Introduction to Data Science

Lecture 03; January 31st, 2013

Ernst Henle
ErnstHe@UW.edu
Skype: ernst.predixion

Agenda

- Social Interactions
  - LinkedIn (UW – Data Science)
    - Homework. I encourage you to collaborate on homework assignments
    - Last week some students organized an after class social meeting
- Homework and Review
  - R
    - Cleaning Data
    - Methods for Cleaning
  - Predixion Install; Please notify me if you haven't been able to install Predixion Insight.
- Predictive Analytics
  - Terminology
  - Supervised vs. Unsupervised
- MATLAB (GNU-Octave)
  - K-means
- AWS codes (you should have received a code for $100.00 of AWS time)

Homework Review (R, Data Cleaning)

- Remove aberrant values
  - Coerce a character vector or a factor to a numeric vector

    as.numeric(x) # convert a character vector

    as.numeric(levels(x)[x]) # convert a factor
  - Remove the NAs:

    v1 <- v1[complete.cases(v1)] # remove NAs from a vector

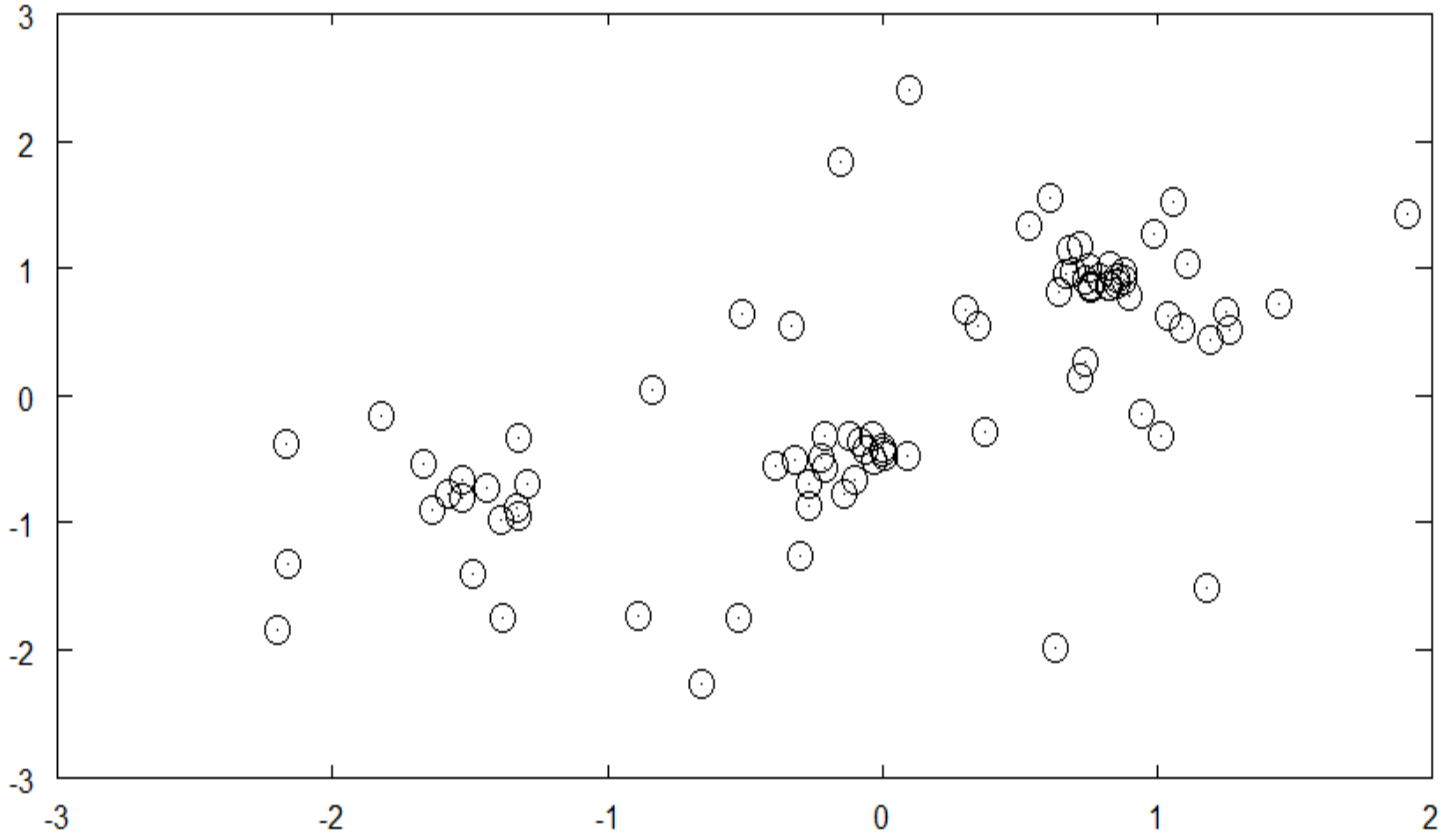
    df <- df[complete.cases(df),] # remove Nas from a dataframe
- Remove Outliers
  - Reset values that are outside of normal distribution
    - Set to limits: Low values are set to lowest allowed value; high values are set to highest allowed value
    - Set to average
    - Set to NA
  - Remove values that are outside of set limits

