ITEC874 — Big Data Technologies

Week 8 Lecture 1: Analysing Big Data

Diego Mollá

ITEC874 2019H2

Abstract

In this lecture we will introduce some of the key concepts about analysing big data.

Update October 1, 2019

Contents

Reading

- Big Data Challenges and Analysis-Driven Data, Chapter 1.
- Text Analytics Using SAS Text Miner, Appendix "Predictive Modeling", sections A1 to A3.

1 Big Data Analytics

Revisiting the Meaning of "Big Data"

Big Data - Wikipedia Definition
Big data is data sets that are so *big* and *complex* that traditional data-processing application software are inadequate to deal with them.

- We know about the four V's of Big Data:
  - Volume
  - Variety
  - Velocity
  - Veracity
- In this and following lectures we will focus on the issue of *variety*.

Before Big Data

- Companies have used large volumes of data before the term "Big Data" was coined.
- Traditional approaches to handle large volumes of data assumed:
  - The data were well structured.
  - The data would often come from internal sources.
  - Data were often used for *descriptive and diagnostic analytics.*
- All of this has changed with the advent of Big Data.

It's About Variety, not Volume.

Structured, Semi-structured, Unstructured Data

Structured Data

- Information stored in relational databases.
- All data and their relations are clearly defined.

Semi-Structured Data

- Data are presented in a loose structure.
- Data and their relations are less clearly defined.
- For example, the contents of fields are free text.

Unstructured Data

- Data are presented as collections of text, videos, images, etc.
- These data are originally created for people.
- Machines have difficulty processing these data.
- Most of the information available is unstructured data.

2 Analysing Data

2.1 Big Data Analytics

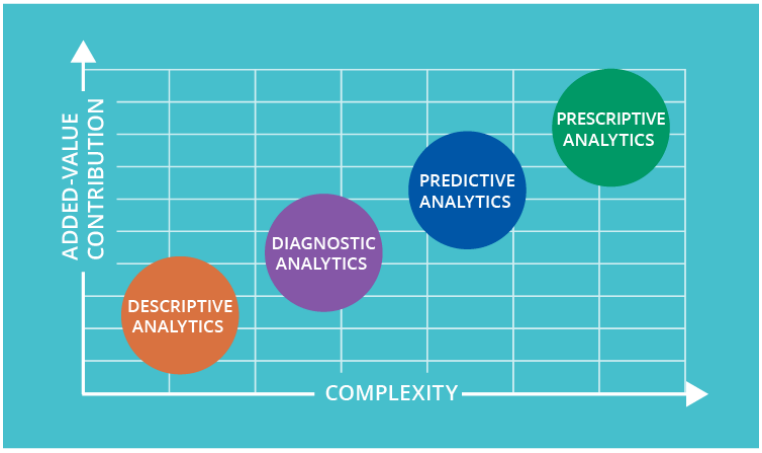
What is Big Data Analytics?

WhatIs.com
Big data analytics is the process of examining large and varied data sets — i.e., big data — to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions.

Benefits of Big Data Analytics
https://www.sas.com/en_au/insights/analytics/big-data-analytics.html

- Cost Reduction. Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data — plus they can identify more efficient ways of doing business.
- Faster, better decision making. With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately — and make decisions based on what they've learned.
- New products and services. With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers' needs.

Four Types of Analytics



https://www.scnsoft.com/blog/4-types-of-data-analytics

Four Types of Analytics

Descriptive Analytics
Analyse past and present data with the aim to understand it.

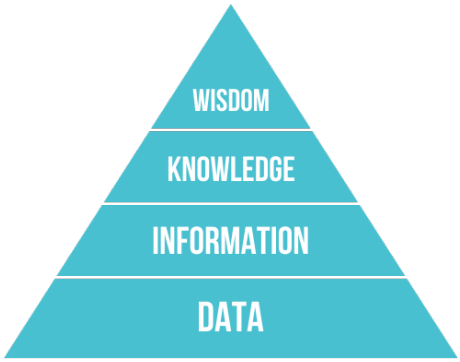
Diagnostic Analytics
Analyse past and present data to determine *what happened* and *why.*

Predictive Analytics
Use models based on past data to predict the future. The deliverables are usually a predictive forecast.

