Data Mining, CSCI 347, Fall 2019
Review for exam 2, Nov. 1

Yellowed material is new

Overview of Data Mining

Be able to describe the purpose of data mining

Be able to describe the difference between causation and correlation, and know that data mining finds correlations, not necessarily causations

Know the 2 purposes for which data mining is typically used

Know the two primary types of attributes

Know what is meant by overfitting and ways to reduce it.

Know the difference between supervised and unsupervised learning

Inputs to Data Mining

Know the typical form of the input to data mining algorithms

Know the 4 levels of measurement, what operations are allowed with each, be able to give examples of each, and know not to make interval claims on ordinal data.

Know what is meant by metadata

Know what is meant by incomplete, noisy and inconsistent data, and that this is the data used in data mining

Know what is meant by data cleaning

Know what is meant by data integration

Know ways to handle missing data

Outputs from Data Mining

Know that knowledge can be in the form of structural patterns such as:

1. Tables (decision tables or Bayesian conditional probability tables)
2. Linear models
3. Trees
4. Rules (classification and association)
5. Clusters

Know the form of decision tables

Be able to describe statistical or probability based learning. Specifically, know how to read P[H] and Pr[H|E], and know that Pr[H|E] can be calculated using Bayes Theorem.

Be able to describe linear models

Know that for linear regression the output depends on the sum of the inputs and parameters

Know that logistic regression is a type of linear regression where the class value is binary.

Know the form of trees and how they are constructed from a high level

Know the form of rules and the terms antecedent and consequent

Know the difference between classification/covering and association rules

Be able to describe Instance-Based Learning

Know what is meant by clustering

Algorithms

Simple Algos

Know the idea behind and be able to predict using 0R

Know the meaning of the mean, mode and median

Know the idea behind and be able to predict using 1R

Naïve Bayes

Know the idea behind Naïve Bayes and be able to predict using Naïve Bayes

Know the meaning of Pr [A|B] and what is meant by the hypothesis and evidence in Bayes Theorem

Know to calculate the probability for each class value and normalize the results

Know what the Laplace estimator is, why it is used, and be able to use it

Decision Trees

Be able to construct a decision tree, using the "gain ratio" recursively for a small dataset.

Know that information is given in bits, and be able to calculate the information needed to classify elements at a node using entropy.

Know what is meant by a "pure" node.

Know the definition of entropy and be able to calculate it

Know the terms and be able to calculate: information gain of a node splitting into subnodes, the intrinsic information in a split, and the gain ratio of a node splitting into subnodes.

Know what is meant by highly branching attributes, and be able to explain why the gain ratio is used rather than the information gain when selecting an attribute to split on

Know how decision trees and rules are similar, but different

Constructing Covering Rules

Know the terms coverage (support) and accuracy (confidence) of a rule

Given a dataset, know how to find a rule set that classifies it completely.

Association Rules

Know what is meant by an item and item set in the context of learning association rules.

Know the purpose of mining association rules and that this is unsupervised.

Given a dataset, know how to generate rules which meet a minimum coverage and accuracy.

Linear Regression

Know that, in data mining "regression" refers to the process of predicting a numeric quantify (i.e. finding a formula).

Know the purpose of linear regression, and that it is finding weights to minimize the squares of the differences between the actual and predicated values in the formula:

$$x = w_0 + w_1 * a_1 + w_2 * a_2 + \ldots + w_n * a_n$$

where there are n attributes $a_1, \ldots, a_n$

That is, minimizing

$$\sum_{j=1}^{m} \left(x^{(j)} - \sum_{i=0}^{n} w_i * a_i^{(j)} \right)^2$$

(You don't need to memorize this second formula, but do be able to explain what each item is.)

Know that linear regression works well when the data is truly linear or near linear.

Given a dataset of linearly separable elements, know how to use a single-level perceptron to find that line.

Synthesis

Based on the class material, class presentations (case studies, special learning, competition and project presentations) be able to make recommendations as what algorithms to use in various situations.

Ethics

Know that the Data Science Association has created a code of conduct and be able to discuss rules within it.

Know that due models being opaque, unregulated and uncontestable they can reinforce discrimination