IntroductionToR3.R

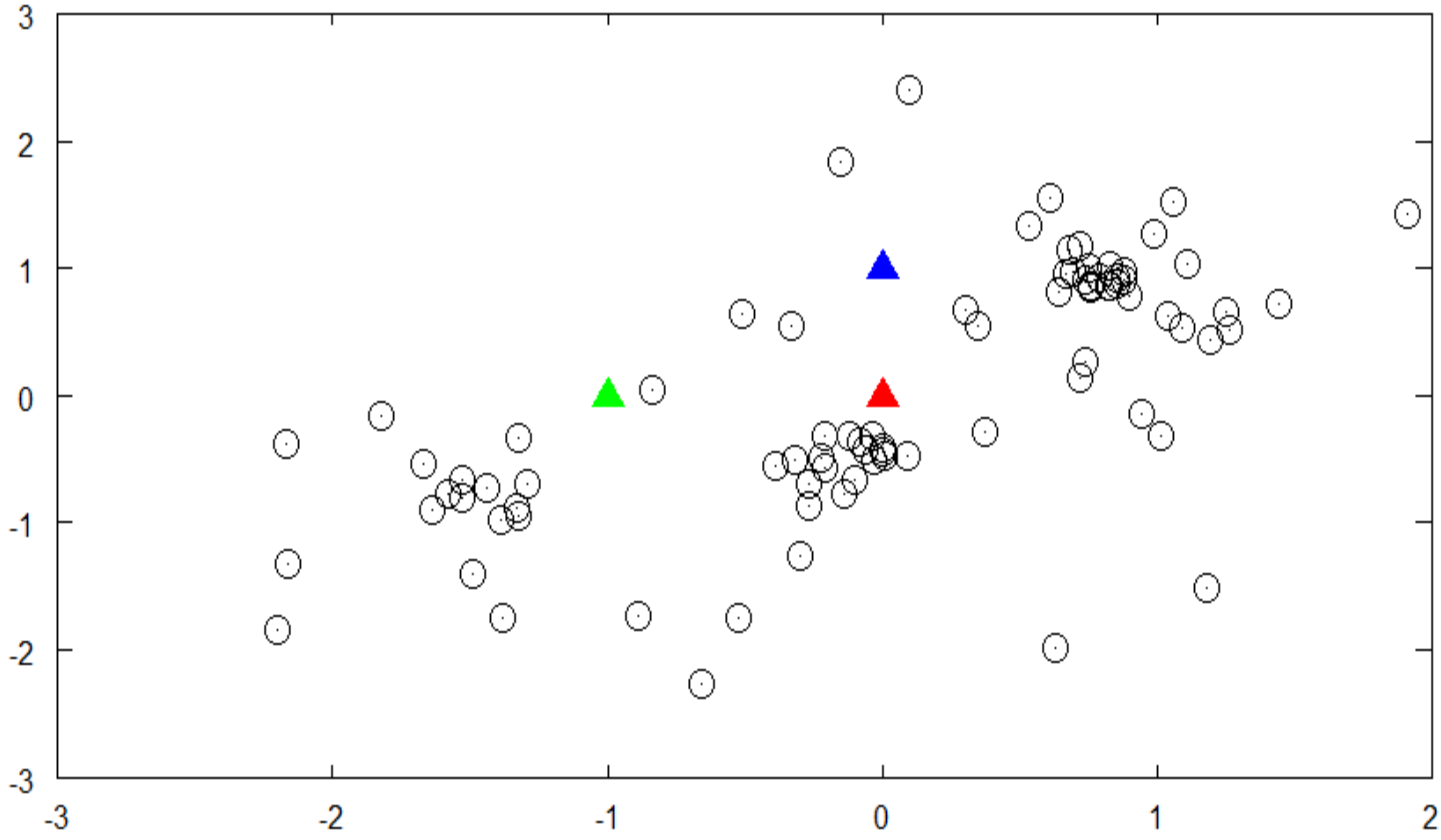
K-means clustering: Algorithm

- Pre-requisites
  1. Get points in multi-dimensional space.
     - table, matrix, rectangular dataset
  2. Specify the number of clusters
     - Weakest point in algorithm (makes algorithm non-deterministic)
  3. Get a random center for each cluster
     - Another weak point in the algorithm
- Repeat until convergence:
  1. For each point, determine its closest cluster center and assign that point to that cluster
  2. Determine the centroid (mean) of the cluster

