Lecture 18: Data Mining

BADM/ACCY 352

Spring 2017

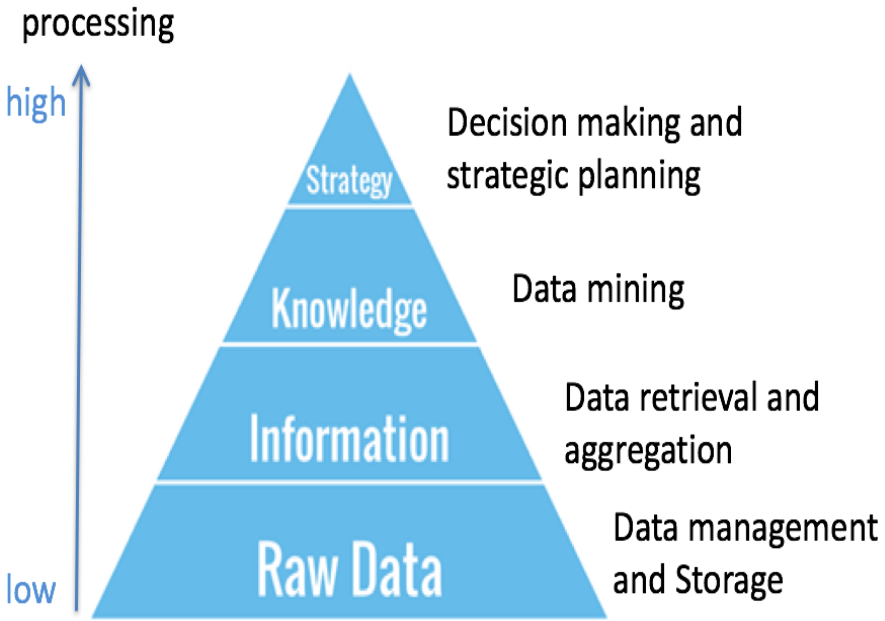
Instructor: Yi Yang, PhD

Last lecture

- NoSQL databases

This lecture

- Data mining
  - Association rule learning
  - Different types of learning



Data mining

- Extract and uncover useful information and knowledge from large volumes of data
- Changes the way you think about data and its role in business

Answering business questions with these techniques

- Who are the most profitable customers?
  - Database querying
- What are the similarities among these customers? Can I characterize them?
  - Data mining (automated pattern finding)
- Will some particular new customer be profitable? How much revenue should I expect this customer to generate?
  - Data mining (predictive modeling)

Beer and diaper

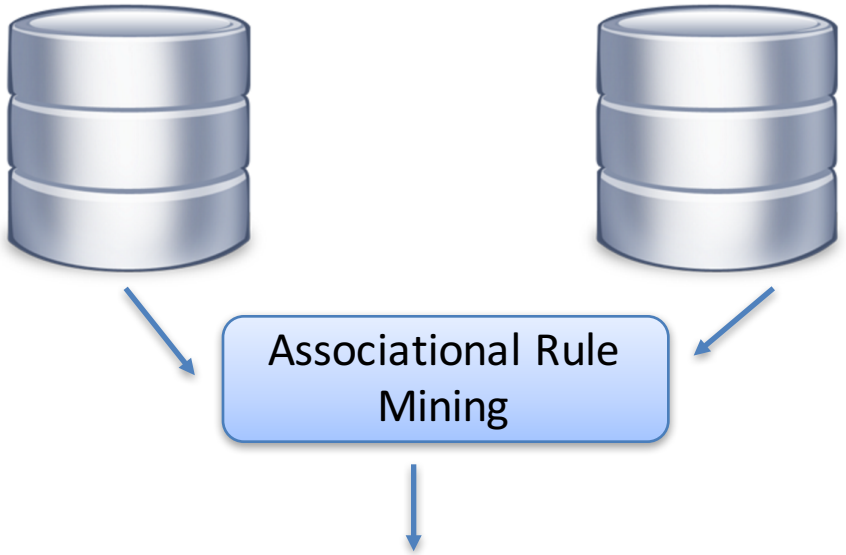
Walmart

Customer Database

(demographic data about its customers)

POS Database

(where, when and what those customers bought)



1. people who buy gin are also likely to buy tonic. They often also buy lemons.
2. On Friday afternoons, young American males who buy diapers also have a predisposition to buy beer.

