PSS718 - Data Mining

Lecture 6 - Building Models

Asst.Prof.Dr. Burkay Genç

Hacettepe University

October 30, 2016

What is modelling?

Modelling

Modelling is the process of taking some data (usually) and building a simplified description of the processes that might have generated it.

Types of data analytics

- Data analytics can be studied under two categories:
  - Descriptive: cluster analysis, association rules
  - Predictive: classification, regression

Building Models

- Building models is a common pursuit in life
- We build models everyday
- Architects and engineers build models to see how things fit, how they work, or even to sell ideas to others

Basic to complex

- We start with a simple model
- See how it fits real life
- We then extend our model to explain/capture more
- Our models are based on data and hence objective
- Models in other fields may be more subjective

Models in different professions

- A computer program is a model
- An accountant's spreadsheet is a model
- A social media application is a model
- Econometric models capture events in economy
- Environmental models picture changes in the environment

Perfect model?

- No model can perfectly represent real world
- ... unless it is really simplistic and trivial
- The real world has so many visible and invisible actors that it is impossible to incorporate every one of them into our model
- Hence our approach is to approximate the real world

Model builders

Rattle supports a number of model builders:

- Clustering
- Association rules
- Decision tree induction
- Random forests
- Boosted decision trees
- Support Vector Machines
- Logistic regression
- Neural networks

What is it?

Descriptive analytics is the task of providing a representation of the knowledge discovered without necessarily modelling a specific outcome.

- Cluster analysis
- Association
- Correlation analysis
- Pattern discovery

Why to do it?

- Similar to "unsupervised learning" in machine learning
- Identify patterns in the data that extend our knowledge and understanding of the world that the data reflects
- There is generally no specific target variable that we are attempting to model

What is it?

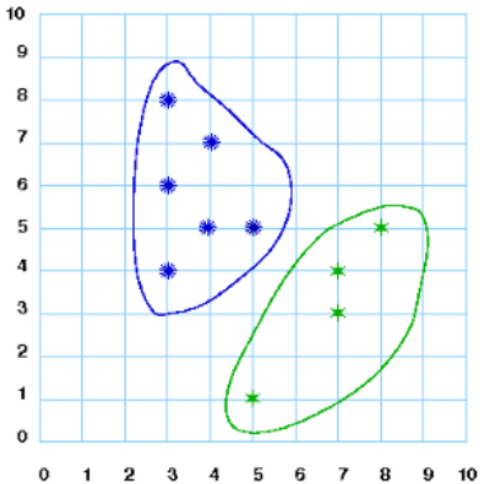
Build a model that can be used to predict the occurrence of an event.

1. Historical data
   - A -> observe -> X
   - B -> observe -> Y
   - C -> observe -> X
2. Model
   - A -> model -> X
   - B -> model -> Y
   - C -> model -> X
3. Predict
   - D -> model -> Y
   - E -> model -> X

Types of predictive models

- Classification models: tries to classify an observation into one of two or more categories, such as *Yes* or *No* for *RainTomorrow*
  - Usually expressed as a series of tests
  - Did it rain today? Is it hot today? Then, it will rain tomorrow.
- Regression: tries to produce a numeric outcome for a given observation, such as the amount of rain for tomorrow
  - Usually expressed as a formula based on the input variables
  - $Y = X_1 + 0.2X_2 - 0.4X_3$

What is it?



Definition (Clustering)

Group observations in a generally unguided fashion according to how similar they are based on a measure of the distance between observations.

- One of the core tools of any data mining project
- Allows the data miner to break data into more meaningful groups and then contrast the different clusters against each other
- Can also be useful in grouping observations to help make the smaller datasets easier to manage

k-means algorithm

A model built using the k-means algorithm represents the clusters as a collection of k means. The observations in the dataset are associated with their closest "mean" and thus are partitioned into k clusters.

Example

```
> library(rattle)
> set.seed(42)
> obs1 <- sample(1:nrow(weather), 5)
> vars <- c("MinTemp", "MaxTemp",
            "Rainfall", "Evaporation")
> cluster1 <- weather[obs1, vars]
```

```
> mean(cluster1)

    MinTemp     MaxTemp    Rainfall Evaporation
       4.74       15.86        3.16        3.56
```

```
> obs2 <- setdiff(sample(1:nrow(weather), 20), obs1)
> cluster2 <- weather[obs2, vars]
> mean(cluster2)

    MinTemp     MaxTemp    Rainfall Evaporation
     6.6474     19.7579      0.8421      4.4105
```

A naive approach

- Create all possible sets of k clusters
  - Each clustering corresponds to a candidate model
- Find a way to measure the "goodness" of each clustering
- Choose the clustering with the best score

Warning

How many *possible sets of k clusters* exist?

