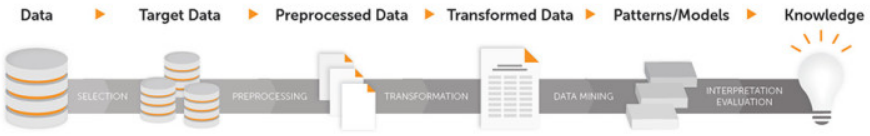
4 Data Mining Techniques for Businesses (That Everyone Should Know)

by **Galvanize** February 8, 2016

Data Mining is an important analytic process designed to explore data. Much like the real-life process of mining diamonds or gold from the earth, the most important task in data mining is to extract non-trivial nuggets from large amounts of data.



Extracting important knowledge from a mass of data can be crucial, sometimes essential, for the next phase in the analysis: the modeling. Many assumptions and hypotheses will be drawn from your models, so it's incredibly important to spend appropriate time "massaging" the data, extracting important information before moving forward with the modeling.

Although the definition of data mining seems to be clear and straightforward, you may be surprised to discover that many people mistakenly relate to data mining tasks such as generating histograms, issuing SQL queries to a database, and visualizing and generating multidimensional shapes of a relational table.

For example: data mining is not about extracting a group of people from a specific city in our database; the task of data

about creating a graph of, say, the number of people that have cancer against power voltage—data mining's task in this case could be something like: is the chance of getting cancer higher if you live near a power-line?

The tasks of data mining are twofold: **create predictive power**—using features to predict unknown or future values of the same or other feature—and **create a descriptive power**—find interesting, human-interpretable patterns that describe the data. In this post, we'll cover four data mining techniques:

- Regression (predictive)
- Association Rule Discovery (descriptive)
- Classification (predictive)
- Clustering (descriptive)

Regression

Regression is the most straightforward, simple, version of what we call "predictive power." When we use a regression analysis we want to predict the value of a given (continuous) feature based on the values of other features in the data, assuming a linear or nonlinear model of dependency.

Here are some examples:

- Predicting revenue of a new product based on complementary products.
- Predicting cancer based on the number of cigarettes consumed, food consumed, age, etc.
- Time series prediction of stock market and indexes.

Regression techniques are very useful in data science, and the

strength of neural networks that use a regression-based technique to create complex functions that imitate the functionality of our brain.

Association Rule Discovery

Association rule discovery is an important descriptive method in data mining. It's a very simple method, but you'd be surprised how much intelligence and insight it can provide—the kind of information many businesses use on a daily basis to improve efficiency and generate revenue.

Our goal is to find all rules (X —> Y) that satisfy user-specified *minimum* **support** and **confidence** constraints, given a set of transactions, each of which is a set of items. Given a set of records—each of which contain some number of items from a given collection—we want to find dependency rules which will discover *occurrence* of an item based on *occurrences* of other items.

For example: Assume you have a dataset of all your past purchases from your favorite grocery store, and I found a

dependency rule (minimizing with respect to the constraints) between these items: {Diapers} —> {Beer}.

This "links" or creates dependencies, based on the specified minimum support and confidence, which are defined as such:

$$\text{SUPPORT} = \frac{\text{number of transactions containing X and Y}}{\text{total number of transactions}}$$

number of transactions containing X

The applications for associate roles are vast and can add lots of value to different industries and verticals within a business. Here are some examples: Cross-selling and up-selling of products, network analysis, physical organization of items, management, and marketing. This was an industry staple for decades in market basket analysis, but in recent years, recommendation engines have largely come to dominate these traditional methods.

Classification

Classification is another important task you should handle before digging into the hardcore modeling phase of your analysis. Assume you have a set of records: each record contains a set of attributes, where one of the attributes is our *class* (think about letter grades). Our goal is to find a model for the *class* that will be able to *predict* **unseen** or **unknown** records (from external similar data sources) *accurately* as if the label of the class was **seen** or **known**, given all values of other attributes.

In order to train such a model, we usually divide the data set into two subsets: *training set* and *test set*. The training set will be used to build the model, while the test set used to validate it. The accuracy and performance of the model is determined on the test set.

Classification has many applications in the industry, such as direct marketing campaigns and churn analysis:

targeting a set of consumers that are likely to be interested in the specific content (product, discount, etc.) based on their revealed past data and behavior.

The method is simply to collect data for a similar product (for simplicity) introduced in the recent past and to *classify* the profiles of customers based upon whether they **did buy** or **didn't buy**. This target feature will become the *class attribute*. Now we need to enhance the data with additional demographic, lifestyle, and other relevant features in order to use this information as input attributes to train a classifier model.

Churn is the measure of individuals losing interest in your offering (service, information, product, etc.). In business it's incredibly important to monitor churn and attempt to identify why subscribers (clients, etc.) decided to stop paying for the subscription. In other words, churn analysis tries to predict whether a customer is likely to be lost to a competitor.

To analyze churn, we need to collect a detailed record of transactions with each of the past and current customers, to find attributes that can explain or add value to the question in hand. Some of these attributes can be related to how engaged the subscriber was with the services and features that the company offers. Then we simply need to *label* the customers as **churn** or **not churn** and find a model that will best fit the data to predict how likely each of our current subscribers is to churn.

Clustering

such that objects within the same cluster are similar to each other, while objects in different groups are not. The Clustering problem in this sense is reduced to the following:

Given a set of data points, each having a set of attributes, and a similarity measure, find clusters such that:

01. Data points in one cluster are more similar to one another.
02. Data points in separate clusters are less similar to one another.

In order to find how close or far each cluster is from one another, you can use the Euclidean distance (if attributes are continuous) or any other similarity measure that is relevant to the specific problem.

A useful application of clustering is marketing segmentation, which aims to subdivide a market into distinct subsets of customers where each subset can be targeted with a distinct marketing strategy.

This is done by collecting different attributes of customers based on their geographical- and lifestyle-related information in order to find clusters of similar customers. Then we can measure the clustering quality by observing the buying patterns of customers in the same cluster vs. those from different clusters.

Preferred Campus *

Please select...

Email *

Phone *

###-###-####

By submitting your information below, you agree to our Terms of Use and Privacy Policy.

Get More Info

Navigation

Home

Blog

Mentors

Press

Venture

Careers

Member Login

Categories

Coworking

Galvanize

Level Up

Series

Software Engineering

Uncategorized

Video

Web Development

Work

Campuses

Austin, TX

Boulder, CO

Denver, CO – Platte

Denver, CO – Golden Triangle

New York, NY

Phoenix, AZ

San Francisco, CA

Seattle, WA

About Us

Galvanize is a collection of modern, urban campuses where people can access the skills and network they need in-person or online to level up in tech. Our culture is shaped by first-time entrepreneurs and growing startups, to Fortune 1000 companies.