Prescriptive Analytics
Use models to specify what actions should be taken. This is the most valuable kind of analysis.

Data, Information, Knowledge, Wisdom

By Longlivetheux - Own work, CC BY-SA 4.0, *https://commons.wikimedia.org/w/index.php?curid=37705247*



Data, Information, Knowledge, Wisdom

Data

- The raw facts generated from observation or activities.
- e.g. samples of observations of rain or no rain in Sydney.

Information

- Patterns, associations, relationships among the data.
- e.g. observation that the temperature drops two degrees when it rains.

Knowledge, Wisdom

- The appropriate combination of information that explains the usefulness of the data and can be used for business decisions.
- e.g. we decide on measures of water restrictions.

Data Mining and Business Intelligence

Data Mining

- Extracting useful information from large data sets.
- The process of exploration and data analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules.
- The process of discovering meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories.

Business Intelligence

- Business intelligence (BI) is a business management term, which refers to applications and technologies that are used to gather, provide access to, and analyse *data* and *information* about company operations.
- Business intelligence systems can help companies have a more comprehensive *knowledge* of the factors affecting their business, such as metrics on sales, production, internal operations, and they can help companies to make better business decisions.

Uses of Data Mining for Business Intelligence

1. From a large list of prospective customers, which are most likely to respond?
2. Which customers are most likely to commit fraud?
3. Which loan applicants are likely to default?
4. Which customers are most likely to abandon a subscription service?

Statistical Learning and Machine Learning

- *Statistical learning* is about inferring rules based on sample data.
- Possible uses of statistical learning are:

  Analysis: Process a data set with the goal to achieve a better understanding of its characteristics.
  Prediction: Learn rules that allow us to predict outcomes.
- We will focus on *machine learning*: Conducting statistical learning automatically.

Question
How does statistical learning / machine learning relate to the four types of analytics, and to the concepts of data/information/knowledge/wisdom?

Example of Machine Learning for Analysis

Analysis
We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

Example of Machine Learning for Prediction

Prediction
We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

Types of Machine Learning

Supervised machine learning

- In supervised machine learning, we have a *training set* with samples where we know the prediction.
- This training set has been annotated, usually manually.
- The training set is used to learn a *model*.
- The model is then used to make predictions on *unseen data*.
  - Unseen data is data that is not part of the training data.

Unsupervised machine learning

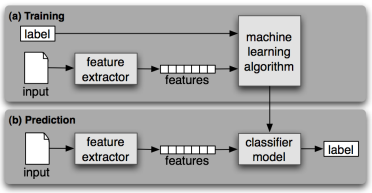
- In unsupervised machine learning, there is no training set.
- We process a data set with the aim to extract useful information from it.
- An example is mining association rules.
- Another example is clustering data.

Supervised Learning

Given
Training data annotated with class information.

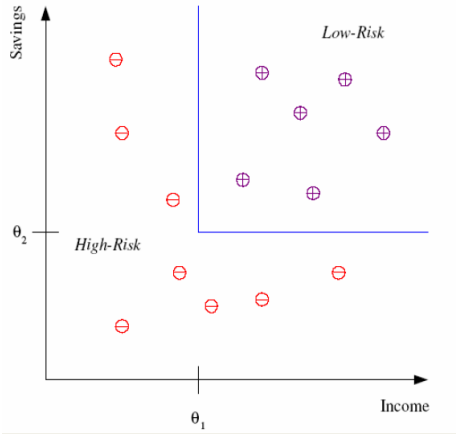
Goal
Build a *model* which will allow classification of new data.

Method



1. *Feature extraction*: Convert samples into vectors.
2. *Training*: Automatically learn a model.
3. *Classification*: Apply the model on new data.