Association Rule Learning

- AR learning is a method for discovering interesting relations between items in large database.
- Also called market basket analysis: what do my customers buy, which products are bought together?

- Goal: Find associations between the different items that customers place in their shopping basket

Formalization

- AR learning:
  - Transaction database **T**: a set of transactions **T** = {t_1, t_2, …, t_n}
  - Each transaction contains an itemset **I**, which is a collection of items **I** = {i_1, i_2, …, i_m}
- Goal:
  - Find frequent/interesting associations among sets of items in databases.
  - Represent the associations in rules: X => Y, where X and Y are both a set of items.

AR learning

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

An example of market basket transactions

Bread, Milk, Diapers, Beer, Eggs, Cola

For d items, the total number of rules is about 3^d

Two challenges:

- Discovering patterns from a large transaction data set can be computationally expensive
- Some of the discovered patterns are potentially spurious because they may happen simply by chance.

Note: An association rule does not necessarily imply causality. Instead, it suggests a co-occurrence relationship between items.

AR learning

- Compute two measures: *support* & *confidence*
- Support (s): number of transactions that contain both X and Y.

  s = #(X, Y) / # transactions.

- A rule will low support indicates the items are not bought together, or may occur by chance.

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

s(Beer => Diapers) = 3/5 = 0.6
s(Milk, Diapers => Beer) = 2/5 = 0.4

AR learning

- Confidence (c): how often items in Y appear in transactions that contain X

$$c = \#(X, Y) / \#(X)$$

- For a given rule X=>Y, the higher the confidence, the more likely it is for Y to be present in transactions that contain X.

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

c(Beer => Diapers) = 3/3 = 1.0
c(Milk, Diapers => Beer) = 2/3 = 0.67

AR learning

A common strategy in association rule learning algorithms has 2 steps:

1. Frequent Itemset Generation: find all itemset that are great than the minimum support threshold.
2. Rule generation: extract all high-confidence rules from the frequent itemset.

AR learning application

How consumers' buying behaviors relate to their support for president candidates.

- younger tech-savvy support base for Bernie Sanders is twice as likely to book travel reservations on Kayak or fly on budget carrier Spirit Airlines versus those who support Hillary Clinton.
- They're also twice as likely to tune in to "The Daily Show,"
- while Clinton backers are far more likely to use a Fitbit or other wearable device to track their activity levels.

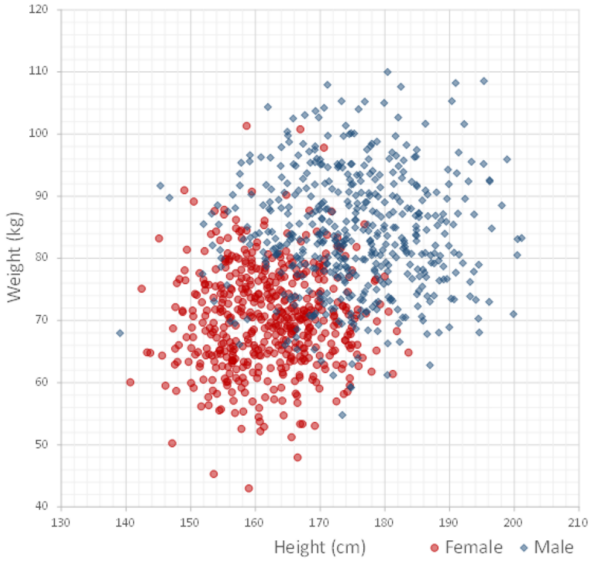
Data mining

- In addition to association rule learning, there are other types of learning methods:
  - Classification
  - Clustering
  - Regression
- ALL business areas can harness the power of big data and data mining to gain insight and knowledge

