Machine Learning Introduction

Data & Models

Machine Learning Data

- Instances
  - Instances are like objects (e.g customer)
  - Instances each have a predefined set of features or attributes.
- Input is a single relational table, often a csv file.
- Rows are instances.
- Columns are features.

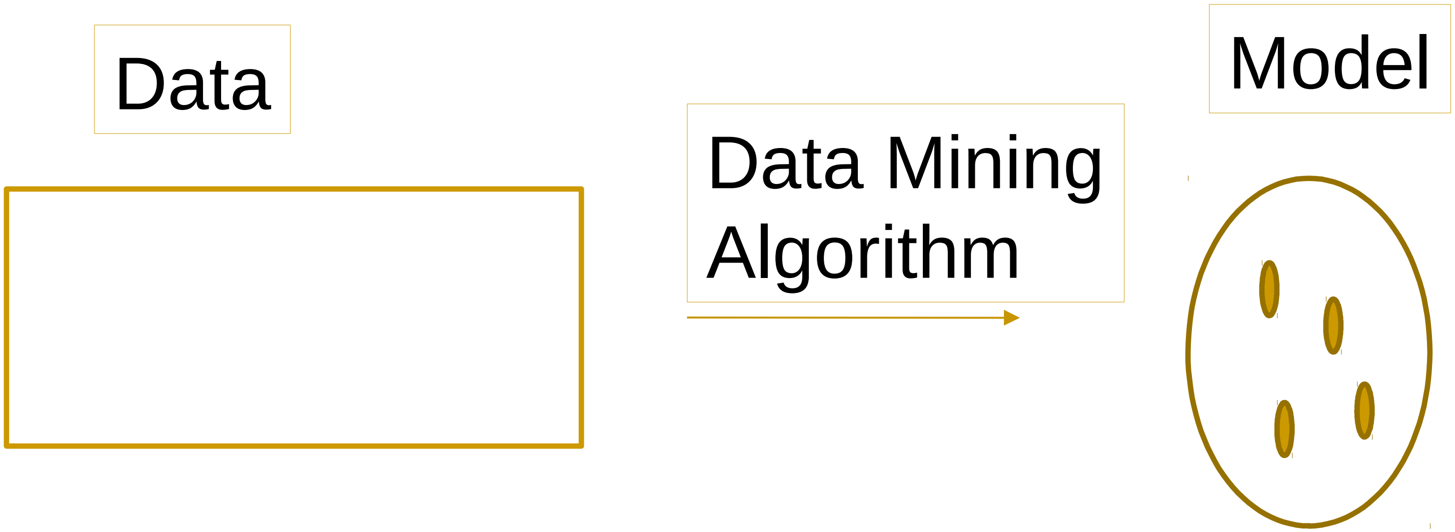
Features

- Features
  - Can be numerical or categorical.
- Non numerical or categorical features have a set of predefined values. Can be nominal (no order) or ordinal
- Examples - Numerical features
  - number calls made, MB downloaded
- Examples - Nominal features
  - make of car
- Examples - Ordinal Features
  - usage of something (low, medium & high)

Data Mining Model

- Data Mining algorithms build models.
- Example is a decision tree.
- For example, a decision tree of size 1 (referred to as 1R)
  - if number calls tech support > 10 then churn
- (A decision tree can have a large number of nodes.)

