COMP8210 — Big Data Technologies

Week 7 Lecture 1: Analysing Big Data

Diego Mollá

Department of Computer Science
Macquarie University

COMP8210 2020H2

Programme

Reading

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

Programme

Revisiting the Meaning of "Big Data"

Big Data - Wikipedia Definition (14 August 2020)

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

- We know 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.

Programme

Programme

What is Big Data Analytics?

WhatIs.com

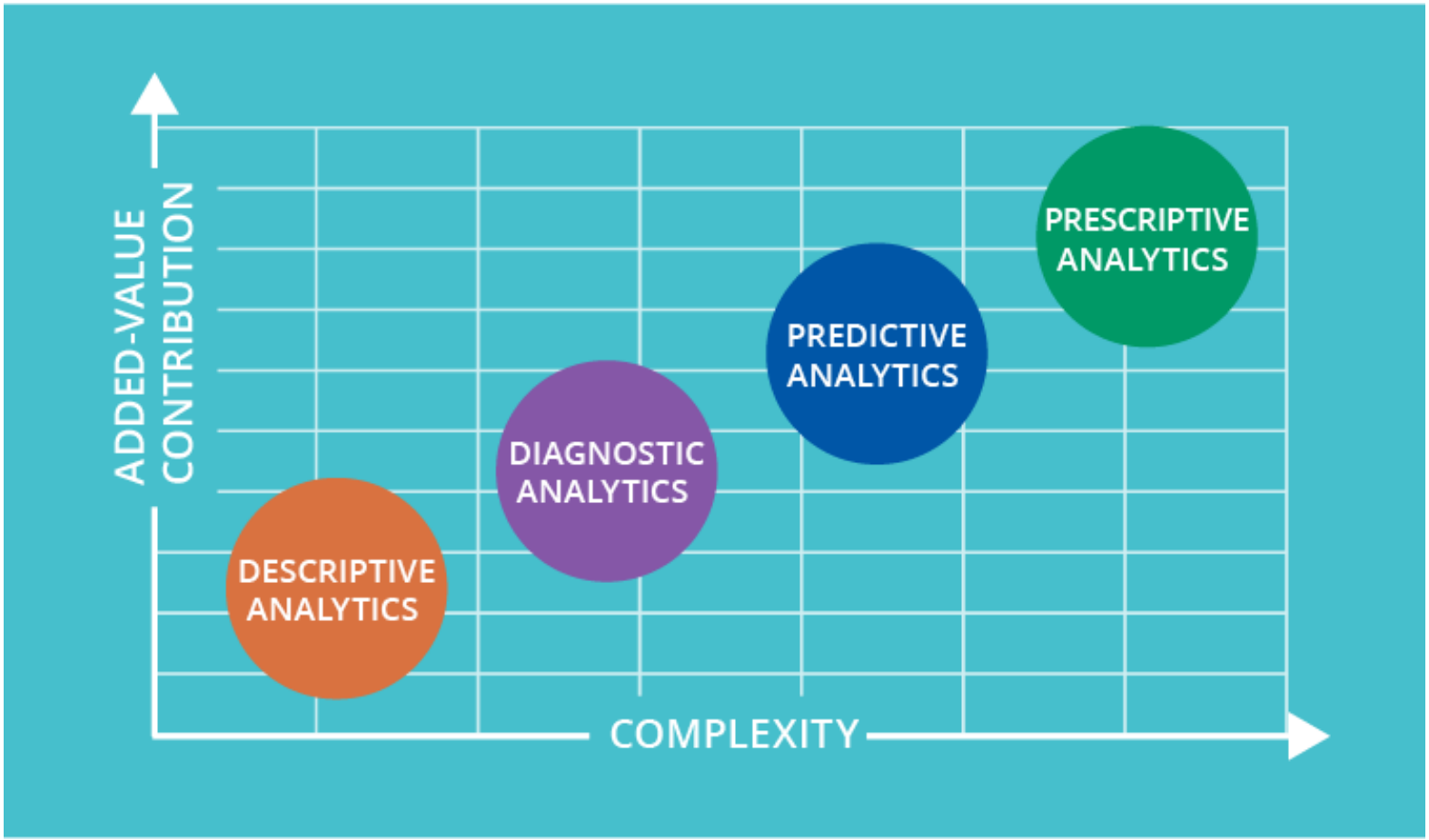
Big data analytics is the process of examining large and varied data sets — i.e., big data — to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions.

