Knowledge Discovery from Data

Knowledge Discovery From Data (KDD)

Knowledge Discovery from Data (KDD): the process of discovering useful **patterns** or knowledge from (large) data sources.

Data sources:

- databases
- text
- images
- World Wide Web
- streaming data (video, audio)

Knowledge discovery from data is often used as a synonym for the term **data mining**. In this course, use the term **KDD** to refer to a wider range of processes. For us, **KDD** incorporates:

- **Data mining**: the techniques, methods and algorithms for finding patterns in structured data.
- **Data warehousing**: the methods and techniques for managing data and processing complex analytical decision-support queries in databases.
- **Information Retrieval:** the techniques, methods, algorithms and data models for finding information in unstructured (primarily, but not always, textual) data.
- **Collaborative Filtering:** the process of filtering for information or patterns using techniques involving collaboration among multiple agents, viewpoints and/or data sources[1]

[1] Wikipedia definition.

Knoweldge Discovery from Data is a multidisciplinary field combining the approaches and methodologies from the following fields:

- **Databases:** KDD activities happen on **very large datasets**. The field of databases deals with efficient storage and management of large quantities of data.
- **Statistics:** the **original** field of *data analysis*. Statistics provides methodology for staging experiments and assessing results. It also provides some basic building blocks for KDD procedures. In addition, a family of KDD methods is based on the use of **probability theory**.
- **Artificial Intelligence:** machine learning, a sub-area of AI studies computer algorithms that improve automatically through experience[2]. The concepts of *supervised learning* (*classification*) and *unsupervised learning* (*clustering*), now and integral part of **data mining**, originated from machine learning and AI.
- **Visualization:** *itself, a multidisciplinary area*, visualization studies the means of clear and understandable representation of information for human consumption.
- **Linguistics:** and **natural language processing** provide rich supply of "building blocks" for analysis of textual data, the same way machine learning and statistics provide building blocks for analysis of structured data.

The Many Faces of KDD

Data is a by-product of human activity. Simple analysis of data can be performed by *querying databases* or *performing statistical analyses* on data. KDD methods seek to provide answers to *more complex* questions about the data.

KDD processes and activities are all around us:

- Google, (yahoo, MS live search);
- Grocery store discount cards;
- Coupons in the mail;
- amazon.com's *"People who bought this book also bought…"*
- last.fm
- Spam filters
- Total Information Awareness
- …

We need to understand a number of aspects of **Knowledge Discovery from Data**:

[2]Tom Mitchell, *Machine Learning*, McGraw Hill, 1997.

- **the technical aspect:** as scientists and engineers we want to know *how KDD works*.
- **the applied aspect:** businesses want to know which KDD methods can help them address their needs.
- **the sinister aspect:** we generate data as a by-product of our activities. We need to be aware of who uses this data and how they are using it.

KDD Process in a Nutshell

The process of **knowledge discovery from data** typically proceeds in **three** steps:

1. **Pre-processing.** Selection of data sources, transformation of raw data into suitable format, data cleaning/filtering,...
2. **Knowledge discovery.** A KDD algorithm is run on the data.
3. **Post-processing.** Output of the KDD algorithm is analyzed, filtered (if necessary), evaluated and visualized.

What we will study

To a large dgree, **Knowledge Discovery from Data** takes a *cookbook* approach to its structure. It is home a large number of diverse problems, which are similar only in that they deal with search for interesting information in large data collections.

Of the various problems that exist under the **extended** KDD umbrella, we will consider the following:

- **On-Line Analytical Query Processing (OLAP) and Data Warehousing**: extensions of database applications for performing analytical tasks on large databases of (typically) business data.
- **Association Rules Mining.** Search of associative patterns in *market basket* datasets.
- **Supervised Learning (Classification).** Determination whether incoming data belongs to a specific class (classes) of objects, based on prior information about these object classes (categories).
- **Unsupervised Learning (Clustering).** Analysis of a collection of data items targeted at combining these items into groups (clusters) based on their percieved similarity.
- **Collaborative Filtering and Recommender Systems.** Formulation of recommendations (predictions) based on similarity patterns discovered in data.
- **Information Retrieval.** Search of textual document collections for documents relevant to user-specified queries.
- **Link Analyisis.** Analysis of graph structures targeted at identifying ”important” components within the graphs.

All of this has be be achieved in a matter of 9.5 weeks!

The course will be *broad* in scope and *shallow* in depth.

Welcome aboard!