Machine Learning for Big Data

Uses for Machine Learning

❑Software

❑Stock Trading

❑Robotics

❑Medicine and Healthcare

❑Advertising

❑Retail and E-Commerce

❑Gaming Analytics

❑The Internet of Things

Languages for Machine Learning

❑Python

❑R

❑Matlab

❑Scala

❑Clojure

❑Ruby

Algorithm Types for Machine Learning

❑Supervised Learning

❑Unsupervised Learning

Supervised learning

❑ *Supervised learning* refers to working with a set of labeled training data.

❑ For every example in the training data you have an input object and an output object.

❑ An example would be classifying Twitter data.

- ❖ you have the following data from Twitter; these would be your input data objects:
  - ▪ Really loving the new St Vincent album!
  - ▪ #fashion I'm selling my Louboutins! Who's interested? #louboutins
  - ▪ I've got my Hadoop cluster working on a load of data. #data

Supervised Learning

❑Supervised learning requires the classification to know the outcome result of each tweet,

❑We have to manually enter the answers

❑ The resulting output object at the start of each line.

music	Really loving the new St Vincent album!
Clothing	#fashion I'm selling my Louboutins! Who's interested? #louboutins
bigdata	I've got my Hadoop cluster working on a load of data. #data

❑Training set is obtained that can be used for later classification of data

Supervised learning

❑Supervised learning problems can be grouped into regression and classification problems.

❑**Classification**: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.

❑**Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”

Analytical Tool - Weka

- Weka (Waikato Environment for Knowledge Analysis)
  - an open source data mining offering
  - fully implemented in Java
  - Primarily developed at the University of Waikato, New Zealand.
- It provides a suite of tools for learning and visualization via the supplied workbench program or the command line.
- Weka also enables you to retrieve data from existing data sources that have a JDBC driver.
- With Weka you can do the following:
  - Preprocessing data
  - Clustering
  - Classification
  - Regression

Another Analytical Tool : Mahout

- Mahout machine learning libraries are an open source project that are part of the Apache project.
    - The key feature of Mahout is its *scalability*
    - it works either on a single node or a cluster of machines.
- It has tight integration with the Hadoop Map/Reduce paradigm to enable large-scale processing.
- Mahout supports a number of algorithms including
    - Naive Bayes Classifier
    - K Means Clustering
    - Recommendation Engines
    - Random Forest Decision Trees
    - Logistic Regression Classifier

Another Analytical Tool: SpringXD

- ❑*Weka and Mahout* concentrate on algorithms and producing the knowledge we need,
- ❑we must also think about acquiring and processing data.
- ❑Spring XD is a “data ingestion engine”
  - ❖ It reads in, processes, and stores raw data. It’s highly customizable with the ability to create processing units.
- ❑Spring XD is relatively new, but it’s certainly useful.
- ❑ It not only relates to Internet-based data, it can also ingest network and system messages across a cluster of machines.

Another Analytical Tool:Hadoop

❑Hadoop is very good for processing Big Data, but it’s not a required tool.

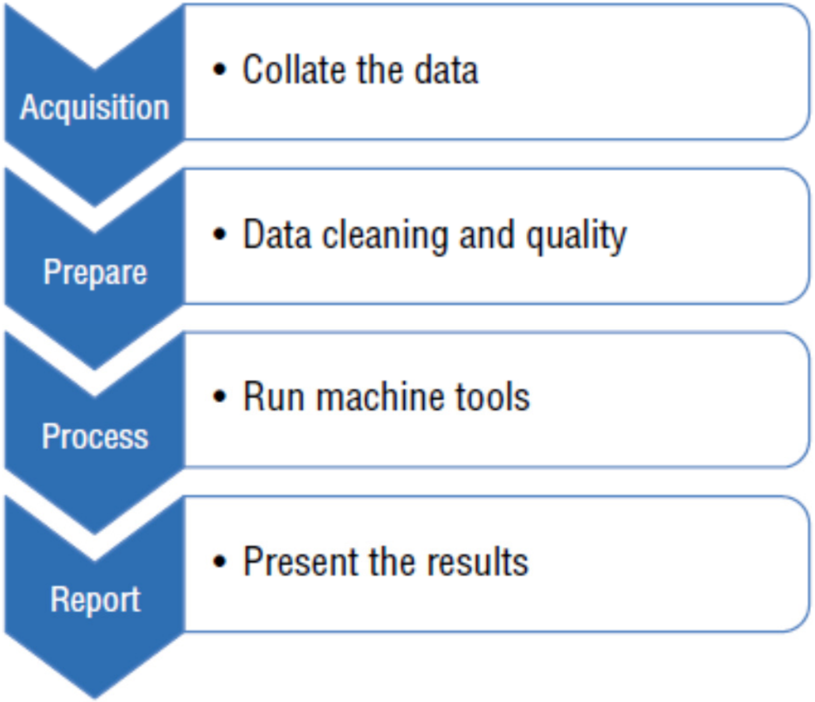
❑Hadoop is a *framework* for processing data in parallel.