Benefits of Big Data Analytics

https://www.sas.com/en_au/insights/analytics/big-data-analytics.html

- Cost Reduction.
- Faster, better decision making.
- New products and services.

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.

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.

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.

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

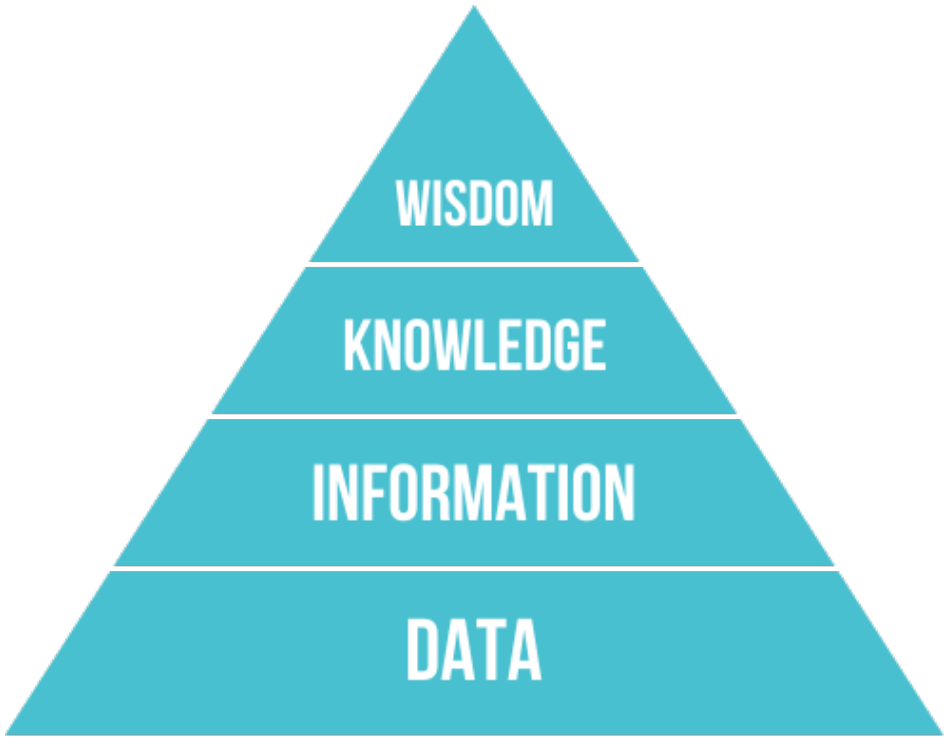
Use models based on past data to predict the future. The deliverables are usually a predictive forecast.

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, 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, 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 I

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.

Data Mining and Business Intelligence II

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?

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?

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?

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 I

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.

Types of Machine Learning II

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

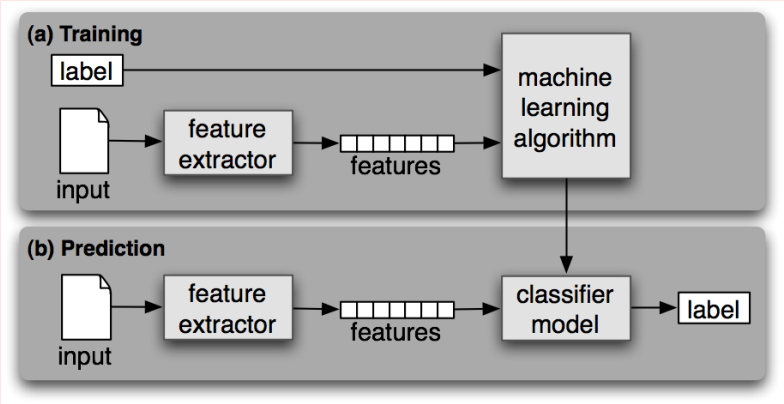
Given

Training data annotated with class information.