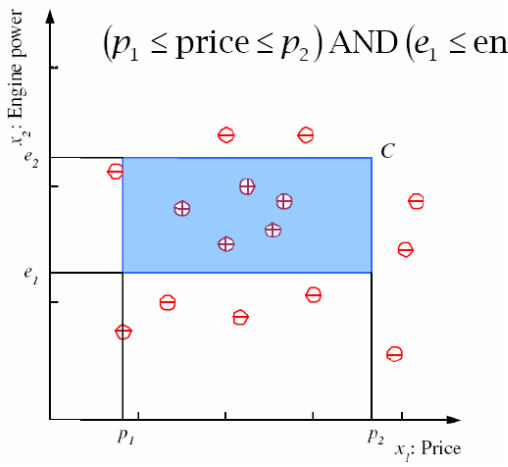
Supervised Learning Example: Bank Customers



(from Alpaydin (2004))

This is an example of a training dataset where each circle corresponds to one data instance with input values in the corresponding axes and its sign indicates the class. For simplicity, only two customer attributes, income and savings, are taken as input and the two classes are low-risk ('+') and high-risk ('−'). An example model that separates the two types of examples is also shown. The model has two parameters, θ_1 and θ_2, which need to be set. These parameters are either set manually by observing the training data, or automatically by training the model using supervised machine learning. Image from: E. Alpaydin. 2004. Introduction to Machine Learning. ©The MIT Press.

Supervised Learning Example: Family Cars



(from Alpaydin (2004))

Another example where now the model has four parameters. The class of family car is a rectangle in the price-engine power space. From: E. Alpaydin. 2004. Introduction to Machine Learning. ©The MIT Press.

Types of Variables (features)

Numeric

- Numeric variables can be continuous (real) or integer.

Categorical

- May be *ordered* (low, medium, high) or *unordered* (male, female).
- May be *binominal* (binary) or *polynominal*.
- Deep learning approaches sometimes convert polynominal variables to vectors of numeric variables by using *embedding*.

Classification and Regression

These are *supervised* learning methods.

Classification

- The goal is to predict an outcome in a *categorical* variable.
- E.g. purchase/no purchase, fraud/no fraud, creditworthy/no creditworthy.
- Target variable is often *binominal* (yes/no) but it may be *polynominal* (fixed and finite number of unordered values).

Regression

- The goal is to predict a *numeric* target.
- E.g. sales, revenue, performance.

Unsupervised Learning

Given
Data *without annotations*.

Goal
Build a *model* which will *find structure* in the data.

Method
There is no separate training stage because there is no training data.

1. *Feature extraction*: Convert samples into vectors.
2. *Modeling*: Find structure from the data.

Association Rules and Clustering
These are *unsupervised* learning methods.

Association rules

- The goal is to produce rules that define "what goes with what".
- E.g. "if X was purchased, Y was also purchased."
- Also called *affinity analysis* and *basket market analysis*.

Clustering

- The goal is to find natural groups or hierarchies among the data.
- E.g. find groups of shoppers with similar interests.

2.2 Steps in a Data Mining Project

Steps in a Data Mining Project

1. Develop an *understanding* of the purpose of the data mining exercise.
2. Obtain the data set, e.g. by *sampling*.
3. Explore, clean, and preprocess the data.
4. Reduce and partition the data.
    - Supervised tasks need a *training* and a *testing* set, and often a *development* set.
5. Determine the data mining task and technique.
6. Iterative implementation and parameter tuning.
7. Assess the results; compare models.
8. Deploy the best model.
9. Evaluate or Monitor Results.
10. Start all over again!

Identify or Formulate the Problem

Examples

1. Improve the response rate for a direct marketing campaign.
2. Increase the average order size.
3. Determine what drives customer acquisition.
4. Forecast the size of the customer base in the future.
5. Choose the right message for the right groups of customers.
6. Target a marketing campaign to maximize incremental value.
7. Recommend the next, best product for existing customers.
8. Segment customers by behaviour.