Data Mining – Model Building

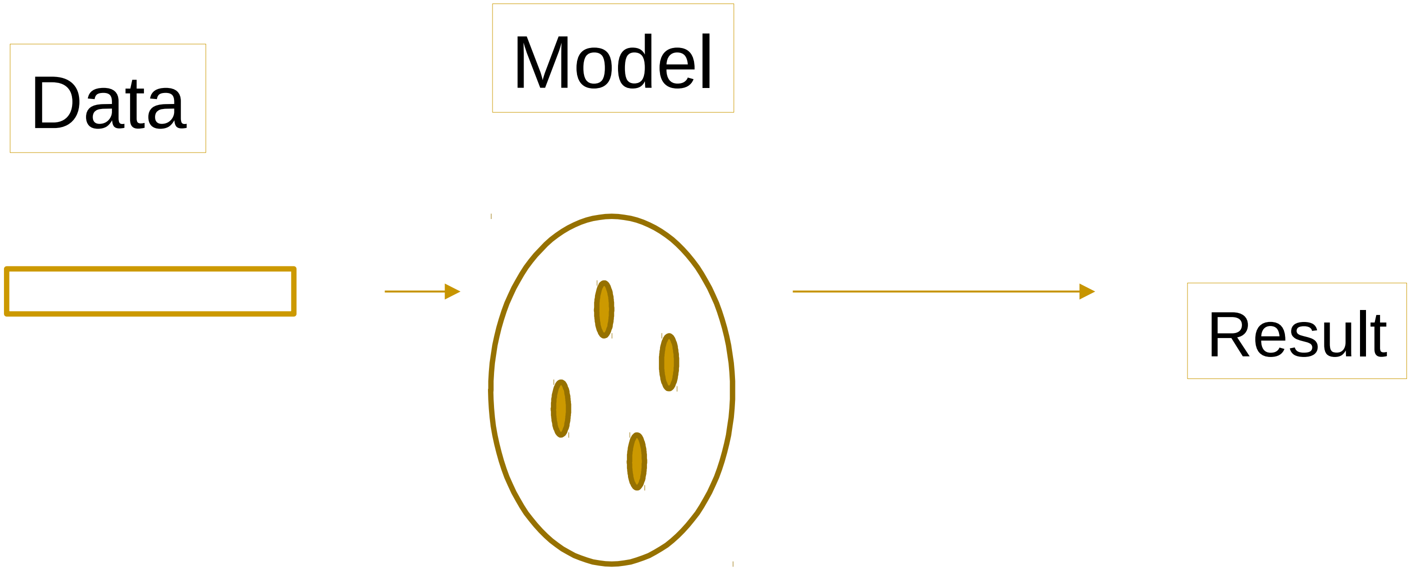


Data

Data Mining Algorithm

Model

Data Mining - Prediction



Data

Model

Result

Supervised v. Unsupervised Learning

- Supervised learning
  - For example, find the group of customers likely to leave after their contract expires.
  - It’s a predefined set.
  - We can use a set of existing training data
- Unsupervised learning
  - Do customers fall into natural groups
  - Not a predefined set
  - Training data not labelled with these groups.

Types of Problems

- Supervised Learning
  - Regression
  - Classification (also probability estimation)
- Unsupervised
  - Clustering

Regression

- Value estimation.
- Estimate the value for a particular attribute.
- For example predict tomorrows temperature at midday.
- Use historical data (trining set) about other instances (days).
- Supervised learning technique.

Classification

- The learning scheme/algorithm is presented with a set of classified examples.
- Expected to learn and be able to classify unseen examples.
- For example cancer diagnosis.
- Mostly assume that instances belong to one class.
- Includes Logistic Regression!

Classification & Prob. Estimation

- Logistic Regression produces numbers which are the probability of an instance being in a particular class.
- Produces a “Probabilistic Classifier”.
- Produces a model giving probability of an individual belonging to particular groups.

Clustering

- Find groups of instances that cluster or belong together.
- [Identify clusters with no specific purpose in mind.]
- Training set not labelled with groups. (Unsupervised technique)
- Success is often gauged subjectively by how meaningful the clusters are to users.

Clustering (cont)

- Can be followed by a classification step.
- The clusters are now treated as classes and results used as input to a classification algorithm.

Data Reduction/ Feature Selection

- Not covered in the module.
- So dont do it!!
- Throw away irrelevant features.
  - For clearer insight.
  - Better accuracy.
  - Performance reasons.
- In a lot of cases dont need to do this.
- Let the model use any features it wants.

Summary

- Supervised Learning
  - Regression
  - Classification
- Unsupervised Learning
  - Clustering