

Data Mining: Introduction

Mike Bowles, PhD

Patricia Hoffman, PhD

<http://machinelearning2010fall.pbworks.com/w/page/30032895/FrontPage>

<http://patriciahoffmanphd.com/>



Machine Learning Consulting / Corporate Training

Available for Consulting - Corporate Training

Currently Teaching at Hacker Dojo

Next Class Starts in January

<http://machinelearning2010fall.pbworks.com/w/page/30032895/FrontPage>

Machine Learning

Statistics

- ◆ Networks – Graphs
- ◆ Weights
- ◆ Learning
- ◆ Generalization
- ◆ Supervised Learning
- ◆ Unsupervised Learning

- ◆ Large Grant = \$1,000,000
- ◆ Conference: French Alps

Snowbird, Utah

- ◆ Models
- ◆ Parameters
- ◆ Fitting
- ◆ Test Set Performance
- ◆ Regression / Classification
- ◆ Density Estimate/Clustering

- ◆ Large Grant = \$50,000
- ◆ Conference:

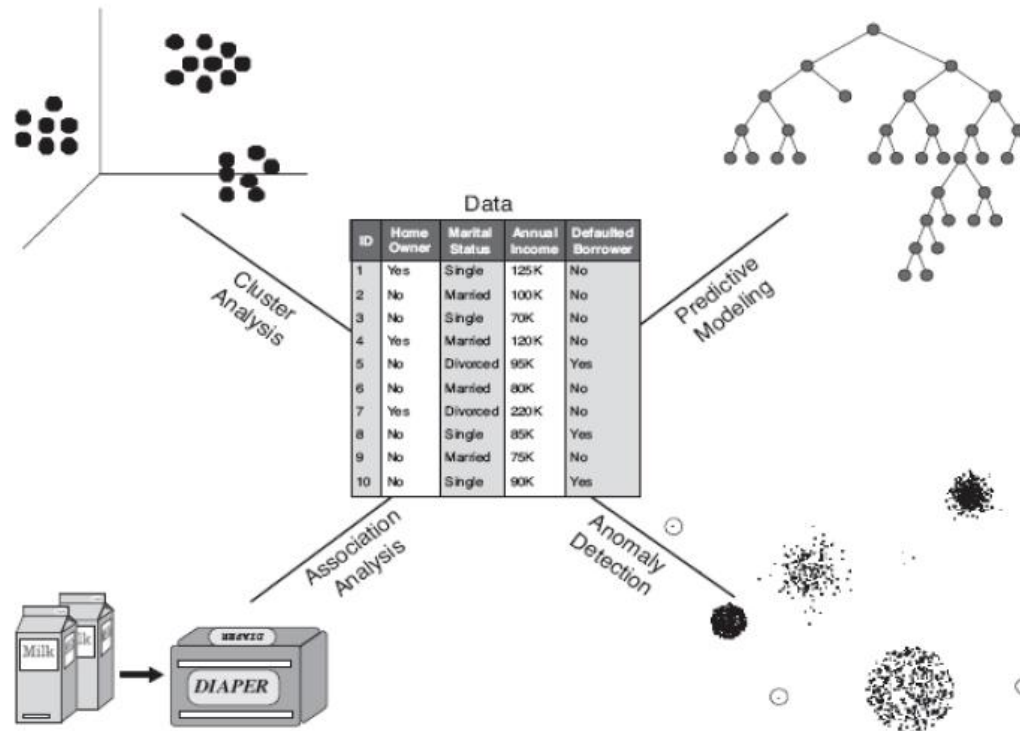
Los Vegas in August

Quote from Professor Robert Tibshirani, PhD

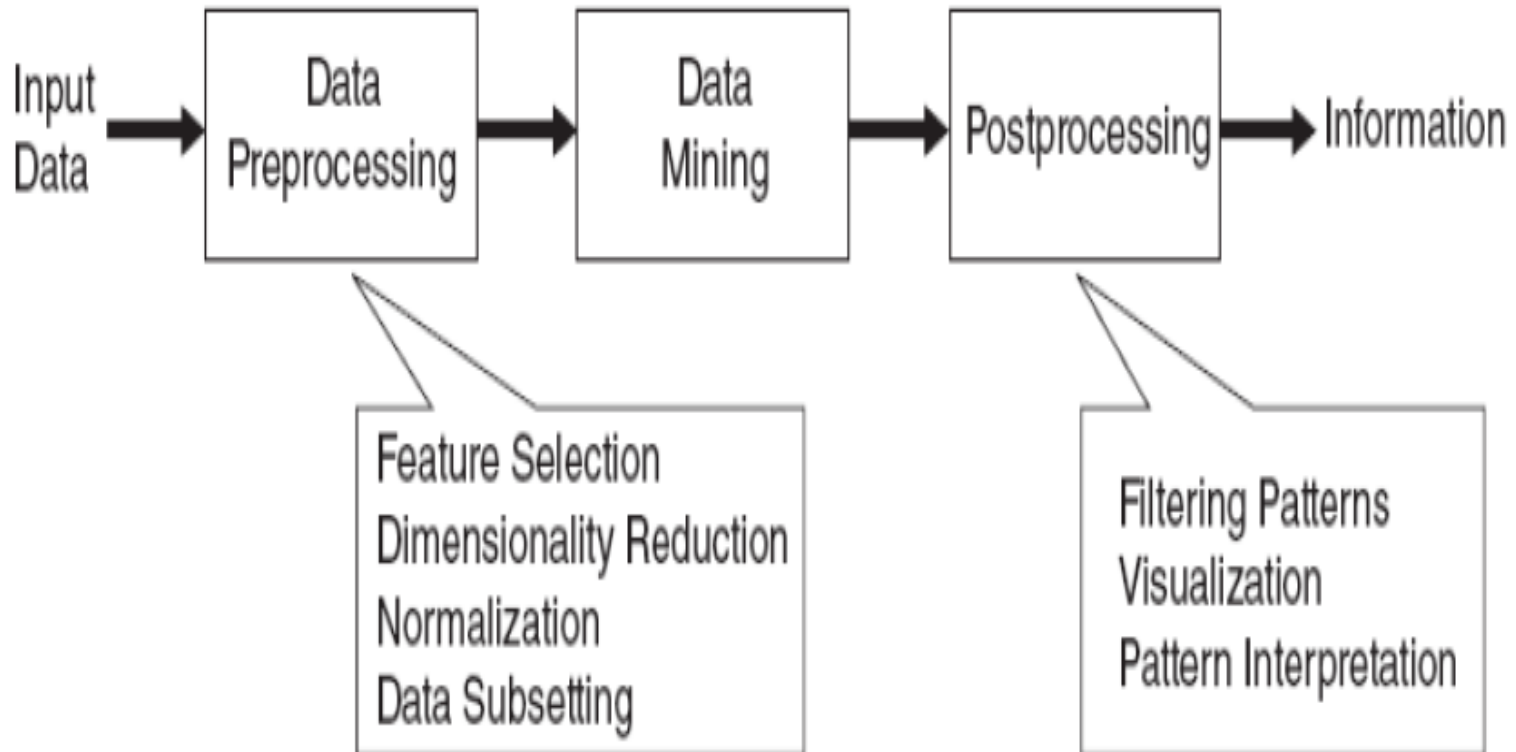
Machine Learning Examples

Problem Domain	Application	Input Pattern	Pattern Classes
Bioinformatics	Sequence analysis	DNA/Protein sequence	Known types of genes/ patterns
Data mining	Searching for meaningful patterns	Points in multi- dimensional space	Compact and well- separated clusters
Document classification	Internet search	Text document	Semantic categories (e.g., business, sports, etc.)
Document image analysis	Reading machine for the blind	Document image	Alphanumeric characters, words
Industrial automation	Printed circuit board inspection	Intensity or range image	Defective / non-defective nature of product
Multimedia database retrieval	Internet search	Video clip	Video genres (e.g., action, dialogue, etc.)
Biometric recognition	Personal identification	Face, iris, fingerprint	Authorized users for access control
Remote sensing	Forecasting crop yield	Multispectral image	Land use categories, growth pattern of crops