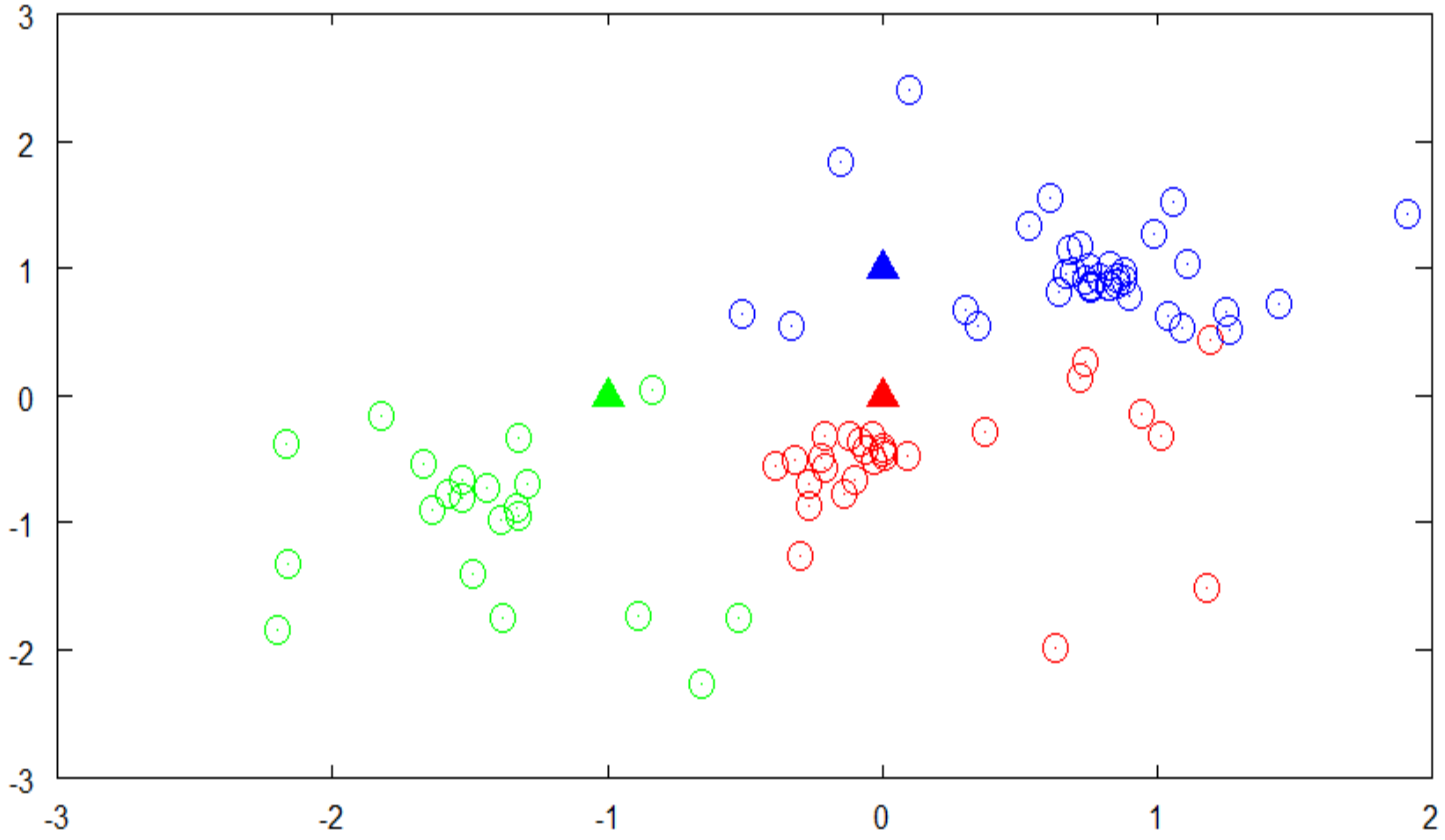
K-means clustering: Points