❑It does this using the MapReduce pattern

❖where work is divided into blocks and is distributed across a cluster of machines.

❑ We can also use Hadoop on a single machine with success

Machine Learning Process



Acquisition

- Collate the data

Prepare

- Data cleaning and quality

Process

- Run machine tools

Report

- Present the results

Machine Learning Projects

- Machine learning projects start with a question
  - Is there a correlation between our sales and the weather?
  - Do sales on Saturday and Sunday generate the majority of revenue to the business compared to the other five days of the week?
  - Can we plan what fashions to stock in the next three months by looking at Twitter data for popular hashtags?

Decision Trees as a Predictive Model

❑Financial institutions use decision trees.

- ❖One of the fundamental use cases is in option pricing, where a binary-like decision tree is used to predict the price of an option in either a bull or bear market

❑In the medical field, decision tree models have been designed to diagnose blood infections or even predict heart attack outcomes in chest pain patients.

- ❖Variables in the decision tree may include *diagnosis, treatment,* and *patient data.*

Building a Decision Tree

Decision trees are built around the basic concept of this algorithm.

❖Check the model for the base cases.

❖ Iterate through all the attributes .

❖Get the normalized information gain from splitting on attribute.

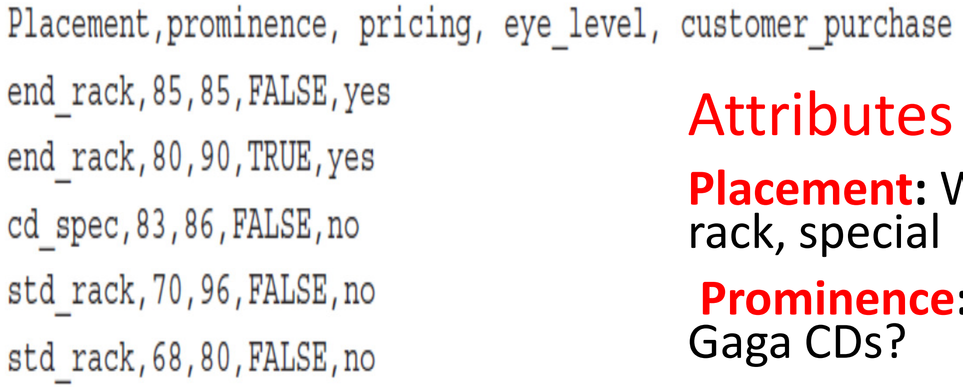
❖ Best attribute will be the attribute with the highest information gain.

❖Create a decision node that splits on the best attribute.

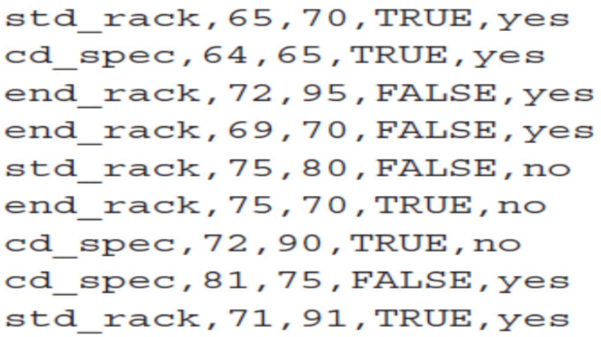
❖Work on the sublists that are obtained by splitting on best attribute and add those nodes as child nodes.

Training Data

The client has given us some in a .csv file



```
Placement,prominence, pricing, eye_level, customer_purchase
end_rack,85,85,FALSE,yes
end_rack,80,90,TRUE,yes
cd_spec,83,86,FALSE,no
std_rack,70,96,FALSE,no
std_rack,68,80,FALSE,no
```



```
std_rack,65,70,TRUE,yes
cd_spec,64,65,TRUE,yes
end_rack,72,95,FALSE,yes
end_rack,69,70,FALSE,yes
std_rack,75,80,FALSE,no
end_rack,75,70,TRUE,no
cd_spec,72,90,TRUE,no
cd_spec,81,75,FALSE,yes
std_rack,71,91,TRUE,yes
```

Attributes

Placement: What type of stand the CD is displayed on: an end rack, special offer bucket, or a standard rack?