A lot of good statistical analysis is directed at solving the wrong business problem.

Obtain the Data Set

Possible questions to ask about the data

- What is available?
- What is the correct level of granularity?
- How much is needed?
- How much history is required?
- Do we want to sample from the data?

Data Exploration

Data exploration is usually required to review the data and help refine the task. It uses techniques of reduction and visualisation.

Data Reduction

- Distillation of complex/large data into simpler/smaller data.
- Reducing the number of *variables* (columns) and/or *records* (rows).

Data Visualisation

- Graphs and plots of data.
- Histograms, boxplots, bar charts, scatterplots.
- Especially useful to examine relationships between variables.

Missing Data

Most algorithms will not process records with missing values.

Omission

- We may want to omit records with missing values.
- We may want to omit variables with many missing values.
- Sometimes we end up omitting too much information!

Imputation

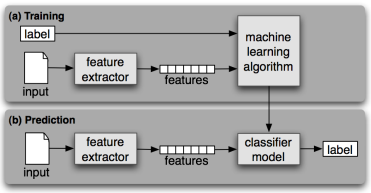
- Replace missing values with reasonable substitutes.
  - e.g. replace with the mean of known values.
- Lets you keep the record and use the rest of its (non-missing) information.

Data Transformation

- Some statistical algorithms require a particular format for their variables.
  - We may need to *bin* numeric variables into categories.
  - We may need to convert polynominal variables into binominal.
- Be careful! Some integer variables are best described as categories.
  - Product codes, (sometimes) month numbers, etc.
  - Ask yourself: is there a natural linear order between the values? For example, is month 2 in some sense "better" or "larger" than month 1? (it may be if time sequence matters).
- Some strings may need to be converted to categories.
  - For example, week days, month names, product type, ...
- In turn, sometimes we may want to convert categorical variables into numbers or vectors, e.g. by applying *embedding* techniques.

Reduction and Partition

- If there are large volumes of data we may want to obtain a sample.
- Often we want to obtain a *random* sample ...
- ... But sometimes we want to keep the linear order, e.g. when performing time series analysis (analysis through time).
- Supervised tasks need a *training* and a *test* set, and often a *development* set.



Parameter Tuning

- Most statistical algorithms have several parameters.
- Many factors affect the choice of optimal parameters.
- Understanding the characteristics of your problem and the properties of each statistical algorithm helps ...
- ... But sometimes you still need to try several parameters.
- Use the candidate models to *score* the validation data. Then compare the results. Select the model with the best performance on the validation data set.

Validation of the Model

- Use the candidate models to *score* the validation data. Then compare the results. Select the model with the best performance on the validation data set.
- Communicate model assessments through the following:
  - quantitative measures (average squared error, misclassification rate, and so on).
  - graphs (cumulative lift, gains, ROC).

Evaluate or Monitor Results

- Compare actual results against expectations.
- Compare the challenger's results against the champion's.
- Did the model find the right people?
- Did the action affect their behaviour?
- What are the characteristics of the customers most affected by the intervention?

Lab results vs. real results

- A common problem is that the results at production time are worse than the results of your experiments.
- Possible causes are:
  - *Poor training data*: Your training data is different from the real data.
  - *The market has changed*: Your training data is obsolete.
  - *Contamination of your data*: Your training data was contaminated with test data.

Begin Again

- Revisit the business objectives.
- Define new objectives.
- Gather and evaluate new data.
  - model scores
  - cluster assignments
  - responses

Example
A model discovers that geography is a good predictor of churn.

- What do the geographies have in common?
- Is the pattern that your model discovered stable over time?

Almost every project raises as many questions as it answers. This is a good thing. It means that new relationships are now visible that were not visible before. The newly discovered relationships suggest new hypotheses to test, and the process can begin again.

A good example of this is performing an analysis and determining that geography is a good predictor of student success in a college program. After you make this determination, you should run an analysis on the different geographies to see whether there is anything that they have in common. You also want to analyze the model over time to see whether this is stable, or to see whether it changed or is changing over time.