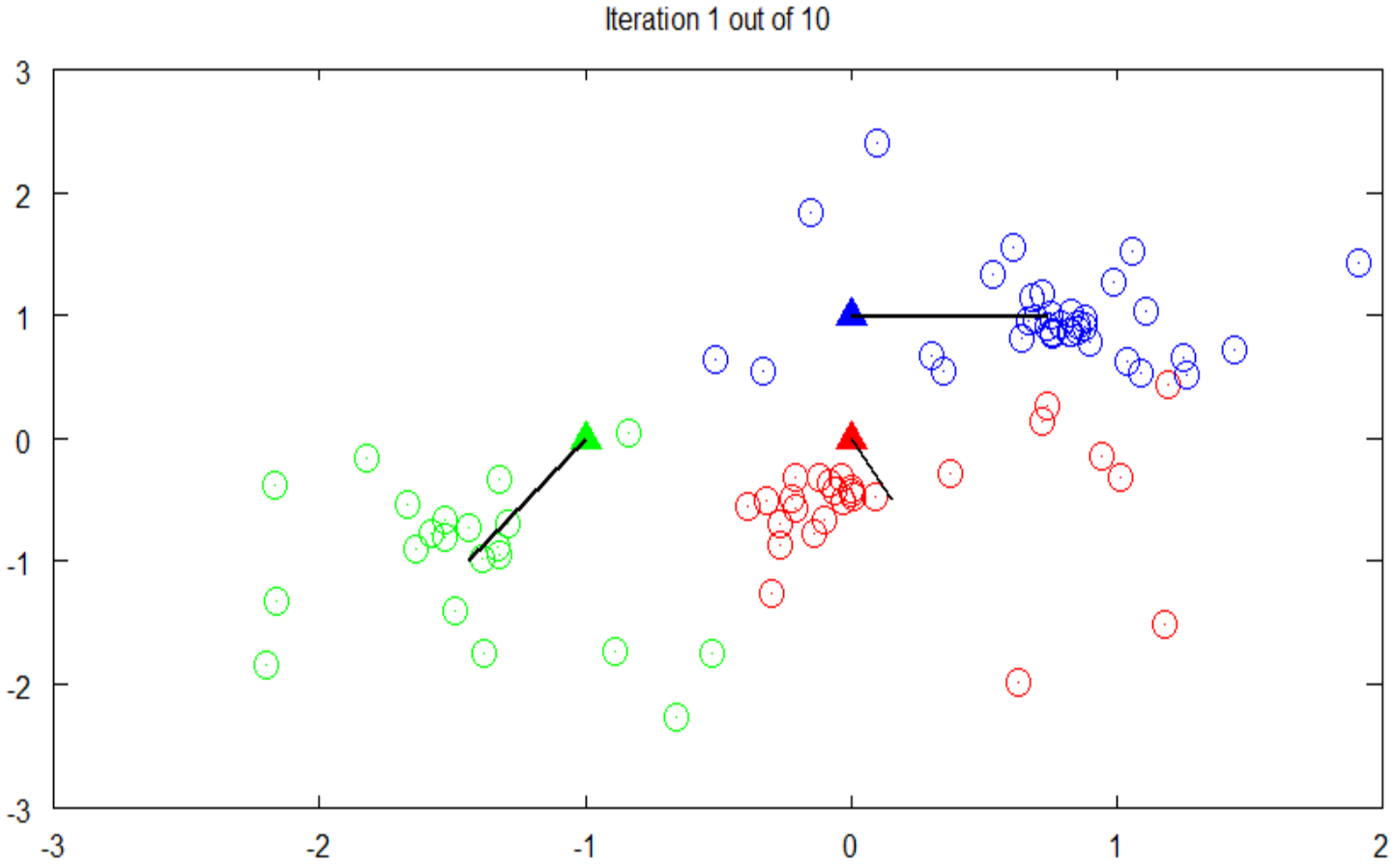
K-means clustering: Cluster Centers (Guess at Centroids)



