

$$sensitivity = \frac{TP}{TP + FN} .$$

- **Specificity** (*True Negative Rate*)
  - the percentage of truly negative instances that were classified as negative

$$specificity = \frac{TN}{TN + FP} .$$

- **Precision**
  - the percentage of positively classified instances that are truly positive

$$precision = \frac{TP}{TP + FP} .$$

- **F-measure**
  - weighted average of the precision and recall

$$F_\beta = (1 + \beta^2) \frac{precision \cdot recall}{\beta^2 \cdot precision + \cdot recall} .$$

Evaluation Measures for Regression

- **Mean absolute error (MAE)**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i)|.$$

- **Mean squared error (MSE)**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2.$$

- **Root mean squared error (RMSE)**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2.}$$

Data Splitting

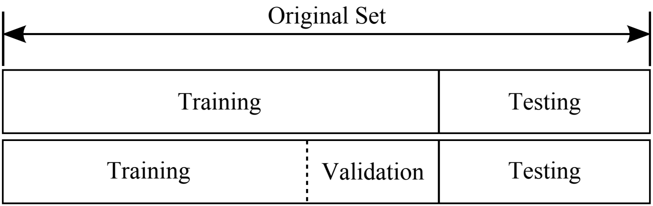


Figure : Two and three way splitting

Training set - a set used for learning and estimating parameters of the model.

Validation set - a set used to evaluate the model, usually for model selection.

Testing set - a set of examples used to assess the predictive performance of the model.

Cross Validation

- A data set is split into k disjoint folds of the same size,
- in each from k turns one fold is used for evaluation and the remaining $k - 1$ folds for model learning,
- the resulting accuracy is the average of all turns,
- typically 10−*fold* cross-validation or Leave-one-out cross-validation