The *SEMMA* Process

1. Develop an *understanding* of the purpose of the data mining exercise.
2. Obtain the data set, e.g. by *sampling*.

   S ample
3. Explore, clean, and preprocess the data.

   E xplore
4. Reduce and partition the data.

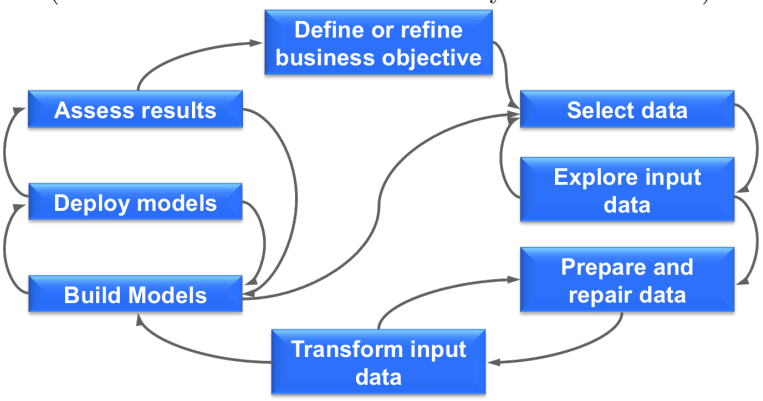
   M odify
5. Determine the data mining task and technique.

   M odel
6. Iterative implementation and parameter tuning.
7. Assess the results; compare models.

   A ssess
8. Deploy the best model.

An Example

(This material is from SAS Text Analytics course notes.)



Business Scenario

At a large university the administration wants to increase the retention rate of new freshmen who enroll. The Office of Institutional Research is given the task of making recommendations for addressing this issue. They made the decision to build models that would predict those students who were most likely not going to return the following semester. They decided to first build a model to predict those new freshmen students who enrolled in the Fall of 2012 and continued to the Spring of 2013. The objective of this analysis is to build a classification model quickly and easily to measure the propensity for these students to not enroll in the Fall of 2013. This will enable intervention with these identified students.

Step 1 — Business Objective

Improve the retention rate of new freshmen who enroll in the university. This will be addressed by building retention models to identify those students who are most likely not going to return the following semester. The first model will be for scoring those new freshmen students who entered in the Fall of 2012 and are currently enrolled in the Spring of 2013. Students in this group who are already receiving special services, such as athletes, will be removed from this study.

Step 2 — Identify and Select Data

- The *data* to build the model will be information about those new freshmen students who enrolled in the Fall of 2011 and continued to the Spring of 2012.
- The *target variable* will be defined by examining students and determining who enrolled and those who did nor enroll in the Fall of 2012. The target variable will take on a value of 1 for those who did not enroll and a value of 0 for those who did enroll.
- The project team will then identify *inputs* that they feel will be good predictors. These inputs include demographic, financial aid, student life, fall semester statistics, admissions, and other data.

Retention Data Inputs

- Course Data
- Demographic Data

- Financial Aid Data
- Alumni Data
- Student Admissions Data
- HR Data
- Some Calculated Inputs

Steps 3 and 4 — Explore Data and Fix Problems

- Explore the data for patterns, unusual values, relationships to target variable, missing values, relationships between input variables, and so on.
- Fix problems related to student IDs as these are used to join tables.

Step 5 — Transform Data

- In this step, new inputs are created, such as distance, age, work-study indicator, transcript indicator, major rate, high school rate, dorm rate, and in-state indicator.
- Those input variables whose distributions are not symmetric may need to be transformed to distributions that are more symmetric.
- Other transformations include binning interval variables, and collapsing levels of a categorical variable.
- Steps 1–5 probably take at least 80% of the project's time.

Step 6 — Model Building

- Modeling tools will be used to build decision tree models, logistic regression models, neural network models, and ensemble models, ...
- The "best" model will be chosen based on an assessment statistic evaluated on the validation, and this model will be used to score the Fall 2012 new freshmen who returned in the Spring 2013 semester.
- The scored data set will identify those students who most likely are not going to enroll in the Fall of 2013.

