Introduction to Data Science

Lecture 05; February 7th, 2013

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

Agenda

- Social Interactions
  - LinkedIn (UW – Data Science)
  - I request more collaboration on homework assignments
    - What do you think about pairing of students to do homework?
- Homework and Review
  - K-means
  - MATLAB (GNU-Octave)
    - Normalization
  - AWS
- Scripting Exercises
  - Data Clean up in R (MPG data)
  - Clustering of MPG data in MATLAB
  - Clustering of MPG data in Predixion Insight
- Time permitting: Mid Course Review
- Time permitting: Classification in Predixion Insight

Homework: Normalization

- Why is normalization important in K-means clustering?
  - Normalization scales the dimensions so they have similar ranges. Non-normalized clustering would always favor separating the points along the dimension that has the largest range.
- Linear normalization
  - Optional: Offset adjustment (e.g. subtract minimum or mean)
  - Required: Scaling (Divide by range or standard deviation)
  - Range:
    - Xnorm <- (Xorig – Xmin) / range
    - Where: range = Xmax - Xmin
  - Z-score:
    - Xnorm <- (Xorig – Xmean) / $\sqrt{\sigma^2}$
    - here: $\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$
- Some considerations
  - Outliers may thwart purpose of normalization so outlier removal may be necessary
  - Non-linear and non symmetric distributions may require non-linear normalizations like Log, exp, Square Root, Square

Homework: Category → Binary (1)

Binary Categories

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

Homework: Category → Binary (2)

Category

ID	Location
835	San Mateo
016	Bellevue
490	Bellevue
835	Tacoma
016	Capitol Hill
490	Tacoma

Homework: Category → Binary (3)

Category

ID	Location
835	San Mateo
016	Bellevue
490	Bellevue
835	Tacoma
016	Capitol Hill
490	Tacoma

→

Binary

ID	Bellevue	Capitol Hill	Tacoma	San Mateo
835	0	0	0	**1**
016	**1**	0	0	0
490	**1**	0	0	0
835	0	0	**1**	0
016	0	**1**	0	0
490	0	0	**1**	0

Homework: (Un)-Supervised

- Training explains the difference between supervised and non-supervised learning. Supervised learning derives patterns from a training set where the answers are given. Unsupervised learning does not need to be trained. There are deterministic and non-deterministic examples of both supervised and unsupervised learning.

Homework: MATLAB

```
% Parameters for normalization and denormalization
   minPoint = min(points);
   maxPoint = max(points);
   range = maxPoint - minPoint;

   % Normalize points
   numberOfPoints = size(points, 1);
   % for each point:
   % subtract away its minimum and then divide by the range
   for (pointNo = 1:numberOfPoints)
      points(pointNo, :) = (points(pointNo, :) - minPoint)./range;
      % y                              = (x                 -  m)            / r
   end
```

See: SimpleKmeans.m

Scripting Exercise

- R
  - IntroductionToR4.R
  - Clean up data from:
    - http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data-original
  - Relabel Column Values
  - Binarize Categories
  - Write out modified file with predictive columns
- MATLAB
  - KMeansDemo3
  - Read cleaned data file
  - Cluster Data
    - Original space
    - Normalized Space
- View Data in Excel
- Read Data and Cluster using Predixion Insight

Mid Course Review

- Data Flow (DFD)
- Scripting Languages
  - R
  - MATLAB
- Data Preparation
  - Data typing
  - Data coercion
  - Missing Data (Incomplete cases)
  - Outlier recognition and removal
  - Relabeling
  - Normalization
  - Replacing Categories with Boolean columns
- Machine Learning
  - Predictive Analytics Overview
  - K-means in depth
- Social Interactions

Assignment

1. Lab 3a and 3b for Predixion Insight. Use AWS and Office 2010 Pro trial version if you need a compatible machine.
   - Make Screen shot of final cost comparison for all methods: Logistic Regression, Neural Net, Naïve Bayes, and Decision Trees. Submit the completed assignment to Catalyst by Monday evening.
2. Review terminology in slide titled: "Assignment Terminology"
3. Review Relational Model, Relational Algebra, and Calculus
   - http://sentences.com/docs/amd.pdf (Pages 35 to 48 only)
   - http://en.wikipedia.org/wiki/Relational_model
   - http://www.youtube.com/watch?v=NvrpuBAMddw

Assignment: Terminology

- Algorithm
- Anomaly detection
- Association
- Attribute
- Binarize Categories
- Binary Column
- Case
- Category Column
- Character Column
- Classification
- Clustering
- Coercion
- Column
- Column Header
- Data
- Data Dimensionality
- Data Frame
- Data Type
- Dataset
- DFD
- Estimation
- Field
- Hypothesis
- Key Column
- Machine Learning
- Market-basket analysis
- MATLAB
- Matrix
- Missing Data
- Model
- Multinomial Column
- Normalization
- Numeric Column
- Observation
- Outcome
- Outlier Removal
- Predictive Analytics
- R
- Rectangular Data
- Relabeling
- Row
- Sparse Multi-Dimensional Matrix
- Standard Deviation
- States
- String
- Supervised Learning
- Supervised Learning
- Support
- Table
- Target Column
- Text Column
- Theory
- Un-structured Data
- Unsupervised Learning
- Z-score