Goal

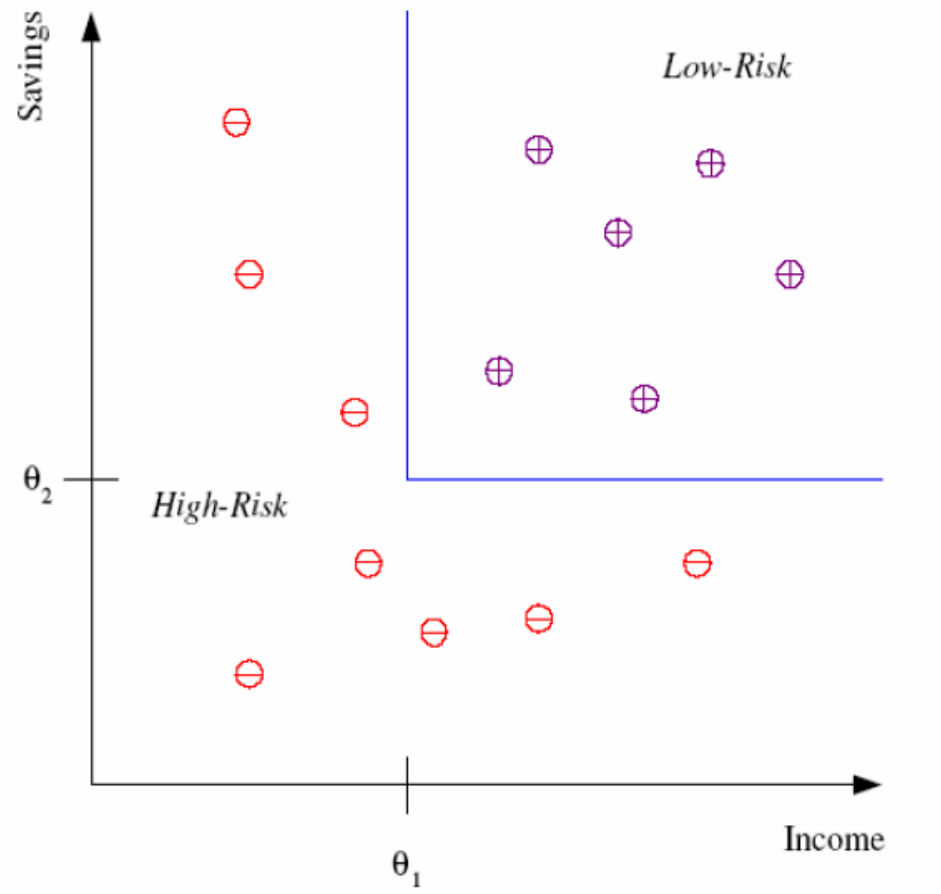
Build a model which will allow classification of new data.