Speech recognition	Telephone directory enquiry without operator assistance	Speech waveform	Spoken words

Data Mining Tasks



Knowledge Discovery Process



Wine Classification Problem

- ◆ This data set comes from UCI Machine Learning Repository
 - Standard Source of Data for Testing Algorithms
- ◆ The red wine dataset has 1599 examples.
 - Each example has numerical values for 11 measured attributes and
 - Quality score assigned by wine tasters
- ◆ If we had the measured attributes could we
 - Stuff predict the score a taster would give the wine?



Example 1:
Inspect the Red Wine Data Set



Regression

- ◆ For each output (“quality” for the wine data set) we have a corresponding set of inputs (alcohol, sulphites, etc.).
- ◆ Arrange the outputs into a column vector $\mathbf{Y} = [y_1, y_2, \dots, y_n]^T$
- ◆ Define a matrix $\mathbf{X} = [\mathbf{1} : (X_1, X_2, \dots, X_m)^T]$

$$\mathbf{X} = \begin{array}{c} | \\ | \\ | \\ | \\ | \\ | \end{array} \begin{array}{ccccc} 1 & X_{11} & X_{12} & \dots & X_{1m} \\ 1 & X_{21} & X_{22} & \dots & X_{2m} \\ 1 & \cdot & \cdot & & \cdot \\ 1 & \cdot & \cdot & & \cdot \\ 1 & X_{n1} & X_{n2} & \dots & X_{nm} \end{array} \begin{array}{l} | \\ | \\ | \\ | \\ | \\ | \end{array}$$

Normal Equation Solution

Find best function (\mathbf{X} is the input $p \times N$ matrix)

$$\tilde{y} = \mathbf{X}\beta$$

Observations: $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p)$

Regression Coef: $\hat{\beta} = (\beta_0 \dots \beta_p)$


$$\text{minimize RSS } \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2$$

$$\text{Solution: } \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$



Ordinary Least-Squares Regression





Example 2: Predict Quality Score
Ordinary Least Squares Regression

Will OLS work for classification?