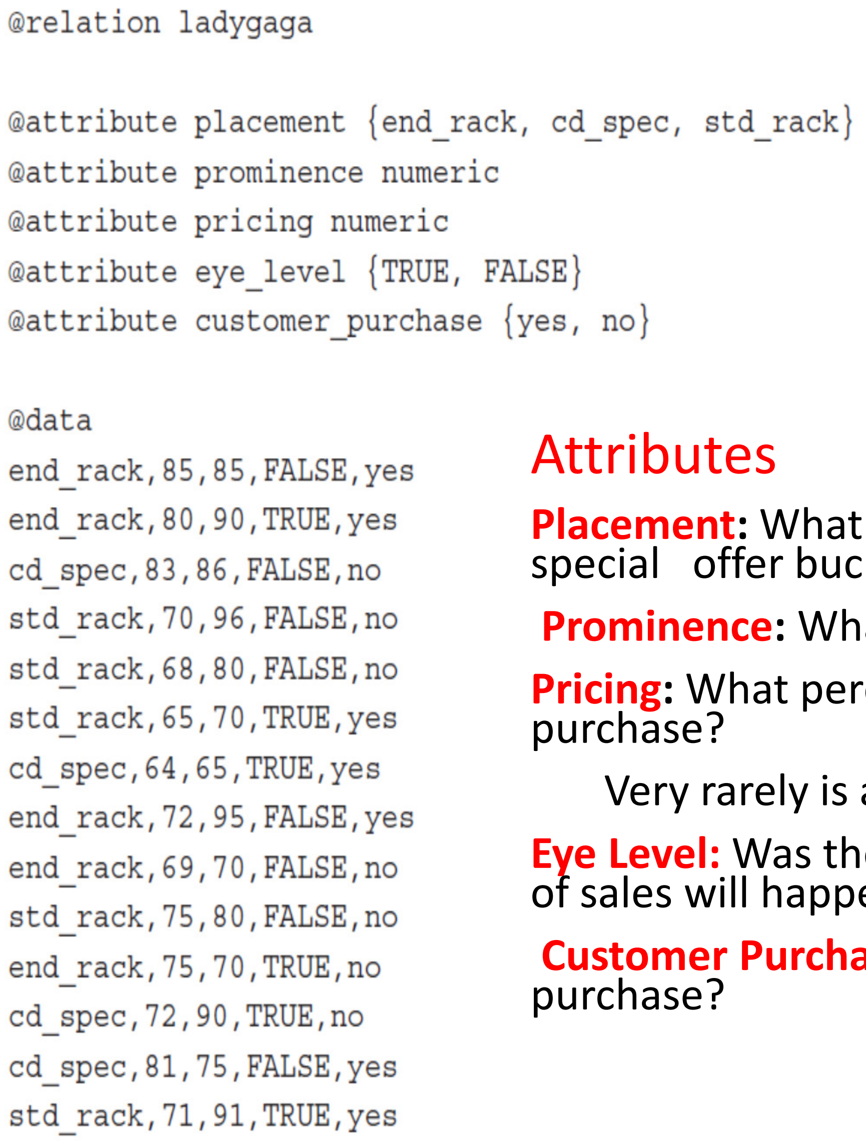
Prominence: What percentage of the CDs on display are Lady Gaga CDs?

Pricing: What percentage of the full price was the CD at the time of purchase?

Very rarely is a CD sold at full price, unless it is an old, back

Eye Level: Was the product displayed at eye level position? The majority of sales will happen when a product is displayed at eye level.

Customer Purchase: What was the outcome? Did the customer purchase?



```
@relation ladygaga

@attribute placement {end_rack, cd_spec, std_rack}
@attribute prominence numeric
@attribute pricing numeric
@attribute eye_level {TRUE, FALSE}
@attribute customer_purchase {yes, no}

@data
end_rack,85,85,FALSE,yes
end_rack,80,90,TRUE,yes
cd_spec,83,86,FALSE,no
std_rack,70,96,FALSE,no
std_rack,68,80,FALSE,no
std_rack,65,70,TRUE,yes
cd_spec,64,65,TRUE,yes
end_rack,72,95,FALSE,yes
end_rack,69,70,FALSE,no
std_rack,75,80,FALSE,no
end_rack,75,70,TRUE,no
cd_spec,72,90,TRUE,no
cd_spec,81,75,FALSE,yes
std_rack,71,91,TRUE,yes
```

Attributes

Placement: What type of stand the CD is displayed on: an end rack, special offer bucket, or a standard rack?