Method



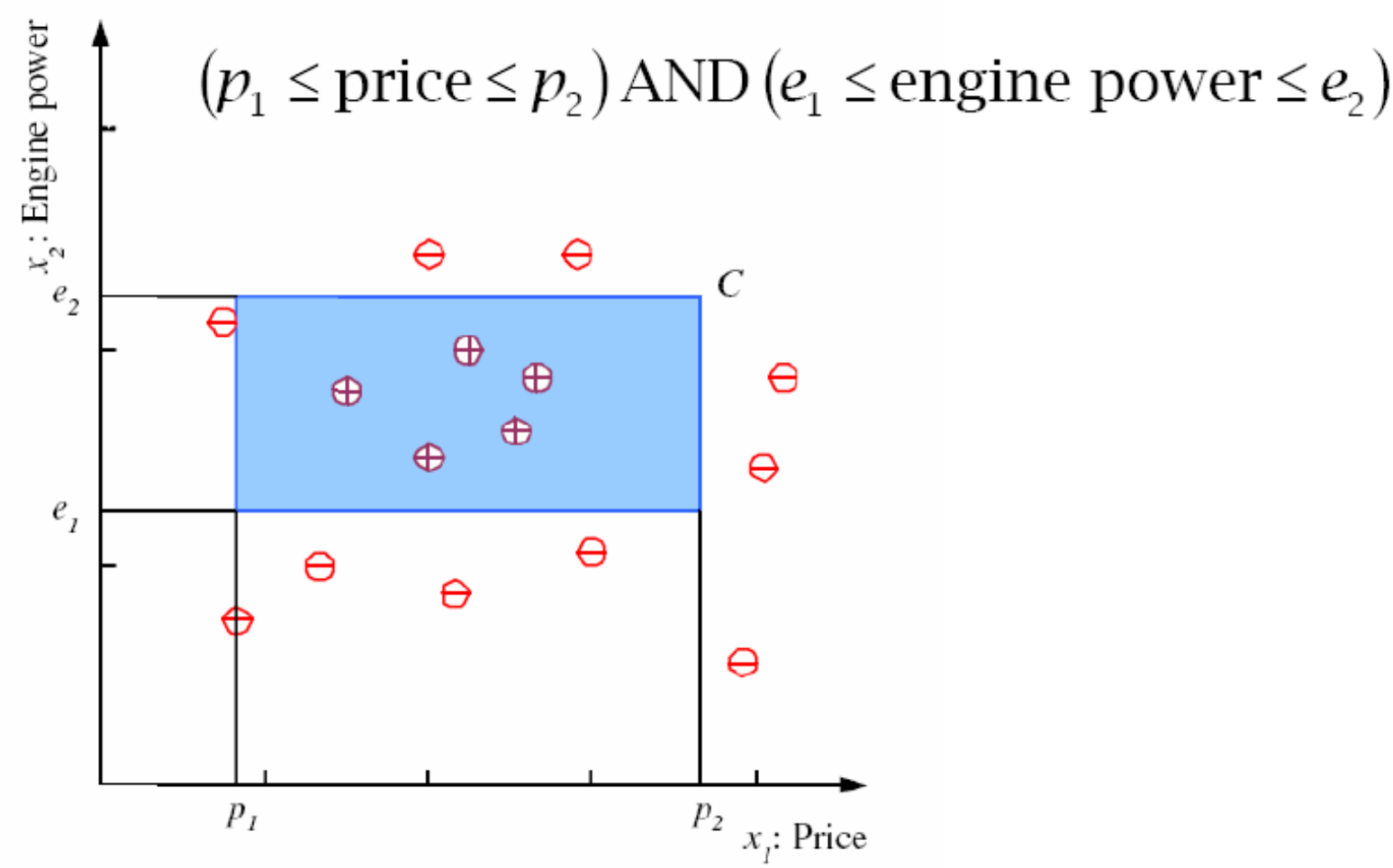
1. Feature extraction: Convert samples into vectors.
2. Training: Automatically learn a model.
3. Classification: Apply the model on new data.

Supervised Learning Example: Bank Customers



(from Alpaydin (2004))

Supervised Learning Example: Family Cars



(from Alpaydin (2004))

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.