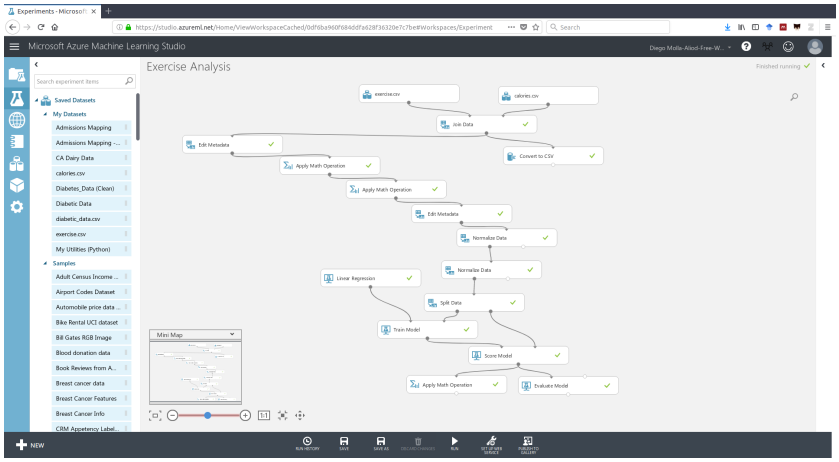
Step 7 — Deploy Model Results

The results of the model scoring will be shared with the appropriate offices on campus that are responsible for retention and enrollment management.

Step 8 — Assess the Model

- After enrollment closes for the Fall 2013 semester, an assessment is made to determine the number of scored students who returned and did not return.
- The model predictions are compared to the actual results and possible recommendations are made concerning the model.
- These recommendations might include the addition of new inputs, using different modeling tools, or perhaps different transformation methods.
- Predictive modeling is an iterative process.

Another Example: Using Microsoft Azure Machine Learning Studio



2.3 Analysing Large Volumes of Data

Machine Learning and Large Volumes of Data

Advantages of Training on Large Volumes of Data

1. Large volumes of training data make machine learning predictions *more accurate*.
   - In *theory*, many machine learning approaches will find the optimal model when trained on infinite volumes of data.
   - In *practice*, the larger the training data, the more accurate the model (up to a point).
2. Large volumes of data allow to build more *complex models*.
   - A key ingredient to the *current success of deep learning* is the availability of large volumes of training data.
   - More complex models, when trained on large training sets, often lead to better results.

Machine Learning and Large Volumes of Data

Problems of Training on Large Volumes of Data

1. Training data sets that are large enough may not fit in RAM.
   - But this problem is becoming less relevant given the current availability of cheap and large RAM.
2. Large volumes of training data make the training process *slow*.
   - MapReduce techniques are less useful here.
     - *Why...?*
   - Clusters of computers and grid help up to a point.
   - Current solutions use dedicated *Graphics Processing Units* (GPUs).

Why a GPU Helps

- Modern computers process data in the CPU and, if available, in the GPU.
- Graphics processing is usually based on *matrix operations*, and GPUs were designed to speed up these matrix operations.
- It turns out that some of the most computer-intensive parts of machine learning involve matrix operations.

  CPU = Central Processing Unit

  - Tens (or less) of computation cores.
  - Single-threaded.
  - Able to perform any computations.

  GPU = Graphics Processing Unit

  - Hundreds (or more) of computation cores.
  - Thousands of concurrent hardware threads.
  - Can only perform simple computations.

Take-home Messages

- What is Big Data Analytics?
- Types of Analytics.
- Data, information, knowledge, wisdom.
- Types of Machine Learning.
- Steps in a Data Mining project.
- Analysing Large Volumes of Data.

What's Next

Assignment 2

- Submission deadline: 6 October.

Week 9

- Text Analytics.
- Assignment 3 will be released.