Classification

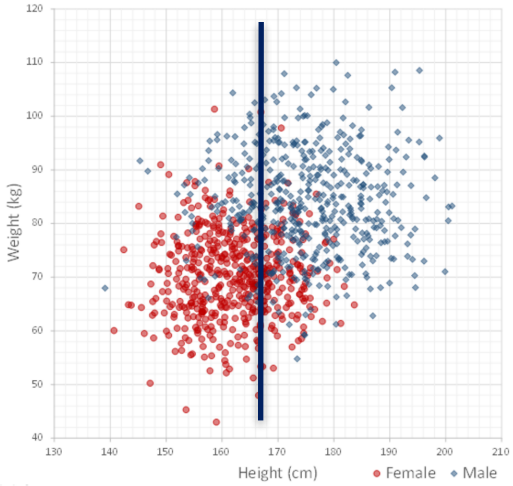
- It predicts, for each individual in a population, which class/category this individual belongs to.
- Eg:
  - Among all the customers, which are likely to respond to a given offer?
  - Is this transaction a fraud transaction?
  - Is this email a spam?
  - Is the student likely to get an A in this course?
  - Is the image cat, dog, or horse? (multi-class)

Classification

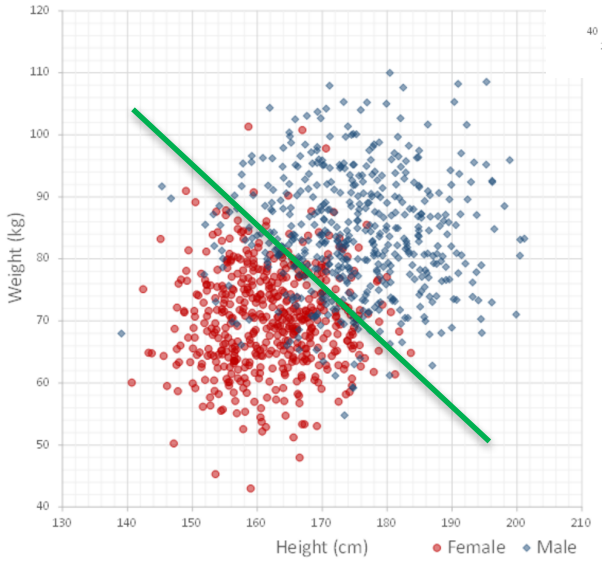


Data with class label

Linear Classification Algorithm



New data
(height = 200cm,
weight = 100kg).



male

Example: credit approval

Deciding approve an application or not?

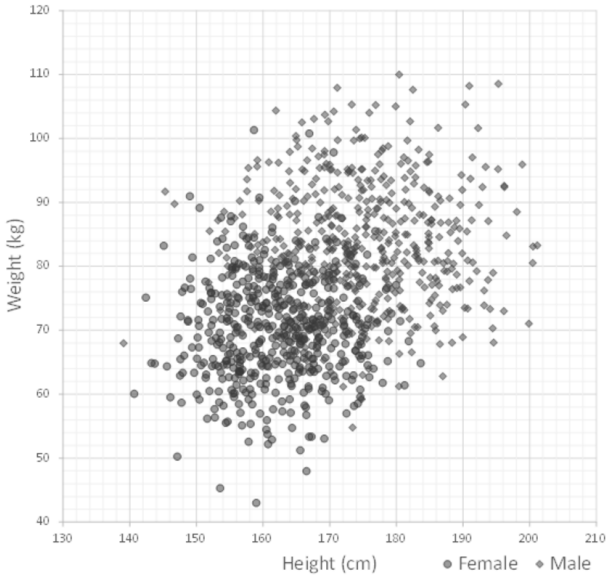
Application information:

item	value
age	23 years
gender	male
annual salary	$30,000
years in residence	1 year
years in job	1 year
current debt	$1000
…	…