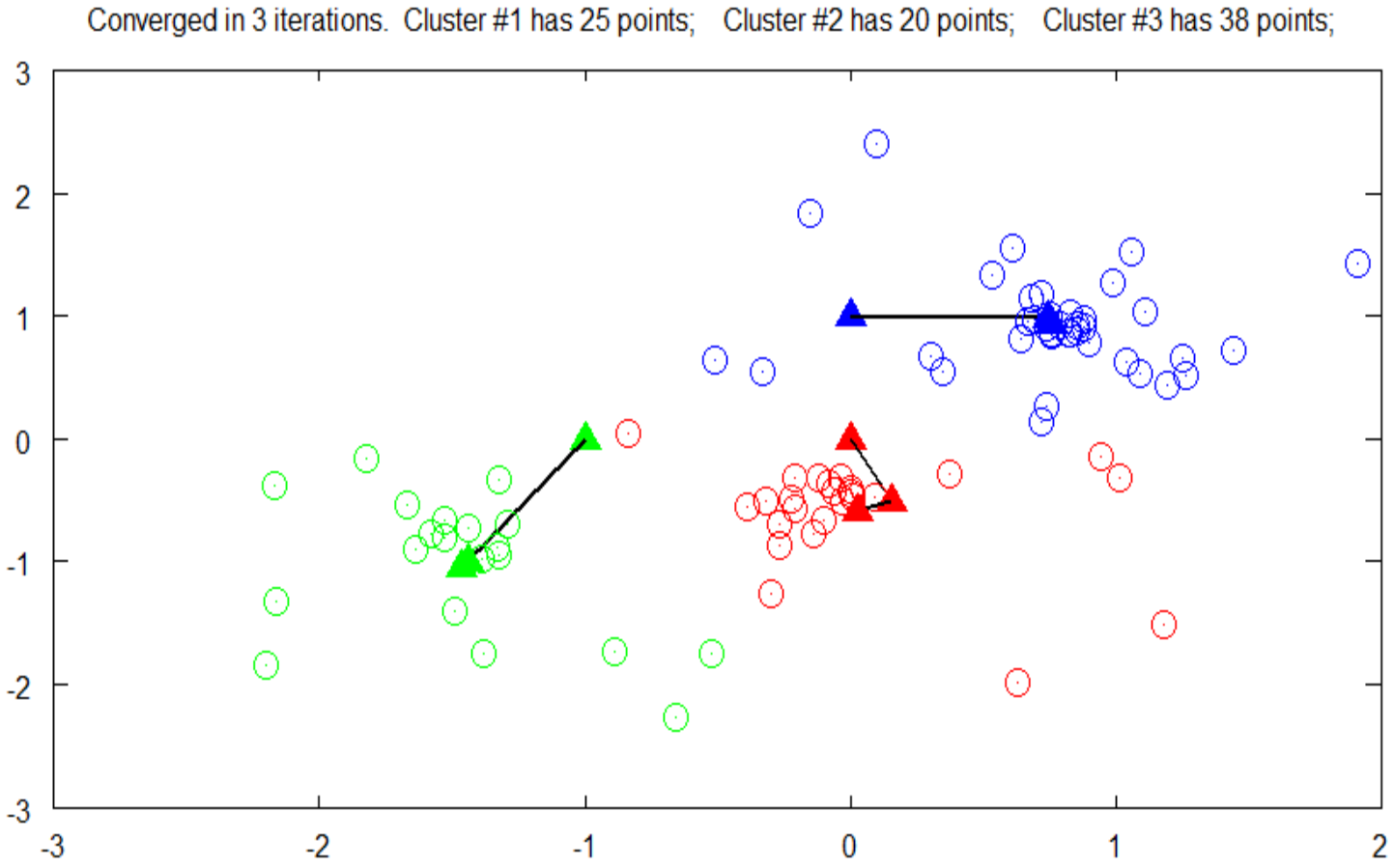
K-means clustering: Cluster Assignments



K-means clustering: Determine Cluster Centroids



K-means clustering: Converged



GNU-Octave (1)

- Assignments
  - a = 17; % simple assignment of a scalar
  - a % without the semicolon, the result appears in the console
  - a = [11, 19, 23]; % create a vector using[]
  - a
  - a(2) % index using ()
  - a = ‘Hello World’; % assignment of characters
  - a % simplest Hello World
  - a(7:9) % ‘Hello World’ is a character array
  - [a(7:9) a(11)]
- In Octave all variables are matrices
  - a = 19;
  - size(a)
  - a
  - a(3,2) = 11; %
  - a’

GNU-Octave (2)

- Create a 1D matrix (aka vector)
  - x = -3:0.1,3;
  - y = exp(-x.*x);
  - plot(x, y)
- Some common operator symbols are for matrix algebra
  - '*' matrix multiplication; Use ".*" for element-by-element multiplication
  - '^' matrix power; Use".^" for element-by-element power
  - '/' matrix right division; Use "./" for element-by-element division

Predictive Analytics

- Predictive Analytics (Police and Amazon.com)
  - http://www.youtube.com/watch?v=brGfLspBAj8
- Trendologist L. Vaughan Spencer on BizIntelligence.TV
  - http://www.youtube.com/watch?v=w4mISk-vMZY&feature=youtu.be
- Why DFD's should be kept simple (Dilbert)
  - http://www.youtube.com/watch?v=_HOUs9cF5uo
- Colbert on Predictive Analytics
  - http://www.colbertnation.com/the-colbert-report-videos/408981/february-22-2012/the-word---surrender-to-a-buyer-power

Rectangular Data (1)

- The data set has columns and rows. Each cell has a value or is null.
- A Rectangular dataset is often called a matrix, data frame, or table.
- Example usage: classifications and estimations

Rectangular Data (2)

- Columns have descriptive headers like: Name, Age, Height, Weight of each student.
- Columns are also called attributes and fields.
- All values within a column have the same data type

Rectangular Data (3)

- Rows generally do not have names. If a row has a name, then the names could be considered another column.
- Rows are also called observations or cases
- The number of rows in a category is called support.