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

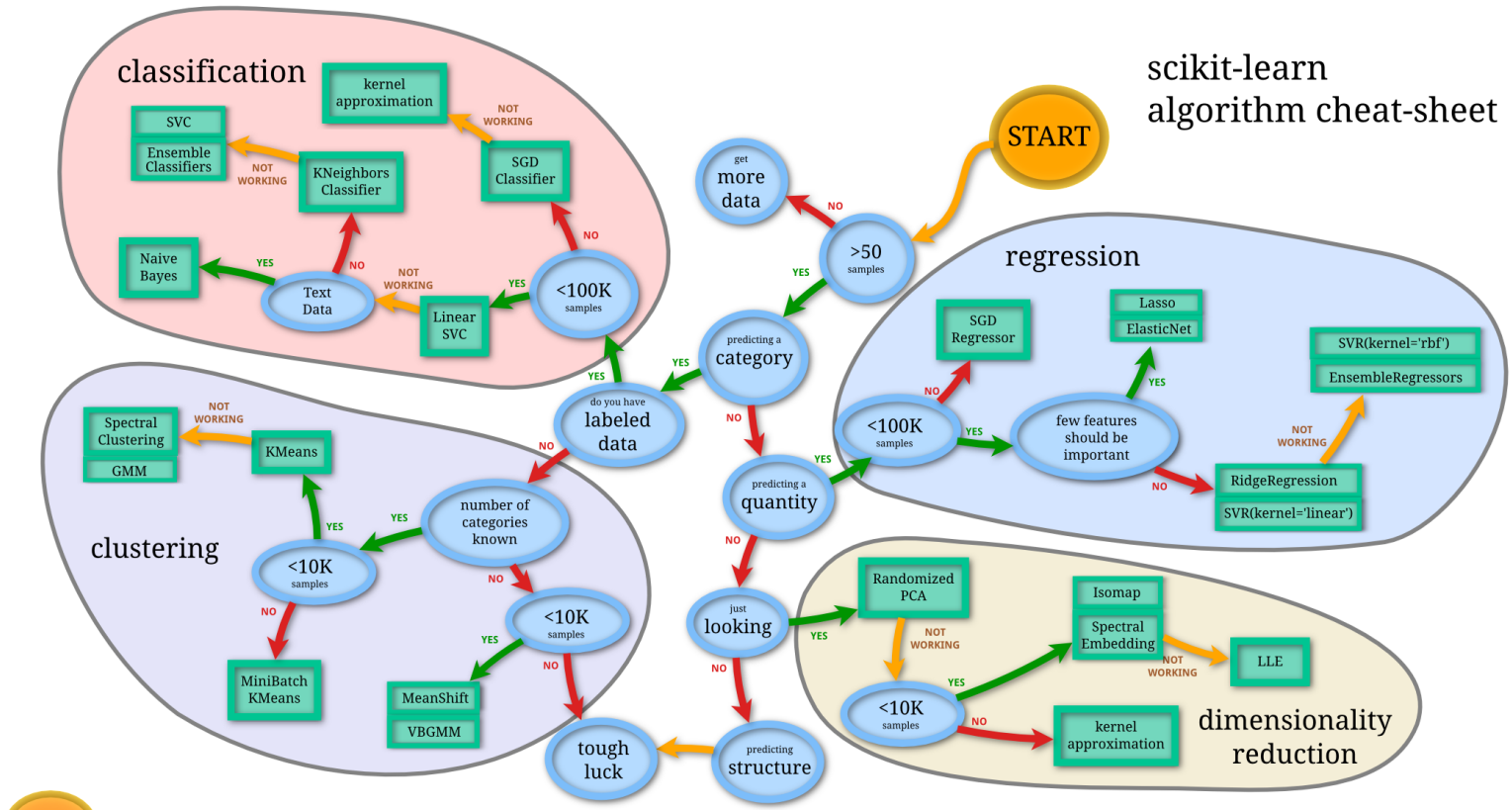
- The goal is to produce rules that define "what goes with what".
- E.g. "if X was purchased, Y was also purchased."
- Also called affinity analysis and basket market analysis.

Clustering

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

Choosing the Right Model

https://scikit-learn.org/stable/tutorial/machine_learning_map



Programme

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.
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!

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.
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.

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.
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!

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?

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.
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!

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.

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.

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

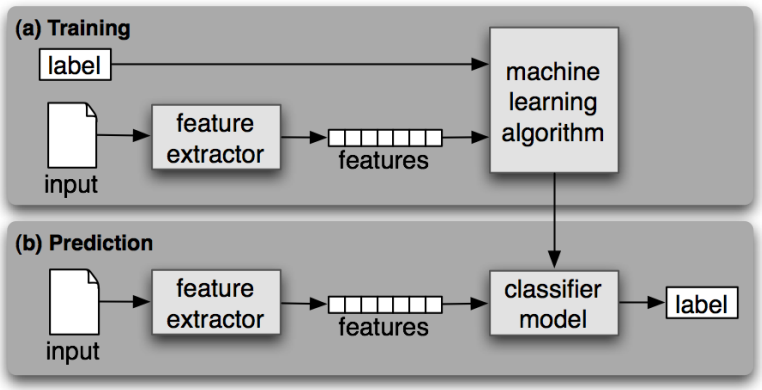
- Some statistical algorithms require a particular format for their variables.
  - We may need to bin numeric variables into categories.
  - We may need to convert polynominal variables into binominal.
- Be careful! Some integer variables are best described as categories.
  - Product codes, (sometimes) 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.

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.
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!

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.



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.
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!

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.
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!

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.

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.
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!

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).

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.
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!

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.
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!

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.

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.
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!

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?

CRISP-DM, SEMMA

- These are two popular standards for data process.
- The process we have covered above subsumes them.

CRISP-DM

SEMMA

- Sample
- Explore
- Modify
- Model
- Assess

Image by Kenneth Jensen - Own work based on: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf (Figure 1), CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=24930610

An Example in Python

In a jupyter notebook we will apply the above steps for a very simple task: Classify flowers using the Iris data set.

Programme

1. Big Data Analytics

2. Analysing Data
   - Big Data Analytics
   - Steps in a Data Mining Project
   - 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

Mid-semester Break

- No classes between 14 and 25 September.

Week 8

- Topic: Text Analytics.
- Assignment 2 deadline.