**Objectives for Midterm 1**

You should be able to answer the questions on the first test if you can do the things listed in the following. The test will only sample your knowledge, and so will be shorter than this list.

You may use your textbook, copies of the lecture slides, and your own notes, but no other materials. You can use a pocket calculator (not programmable!) if you want. You cannot use a laptop.

* Distinguish data mining tasks from ordinary data manipulation tasks, and explain the difference.
* List, explain, and identify the different types of data attributes (nominal, ordinal, interval).
* Explain the nature and significance of noise and outliers in data analysis.
* List, explain, compare, apply, and evaluate methods for handling missing data values.
* List, explain, and compare different methods for reducing the volume of data.
* List, explain, apply, and compare different methods for converting continuous attributes into discrete attributes.
* List, explain, compare, and compute different types of measures of similarity and dissimilarity between data points.
* Verify if a distance measure is a metric.
* Explain when to use asymetric similarity measures such as Jaccard measures rather than symmetric measures.
* Explain and compute summary statistics such as the mean, median, and mode of a data set.
* Explain and draw boxplots to depict data.
* Critique visualizations with respect to representation, arrangement, and selection.
* Explain and give examples of the notion of decision trees.
* Explain the process for top-down construction of decision trees, including the main splitting and stopping criteria.
* Explain and compute Gini index, entropy, and misclassification error.
* Compare alternative splitting attributes in decision tree construction by applying node impurity measures.
* Construct decision trees from small data sets.
* Explain the problems caused by underfitting and overfitting data.
* List and describe different techniques to evaluate the performance of a classifier.
* Explain and compute optimistic and pessimistic error estimates for decision trees.
* Explain and compute confusion matrix, and .basic performance measures for binary classifiers.