Search heuristic

- In general, the above approach is not viable
- Instead, the k-means algorithm uses a search heuristic
  - Start with a random clustering
  - Re-associate each observation with the closest cluster
  - Re-compute cluster means
  - Repeat until stabilized

How to measure?

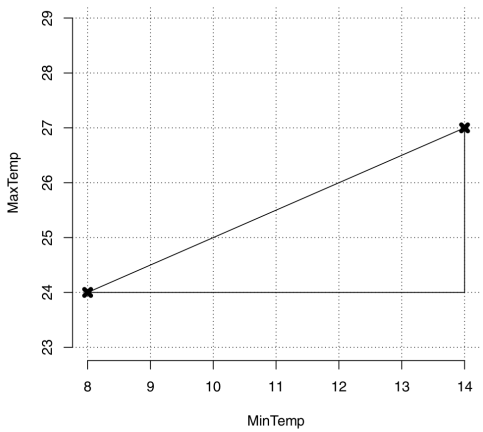
- We need to measure the **distance** (*dissimilarity*) between two observations
- Any measurement method should satisfy the following:
  - Distance is nonnegative: $d(a, b) \geq 0$
  - Distance to self is 0: $d(a, a) = 0$
  - Distance is symmetric: $d(a, b) = d(b, a)$
  - Triangular inequality: $d(a, b) \leq d(a, c) + d(c, b)$

Minkowski Distance

Minkowski Distance

The Minkowski Distance between points a and b is

$$d(a,b) = \sqrt[q]{|a_1 - b_1|^q + |a_2 - b_2|^q + |a_3 - b_3|^q + \ldots + |a_n - b_n|^q}$$



- When $q = 1$ -> Manhattan distance
  $d_M(a,b) = |a_1 - b_1| + |a_2 - b_2| + \ldots + |a_n - b_n|$
- When $q = 2$ -> Euclidean distance
  $d_E(a,b) = \sqrt{|a_1 - b_1|^2 + \ldots + |a_n - b_n|^2}$

Scale is a problem

Warning!

Do not forget that different scales of variables is a problem. If necessary, rescale your variables prior to computing distances.

Other variable types

- Minkowski Distance works fine for numeric variables
- Other variable types need special attention
  - Binary
  - Categorical
  - Mixed

Binary Variables

- Binary variables have two values: 0 and 1
- Consider the following table for two binary input variables x_i and x_j

	1	0	sum
1	a	b	a+b
0	c	d	c+d
sum	a+c	b+d	n

- Simple matching coefficient (symmetric variable)

$$d(x_i, x_j) = \frac{b + c}{a + b + c + d}$$

- Jaccard coefficient (asymmetric, 1 is more important)

$$d(x_i, x_j) = \frac{b + c}{a + b + c}$$

Categorical Variables

- Generalization of binary variables
- Simple matching

$$d(x, y) = \frac{n - p}{n}$$

where n is the total number of variables and p is the matched categorical variables between x and y.

- Recode into *indicator variables*: Create a binary variable for each level of the categoric variable

Mixed variables

- A dataset may contain many types of variables
- Use a weighted formula to combine the different normalized distances, where the weights are used to express the relative importance of the variables:

$$d(x_i, x_j) = \sum_k w_k d_{ij}^{A_k}$$

where w_k is the weight of variable A_k, $d_{ij}^{A_k}$ is the dissimilarity of the i^{th} observation and the j^{th} observation on variable A_k. $d_{ij}^{A_k}$ is normalized to $[0, 1]$.

Types of quality

- Our goal is to obtain
  - high intra-class similarity
  - low inter-class similarity
- Within Sum of Squares: sum the square of the distances between the observations within a cluster, and total this up for each of the clusters.
- Between Sum of Squares: sum the squares of the distances between the observations in different clusters

Measuring quality

- Using withinss to measure intra-class similarity

```
model$withinss
```

```
##  [1] 172.1 219.2 237.6 217.2 336.6 228.4 254.2 291.9 310.5 126.0
```

```
model$tot.withinss
```

```
## [1] 2394
```

- Using betweenss to measure inter-class similarity

```
model$betweenss
```

```
## [1] 3446
```