An Introduction to Data Mining

for Wind Power Management

Spring 2015

The University of Iowa

Big Data World

Every minute:

- Google receives over 4 million search queries
- Facebook users share almost 2.5 million pieces of content
- Instagram users post almost 220,000 photos
- 72 new hours of video are uploaded to YouTube
- Twitter users tweet 300,000 times

Every day:

- we create over 2.5 quintillion bytes of data (2.5×10^{18} bytes)

90% of data in the world today was created in the last two years

The University of Iowa

What is Data Mining?

"The process of discovering patterns in data." – (Witten et al.)

Data mining provides the capabilities to:

- predict the outcome of a future observations
- uncover relationships between data attributes (features)
- uncover relationships between observations
- learning how to best react to situations through trial and error (reinforcement learning)

The University of Iowa

Related Fields and Disciplines

- Machine Learning
- Computational Intelligence
- Big Data
- Knowledge Discovery
- Artificial Intelligence
- Data Analytics
- Computer Science
- Engineering
- Statistics
- Mathematics
- Data Visualization
- High Performance Computing

The University of Iowa

Applications of Data Mining

- Medical Diagnostics
- Speech Recognition
- Market Prediction
- Sports
- Fraud Detection
- Online Dating
- Credit Worthiness
- Surveillance
- Cosmology
- Law Enforcement/NSA
- DNA Sequence Mapping
- Equipment Condition Monitoring
- Pharmaceutical Research
- Sales Prediction
- Product Recommendation
- Image Recognition

The University of Iowa

Applications in Wind Industry

Wind Power Forecasting

- Long term forecasts for planning
- Medium and short term forecasts for power generation commitment
- Time series, neural networks, fuzzy intelligent systems

The University of Iowa

Applications in Wind Industry

Wind Power Firming

- Securing availability of wind power source at a defined output level and time duration
- Uses energy storage system to capture excess energy to release later, or
- Uses gas system that can be deployed quickly
- Avoids rapid voltage and power swings on grid

The University of Iowa

Types of Learning

Supervised Learning
- Learns from labeled data
- Regression
- Classification

Unsupervised Learning
- Finds groups or structure of unlabeled data
- Clustering
- Dimensionality reduction

Reinforcement Learning
- Learns best reaction through trial and error

The University of Iowa

Common Algorithms

- Neural Networks
- Decision Trees
- K-Nearest Neighbor
- K-Means Clustering
- Support Vector Machines
- Extreme Learning Machines
- Naïve Bayes
- Logistic Regression
- Principle Component Analysis

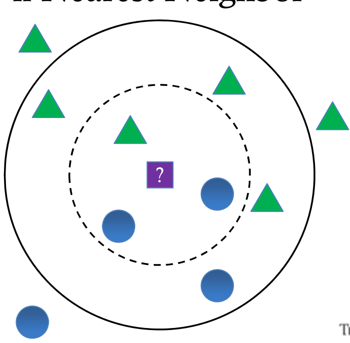
THE UNIVERSITY OF IOWA

k-Nearest Neighbor

- Type of supervised learning
- Can be used for classification or regression
- One parameter to tune for (k)
- Parameter k is an odd number
- Can suffer from curse of dimensionality in high dimensional space

THE UNIVERSITY OF IOWA

k-Nearest Neighbor



THE UNIVERSITY OF IOWA

k-Nearest Neighbor

Must define what is meant by “nearest”
- $d(x, y)$ should be large for dissimilar objects

- Numbers
  - Real number
  - Binary number
- Strings
- Images
- Videos
- Documents

THE UNIVERSITY OF IOWA

k-Nearest Neighbor

Non negativity
- $d(x, y) \geq 0$

Isolation
- $d(x, y) = 0 \;\; if \; x = y$

Symmetry
- $d(x, y) \geq d(y, x)$

Triangle Inequality
- $d(x, y) \leq d(x, z) + d(z, y)$

THE UNIVERSITY OF IOWA

k-Nearest Neighbor

- Euclidean distance
  - $d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$
- Manhattan distance
  - $d(x, y) = |(x_1 - y_1)| + |(x_2 - y_2)| + \cdots + |(x_n - y_n)|$
- Minkowski distance
  - $d(x, y) = (|(x_1 - y_1)|^p + |(x_2 - y_2)|^p + \cdots + |(x_n - y_n)|^p)^{1/p}$
- Hamming distance
  - Number of positions of two strings that are different (if strings are of equal length)
- String edit distance