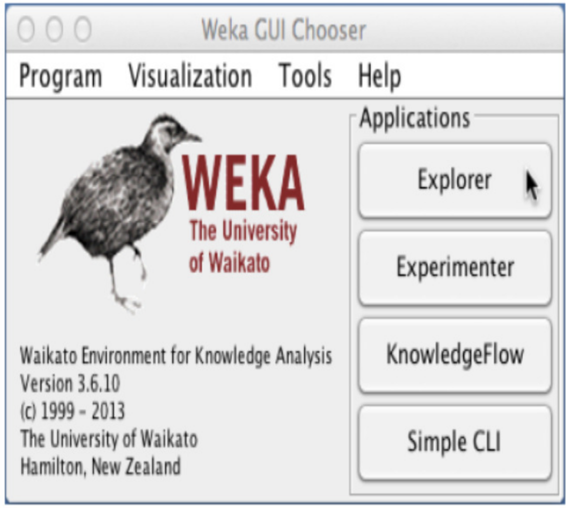
Prominence: What percentage of the CDs on display are Lady Gaga CDs?

Pricing: What percentage of the full price was the CD at the time of purchase?

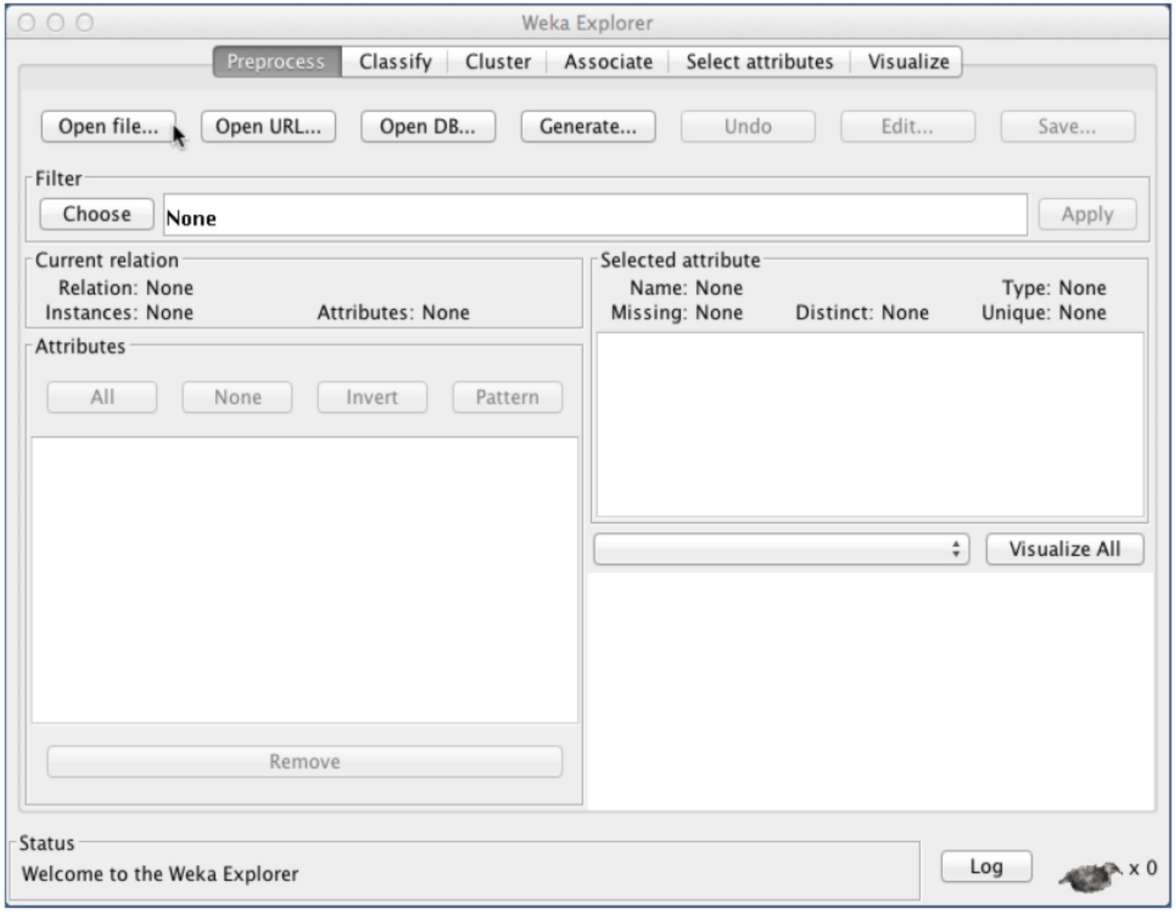
Very rarely is a CD sold at full price, unless it is an old, back

Eye Level: Was the product displayed at eye level position? The majority of sales will happen when a product is displayed at eye level.

Customer Purchase: What was the outcome? Did the customer purchase?



Waikato Environment for Knowledge Analysis
Version 3.6.10
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand



Filter: None

Relation: None
Instances: None
Attributes: None

Name: None
Type: None
Missing: None
Distinct: None
Unique: None

Welcome to the Weka Explorer

Weka saves the file as a .arff file to set up the attributes