Iheanyi Ekechukwu
Data Mining
September 9, 2014

Assignment 1

1. Choose the data technology (Q, H, U, or S) that is most appropriate for each of the following business scenarios. Q - SQL Querying, H - Statistical Hypothesis Testing, U - Unsupervised Data Mining/Pattern Finding, S - Supervised Data Mining/Pattern Finding
    a. H
    b. Q
    c. Q
    d. U
    e. S
    f. U
2. Label each use case as describing either data mining (DM), or the use of the results of data mining (Use)
    a. Use
    b. DM
    c. DM
    d. Use
3. The target variable for MegaTelCompany should be whether the customer canceled their cellphone contract or terminated it, this would allow them to then build a predictive model for identifying who is more likely to terminate their contract.
4. In terms of Joe for Plumbing, Inc. trying to sell gardening tools to his products, Joe is not ready to apply data mining to his customers. This is because although he has data mining for building predictive models for identifying customers for special plumbing

offers, his data from his plumbing transactions are not good enough to accurately identify a subset of customers.

5. The Inductive Learning Hypothesis is given known information, at best we could create a hypothesis around this information we already know. An inductive bias is a different way of finding a solution to a problem, such as using a different predictive model. The difference between induction and deduction is that deduction uses already available information/evidence to support a specific conclusion. On the other hand, induction uses the supporting information/evidence to make a general statement, without necessarily making a firm conclusion.
6. Steps for Developing a Data Mining Task
    a. Problem Statement
    b. Research/background knowledge
    c. Specify the task
    d. Get all of the data
    e. Understand your data
    f. “Janitor” work - data sanitizing
    g. Linking/normalization/consolidation of your data
    h. More “janitor” work
    i. Dataset
    j. Create target Function/machine learning model/training
    k. Validation of data
    l. Interpretation of data
    m. Deployment
    n. Monitoring

Iheanyi Ekechukwu
Data Mining
September 9, 2014

7. 3-5 risks or challenges that data mining/analytics applications may face in deployment

    a. Overfitting data

    b. Data generalization

    c. Lack of optimal model to compare to

    d. Having to rebuild models or build multiple models as more data becomes available and/or necessary