Regression

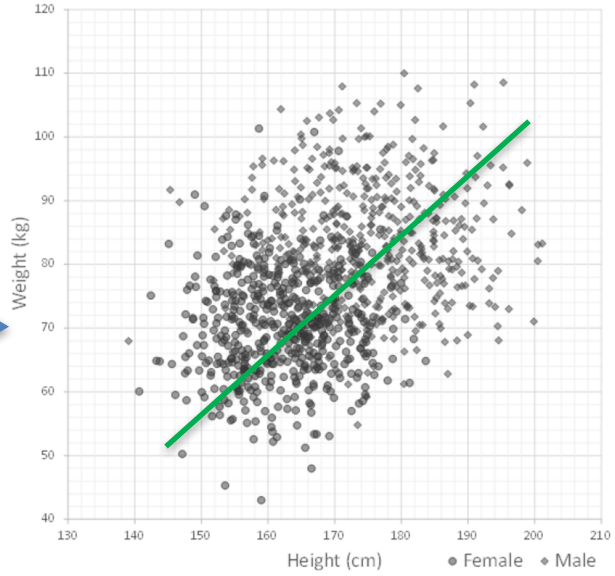
- It attempts to estimate or predict, for each individual, the numerical value of some variable for that individual.
- For example:
  - how much will a given customer use the service?
  - how likely will a given voter support a candidate?
  - how much will the product demand change if the price increases by 3%?

Regression



Regression Algorithm

New data
height = 200cm



weight = 100kg

Deloitte.

Western Australia Rental Prices

Predict rental prices for properties across Western Australia

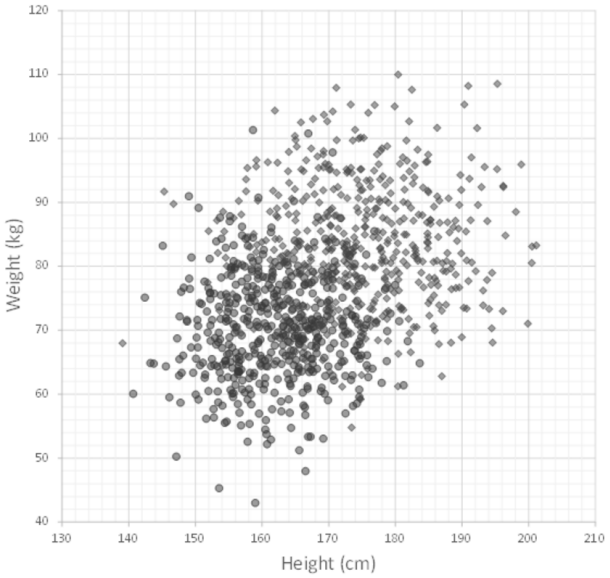
19 hours
59 teams
$100,000

Using data on location, property, zoning, past sales, and more, the goal of this competition is to improve on existing models by accurately estimating the weekly market rental value for residential properties across Western Australia.

Clustering

- It groups objects into different sets so that the objects in the same group are more similar to each other than to those in other groups.
- Example:
  - market segmentation. It groups customers, with purchasing power, according to their similarity in several dimensions (age, gender, income, education..) related to a product.

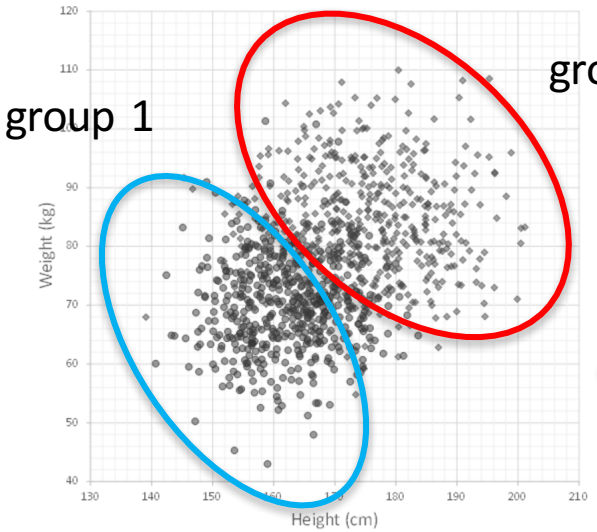
Clustering



Data with no class label

Clustering Algorithm

group = 2



New data
(height = 200cm,
weight = 100kg).

Belongs to group 2

Summary

- Data mining in business intelligence.
- Association rule learning
- Different types of data mining