- ◆ OLS worked okay for numeric input and numeric output. What about if the output is binary – a classification problem?
- ◆ Sonar Data – mines vs rocks
 - 208 instances,
 - 97 instances of rocks
 - 111 from a metal cylinder approximating a mine.
 - Chirped sonar,
 - 60 measure of signal strength.
- ◆ Classify as rock vs mine.




Example 3: Classify Sonar Data using OLS Regression




Cross-Validation

We used 78/208 of our data to check for overfitting.
Here's an orderly process
to incorporate testing alongside training.





Example 3 Continued: Using Cross-Validation



What to do about Over-fitting

- ◆ Sonar data showed that we can reduce over-fitting with more data
- ◆ But we ran out. Now what?
- ◆ How can we systematically control model complexity?
- ◆ We'll explain a technique called “ridge regression” provides control

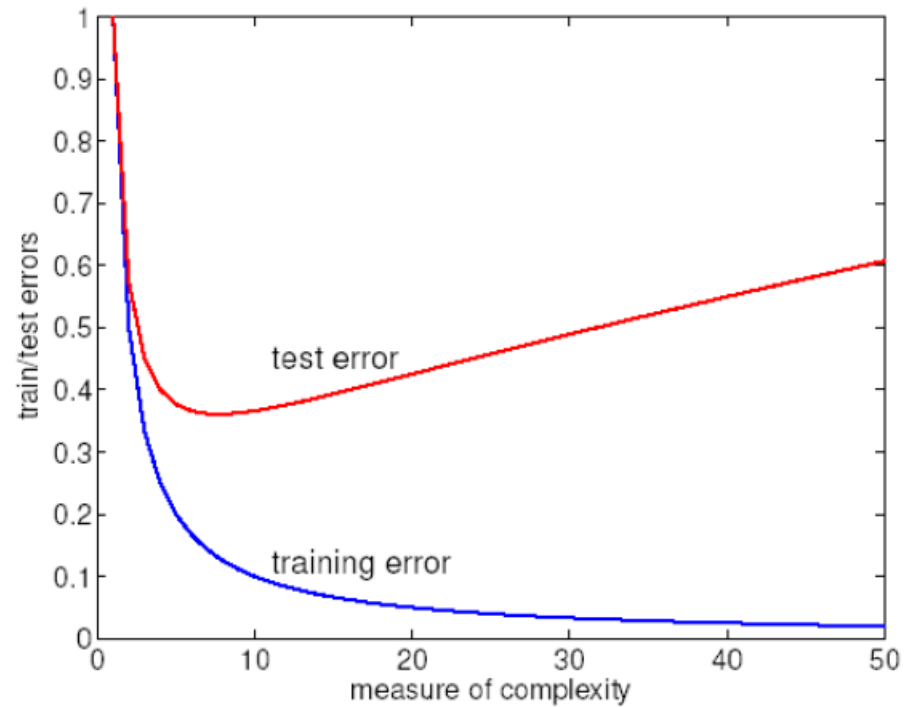
Ridge Regression

- $\lambda \geq 0$ complexity parameter controls shrinkage.
- $\lambda = 0 \Rightarrow$ solution is the same as regular regression.
- λ penalizes the sum-of-squares of the parameters.

Minimize

$$\sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

Complexity – Model Fitting





Example 4: Ridge Regression Sonar Data



Multiclass Problem

- ◆ The Target $Y = \{a, b, c, \dots\}$
 - multiple values – Iris Data
- ◆ One-Against-Rest (1-r)
 - Solve a separate binary classification problem for each possible classification
- ◆ Test Instances Classified
 - combine predictions from binary classifiers
 - voting scheme or probability estimate

Multi-class Problems

With two-class data we could use linear regression by calling $\text{class1} = 1$ and $\text{class2} = -1$

What about data that fall into more than one class?

For example the iris data

One way is “one versus the rest”



Example 5: Using regression for Iris Data



Wrap up

- ◆ If have questions feel free to contact us
- ◆ Join us at Hacker Dojo class in January
- ◆ Eager to help with Startups
- ◆ Provide Consulting
- ◆ Provide Training