Data Mining Techniques and Applications

This compendium lists different data mining tasks and how they can be used. It also presents a list of specific techniques that "solve" each data mining task. Both lists are extensive, but neither one is exhaustive.

This document is ***not*** a substitute for what we do in the class or the book. Knowing about these techniques is good but simply memorizing them will not be useful in the long run. Understanding what each technique does and *how / when* to apply it is far more important.

Data Mining Tasks Shortlist

- Classification (or class probability estimation)
    - Does data instance X **belong to class A or class B** ?
        - Typically, we deal with scenarios with binary classes ("yes" / "no"), but we can have more classes than 2
    - **Example:** Will a *particular* user click on the ad link that we're sending them?
        - Data Instance: the *particular* user, i.e. that individual's traits / features
        - Target Variable: "click" / "no click"

- Regression
    - What is the *predicted* **numerical value** of a *data instance*?
    - **Example:** How many units of product Y do we expect to sell in February?
        - Data instance: product Y, i.e. its characteristics / features
        - Target Variable: Number of units to sell

- Similarity Matching
    - Given an *item I* of interest, find its top-3 ***most similar*** items.
        - Similarity is in the eye of the beholder: many ways to define similarity
    - **Example 1:** Find people to hire that are like my current top performers
    - **Example 2:** Find new customers that are like my current customers
    - **Example 3:** Here's a book that I like. What are similar books?

- Clustering
    - *Does my data fall in "natural" groups?*
    - Characterizing the clusters, i.e., giving them a comprehensive name, is secondary and *entirely up to us*.
        - Data Instance: An item / user / record that will become part of a group

- Co-Occurrence Grouping / Association Rules / Market-basket Analysis
    - Identify items / products / people that frequently appear together
        - **Example 1:** What items are *frequently* purchased together?

            - E.g., milk & bread, chips & beer, noodles & vegetables
        - **Example 2:** Which companies are frequently mentioned in the news together?
        - **Example 3:** Which individuals frequently sign up for the same meetups?

- Profiling
    - What is “normal” behavior in my data?
        - We often ask that question to figure out “atypical” behavior
        - Though not its sole purpose, profiling is often used in domains where acquiring labelled data is nearly impossible and that prevents us from doing classification.
    - **Example:** Is this insurance claim fraudulent ?

- Data Reduction
    - Is the data somehow connected through some *latent* aspect(s)?
    - Also applies as a general “simplification” approach
    - Often, the purpose is to use this / these latent aspect(s) to describe the entire dataset, instead of using the whole thing.
        - **Note 1:** Nowadays, data reduction is incorporated in other more advanced methods. The technique can still be applied on its own, though.
        - **Note 2:** Like clustering, the data mining task does *not* characterize the latent dimension(s) that connect the data. It simply identifies them.
    - **Example:** What “genres” - without knowing their names in advance - do these movies form?

- Link Prediction
    - In a graph / (social) network, figure out if graph nodes should be connected.
        - This task feels like a “classification” problem. The reason for the distinction has to do with the setting (graphs) and the fact that there are techniques which rely on other approaches, not based on classification.
    - **Example 1:** Should there be a link between User A and User B?
    - **Example 2:** Should there be a link between User A and webpage B?

- Causal Modeling
    - Understanding why things take place in a statistically robust way
        - Another way to pose this question is: Does factor X affect the outcome?
    - **Example 1:** Why are users leaving my service?

Data Mining Techniques and Applications

Task	Techniques	Known Applications
Classification	● Logistic Regression ● Decision Trees ● Random Forests ● SVM ● Bayesian Models ● Neural Networks ● XG Boosted Trees ● Deep Learning	- Customer Churn - Recommendation - Targeted Advertising
Regression	● Linear Regression (OLS, GLS) ● Moving-Average model ● ARIMA ● VAR	- Logistics
Similarity Matching	● Lp Distances ● Cosine Similarity ● LCSS ● Dynamic Time Warping	- *Used in clustering and classification* - Find similar items
Clustering	● k-Means ● k-Medoids (PAM) ● Hierarchical Clustering ● DBScan	- Data Understanding - Data Summarization
Co-Occurrence Grouping	● A Priori ● FP-Growth	- Market-Basket Analysis - Promotions - Physical Item placement
Profiling	*Depends on the application. Use of statistical information*	- Credit Scoring - Fraud
Data Reduction	● Matrix Factorization ● Smoothing & Interpolation (discrete -> continuous)	- Visualization - Recommendations (again, it happens implicitly) - Identifying topics / genres without knowing what they are
Link Prediction	● Graph Distance ● Common Neighbors / Jaccard ● Adami-Adar model ● Preferential Attachment ● Personalized PageRank	- (Social) Recommendations - Social Research - Personalization - Ad-Campaigns

Causal Modeling	• Counterfactuals	Well, if you identify “cause and effect” relations, you can do whatever you want :-)

The techniques for each data mining task are interchangeable, for the most part. This means that if you can understand the type of problem you must solve for a business, you can then pick and choose a technique from the respective list (or others). Trying out the various techniques without a good comprehension of the problem at hand is not going to work.

Now, there are differences between the *techniques* that solve the same task and we will be discussing some of them during class. Those are a lot more technical and a lot of people may not be interested in them. If you are, let me know and I will point you in the right direction.

Also, the “applications” column has to do with where these techniques are often used. You must also keep in mind that this is *not* a clear cut separation. If you understand **what** each technique does, i.e., what it uses as input and what output it provides, you should be able to figure out how to use it. For example, if you understand the latent factor that connects movies that someone watches, you can use that to improve the recommendations. If items are very frequently bought together, you can use that information for targeted ad-campaign and so on.

What the technique does (learning a model) and how to use its output (using / applying a model) are two entirely different things!

Also, the following matrix can often help you decide what type of problem you are dealing.

Task	Supervised Methods	Unsupervised Methods
Classification	✓	
Regression	✓	
Causal Modeling	✓	
Similarity Matching	✓	✓
Link Prediction	✓	✓
Data Reduction	✓	✓
Clustering		✓
Co-occurrence Grouping		✓
Profiling		✓

Side note & Example

These tasks do not live in isolation of each other. They can be combined to produce MORE meaningful results.

Business Goal: Let's say that our business sells *niche market* products. The marketing department wants to send out some discounts / offers to "extreme" customers. That's their best explanation of what they are looking for. They ask you to send them a list of customers that meet the description.

How do you approach the problem?

Give it some thought. I'm opening a Forum thread where we can discuss this.