Data Mining Review

CS 584: Big Data Analytics

Course Logistics: Piazza

Signup link http://piazza.com/emory/spring2016/cs584001

- Announcements
- Dataset resources
- Questions and Discussions

Course Logistics: Class Presentations

- 1 student per presentation, each presentation is roughly 30 minutes (instead of the full class time)
- Max of 3 presentations a week
- Google Spreadsheet Signup: <u>https://docs.google.com/spreadsheets/d/</u> <u>1pyo7gaMnCzqNYK26QkT64S3rUuvAwLDwHUBkZ1cw</u> <u>WCI/edit?usp=sharing</u>

What is Data Mining?

- Definition (Fayyad, Piatetsky-Shapiro and Smyth, 96)
 "Knowledge Discovering in Databases (KDD) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."
- Patterns can be anything from associations, groupings, trends, anomalies, etc.
- "Extraction of useful information from large data sets"

Types of Models

Consider a large collection of fruits with attributes or characteristics such as weight, volume, color, shape...

- **Regression**: predicting weight based on other attributes
- **Classification**: what type of fruit is it?
- **Clustering**: how many different types of fruits are there?
- Anomaly detection: is this fruit different than its other types?

Degrees of Supervision in Learning Algorithms

• Unsupervised

- Model is not provided correct results during the training
- Cluster based on input data and statistical properties
- Supervised
 - Training data includes input and desired results
- · Semi-supervised
 - Tasks that make use of unlabeled data for training with a small amount of labeled data

Data Mining Algorithms

<u>Unsupervised</u>

Continuous

Categorical |

 Clustering & Dimensionality Reduction SVD PCA K-means 	 Regression Linear Polynomial Decision Trees Random Forests
 Association Analysis Apriori FP-Growth Hidden Markov Model 	 Classification KNN Trees Logistic Regression Naive-Bayes SVM

https://nyghtowlblog.files.wordpress.com/2014/04/ml_algorithms.png?w=535&h=311

<u>Supervised</u>

Clustering: K-Means

- Partition the data into k clusters based on their features
- Each cluster is represented by its centroid, defined as the center of the points in each cluster
- Each point is assigned to the cluster whose center is nearest
 Original Data
 K-Means Clustering



CS 584 [Spring 2016] - Ho

Clustering: PCA and SVD

- Linear projection of high dimensional data into a lower dimensional subspace
 - Variance retained is maximized
 - Least square reconstruction
 error is minimized
- Baseline matrix factorization method: best possible matrix approximation given number of components



https://prateekvjoshi.files.wordpress.com/2014/10/2-pca.png

Regression

- Predict new values based on the past
- Compute new values of a dependent variable based on values of one or more measured attributes



Classification: Decision Trees

- A simple set of rules to classify your data
- Splitting criteria determines the rules that will be derived from the data
- Non-parametric because there are no assumptions about the distribution of the variables of each class

Should i play tennis?



http://blogs.msdn.com/blogfiles/chrsmith/WindowsLiveWriter/AwesomeFDecisionTreesPartl 131F5/image 3.png

Classification: Logistic Regression

- Regression model with the dependent variable is categorial (known as a generalized linear model)
- Logistic function to transform linear model
- Produces log-odds ratio as a linear function of predictors



https://en.wikipedia.org/wiki/Logistic regression

Classification: SVM

- Use optimal hyperplane in a suitable feature space for classification
- Flexible représentation of class boundaries
- Allows nonlinear in original features using the "kernel trick"



Classification: k-Nearest Neighbors (KNN)

- Example of instance-based learning, or lazy learning
- Find the k nearest neighbors based on distance metric and classify by assigning the label which is most frequent in the k samples
- Smaller k local, larger k global



https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

Data Mining Books



CS 584 [Spring 2016] - Ho

Data Mining & Python Resources

- Continually updated data science Python notebooks
 <u>https://github.com/donnemartin/data-science-ipython-notebooks</u>
- Applied Predictive Modeling Book & Python <u>http://nbviewer.jupyter.org/github/leig/Applied-Predictive-</u> <u>Modeling-with-Python/tree/master/notebooks/</u>
- Data Science Resources
 <u>https://github.com/jonathan-bower/DataScienceResources</u>
- Interesting iPython Notebooks <u>https://github.com/ipython/ipython/wiki/A-gallery-of-interesting-</u> <u>IPython-Notebooks</u>