THE UNIVERSITY OF IOWA

Example

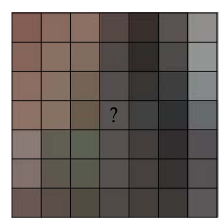
Skin Detection

- Objective: predict skin/nonskin for each pixel in testing set using training set of known classes

THE UNIVERSITY OF IOWA

Data Features

- A skin pixel is adjacent to other skin pixels
- A surrounding 7x7 grid is considered for each pixel
- The RGB color values for each pixel in the grid is assembled as a feature set
- 7x7x3 = 147 total features for each pixel
- The pairwise Euclidean distance is used to determine the k-nearest neighbors



THE UNIVERSITY OF IOWA

k-Nearest Neighbor

- Easy to understand and implement
- Highly nonlinear separator
- Only two parameters to tune (distance and k)
- Model can be updated easily
- Sensitive to noise or irrelevant attributes
- Expensive testing of each instance because all pairwise distance must be calculated to determine k-nearest neighbors
- The model is the data set and must be stored for future predictions

Decision Tree Learning

- Can be used for classification or regression
- In classification, leaves of tree represent predicted class
- Decisions trees are often used for decision making representation, but decision tree learning results in a prediction (that can be used for decision making)
- Interior nodes filter features until a leaf node is reached and a prediction is made

Example

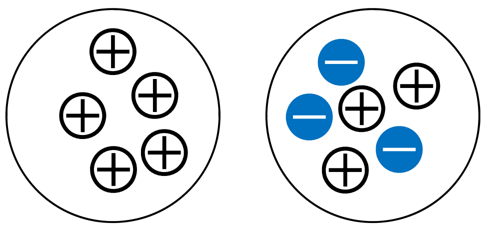
Decision Tree Learning

- Learning is achieved by splitting the training set into subsets based on an attribute value test
- At each step, a variable is chosen to "best" split the set
- One criteria for "best" is information gain (Kullback–Leibler divergence)
- Entropy measures the level of "impurity" of a split
  - Minimum impurity (all one class) has entropy of 0
  - Maximum impurity (equal number of all classes) has entropy of 1

$$H(x) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

where p_i is the probability of class i

Decision Tree Learning



Minimum Impurity
$H(x) = -1 \times \log_2(1) = 0$

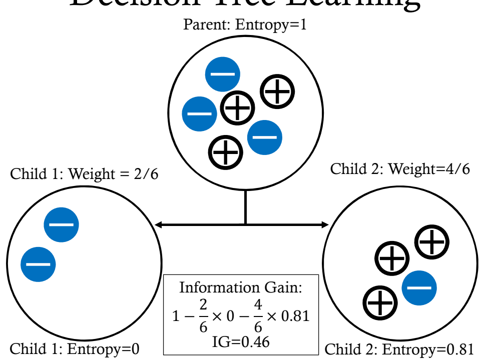
Maximum Impurity
$H(x) = -0.5 \times \log_2(0.5) - 0.5 \times \log_2(0.5) = 1$

Decision Tree Learning

- Information gain is used to decide the order of each attribute split
- Tells us the importance of an attribute of a given feature vector for discriminating between classes
- Information gain = (entropy of parent) – (weighted average entropy of children)
- Choose split with the highest information gain

The University of Iowa

Decision Tree Learning



Decision Tree Learning

Advantages
- Easy to understand and interpret results
- Requires less data preprocessing
- Can handle discrete and continuous data
- Can quickly predict new instances

Disadvantages
- Based on heuristics such as greedy algorithm, therefore cannot guarantee globally optimum solution
- Pruning is necessary to avoid overfitting
- Information gain is biased towards features with more levels, however methods exist for avoiding this bias

The University of Iowa

Ensembles

- Some algorithms are best suited for specific characteristics in data (for example linear or nonlinear)
- Data may have a combination of characteristics
- A single algorithm might not be able to model this combination
- Ensembles combine the results of individual algorithms to obtain better performance than can be achieved by each algorithm alone
- There are several ways to combine the results
- Many data mining competitions are won by ensembles