Rectangular Data (4)

ID	IQ	Parent Income	Moral Support	Gender	College Plans
835	107	40,000	Yes	Female	Applied
016	99	53,000	Yes	Male	Applied
490	105	60,000	No	Male	Did not apply

Terminology and Concepts (1)

- Data
  - Dataset is a set of Data. A set implies a commonality. The commonality is expressed as a type or a relation.
  - A data type provides structure and meaning to the data. Just like there is no such thing as un-structured data, there is no such thing as un-typed data. Data can be insufficiently typed and structured.
- Rectangular Data
  - Datasets are often 2D matrices, which are organized into rows and columns. The column and row order is not important .
  - Columns are named with a header; A columns may be also referred to as an attribute or field. The number of columns is often called the dimensionality of the data.
  - Rows are not named. A rows is often referred to as a case or observation. Number of rows in a category is called support.
- Data dimensionality
  - A data frame or a table can be considered a sparse multi-dimensional matrix
  - The dimensionality for un-supervised learning is #columns
  - The dimensionality for supervised learning is #columns - 1 because one column represents the value and not the dimension. This structure is very similar to a star schema

Terminology and Concepts (2)

- Algorithm
  - Numerical recipe; A procedure; Set of instructions;
- Hypothesis
  - A statement. Predicts an outcome (not always correctly). An untested guess.
- Theory
  - A theory is a hypothesis that has been tested and verified. A theory correctly accounts for the available observations.
- Model
  - A hypothesis that explains some observations. A theory based on a subset of data

Terminology and Concepts (3)

- Predictive Analytics (Machine Learning)
  - Supervised Learning: The model needs to be trained. In other words, you have to tell the machine the results in a training set. Then, the machine will learn apply the same pattern to data that do not contain the results.
    - Classification
    - Estimation
  - Unsupervised Learning: The model does not need to be trained. In other words, you don't have to tell the machine the results
    - Clustering
    - Association
    - Market-basket analysis
    - Anomaly detection

Assignment

1. If you did not do this last week then explain:
   - Why is normalization important in K-means clustering? Provide an example that shows how non-normalized data will lead to errors.
   - How do you encode categorical data in a K-means clustering? E.g. consider an attribute with 2 categories like “Female” and ”Male”. Consider an attribute with 4 categories like “Bellevue”, “Capitol Hill”, “Tacoma”, “Sam Mateo”
   - Why is clustering un-supervised learning as opposed to supervised learning?
2. Octave: Add code to simpleKMeans that normalizes the input by range (min max)
   - Normalization by range works like this:

     (x – min(x))/(max(x) – min(x))
   - Add similar code to simpleKMeans.m to normalize the points
   - Normalization is based on points but is applied to both points and centroids. Apply the normalization to the centroids too.
   - When the clustering is complete, you have cluster centroids that are in normalized space. Add code to the end of simpleKMeans.m that will de-normalize these centroids to their original space. (Undo the normalization)
3. Figure out how to create a Windows 2008 R2 virtual machine with SQL Server 2008 R2
   - http://aws.amazon.com/windows/
   - http://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/Welcome.html
   - http://awsdocs.s3.amazonaws.com/EC2/latest/ec2-wg.pdf
   - http://www.youtube.com/watch?v=uNuhyumosOA1
   - http://www.youtube.com/watch?v=PpWwruWPInY
   - http://www.youtube.com/watch?v=3xEge7Sfzfg
   - https://aws.amazon.com/amis?platform=Windows&selection=platform
4. Submit the completed assignment in Catalyst by Sunday evening.

Preview

- Relational Algebra and Calculus
- Relational Model and Associate Model
- http://sentences.com/docs/amd.pdf
- http://en.wikipedia.org/wiki/Relational_model
- http://www.youtube.com/watch?v=NvrpuBAMddw