The University of Iowa

Training and Validation

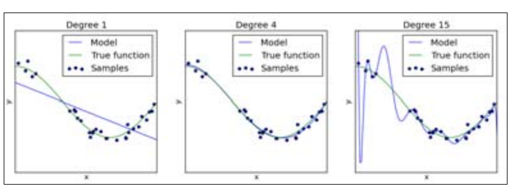
What is the "best" model?
- It depends on the evaluation criteria
- Some errors may be more costly than others
  - Credit card fraud
  - Medical diagnosis
- Confusion matrix

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

The University of Iowa

Training and Validation

- Overfitting versus underfitting
- Model should work well for new (unused)



The University of Iowa

Training and Validation

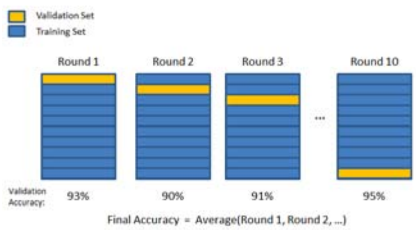
- Cross Validation

Training Set	Validation Set
Used to build model/select parameters	Estimates Error/Model Selection

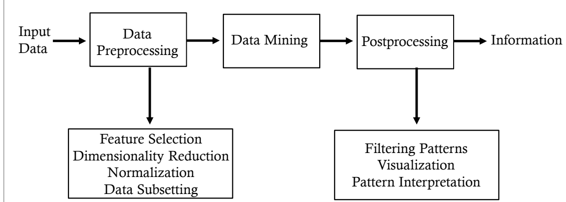
The University of Iowa

Training and Validation

- k-Fold Cross Validation



Knowledge Discovery in Databases



Software

- Weka (free)
- R (free)
- MATLAB
- Statistica
- Many others

Data Sets/Resources

- UCI Machine Learning Repository
  - http://archive.ics.uci.edu/ml/
- UCI KDD Archive
  - http://kdd.ics.uci.edu/
- Carnegie Melon University Statlib
  - http://lib.stat.cmu.edu/index.php
- University of Toronto Delve Datasets
  - http://www.cs.toronto.edu/~delve/data/datasets.html

Classes at UI

- Knowledge Discovery
  - Professor Nick Street
  - CS:6421:0001 or MSCI:6421:0001
- Computational Intelligence
  - Professor Amaury Lendasse
  - IE:6350:0001 or NURS:6900:0001
- Big Data Analytics
  - Professor Amaury Lendasse
  - IE:4172:0001
- Statistical Pattern Recognition
  - Professor Yong Chen
  - IE:6760:0001
- Information Visualization
  - Professor Amaury Lendasse
  - IE:3149:0001

The University of Iowa

References

Tan, P. N., Steinbach, M., & Kumar, V. *Introduction to Data Mining*. Boston: Pearson. 2006.

Witten, I. H., Frank, E., & Hall, M. A. *Data Mining: Practical Machine Learning Tools and Techniques.* Amsterdam: Morgan Kaufmann. 2011.

Santoso, S., Negnevitsky, M., & Hatziargyriou, N. "Applications of Data Mining and Analysis Techniques in Wind Power Systems." *Power Engineering Society General Meeting, 2006. IEEE.*

http://en.wikipedia.org/wiki/Data_mining

The University of Iowa

Figures

http://www.telegraph.co.uk/finance/newsbysector/energy/8575639/Green-taxes-hit-wind-turbine-makers.html

http://new.abb.com/substations/energy-storage-applications/capacity-firming

http://www.proscoutleadgeneration.com/prospect-list-generation-data-mining/

http://www.philippe-fournier-viger.com/spmf/index.php?link=documentation.php

http://aimotion.blogspot.com/2010/08/tools-for-machine-learning-performance.html

The University of Iowa

Figures

http://scikit-learn.org/stable/auto_examples/plot_underfitting_overfitting.html

https://chrisjmccormick.wordpress.com/2013/07/31/k-fold-cross-validation-with-matlab-code/

http://fiji.sc/Trainable_Weka_Segmentation_-_How_to_compare_classifiers

The University of Iowa