**Objectives for Final**

You should be able to answer the questions on final 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.

The test will cover material presented throughout the course, but not from the poster presentations.

You may use your textbook, copies of the lecture slides, a one page HANDWRITTEN cheat sheet, and your own class notes, but no other materials. You can use a pocket calculator (not programmable!), but you cannot use a laptop.

# New Objectives

* (P491) Explain and compare the different types of clusters.
* (P501) Explain the notion of clusters defined by an objective function.
* () Explain, apply, analyze, and compare k-means, bisecting k-means, hierarchical, density-based, and graph-based clustering methods.
* Explain, compare, and apply different methods for measuring the similarity of clusters.
* Define, explain, compare, and use the different graph notations that are used by graph based clustering algorithms
* Explain, compare, and apply methods for evaluating the quality/validity of clusters produced by clustering methods.
* (P651) Explain the importance of outliers and anomalies for computational tasks.
* Describe and compare anomaly detection scenarios and challenges.
* Describe, compare, and apply the anomaly detection approaches covered in class.

# Objectives from prior Midterms

# Midterm 2

* Use confusion and cost matrices to compute which of two classifiers is better for a data set.
* Explain the notion of an ROC curve and its meaning for classifier performance.
* Use ROC curves to compare performance of different classifiers.
* Construct sets of decision rules from decision trees, and decision trees from sets of decision rules.
* Explain and compare the notions of coverage and accuracy of classification rules.
* Explain how to construct sets of classification rules, and compare this process to methods for constructing decision trees.
* Explain and apply instance based classifiers.
* State and explain Bayes theorem and its use in updating probability distributions to incorporate new evidence, and use it to compute probabilities.
* Explain and use Bayes classifiers.
* Diagram and explain the naive Bayesian classification method.
* Compute probabilities in small data sets and use these values in a naive Bayesian classifier to classify data items.
* Explain and evaluate artificial neural network classification methods.
* Explain the notions of maximum margin hyperplanes and support vectors and their roles and uses in support vector machines.
* Explain and compare bagging and boosting ensemble methods for improving classifier performance.
* Explain and compare the notions of association and classification rules.
* Explain and compare the notions of support, and confidence of association rules.
* Describe and use the Apriori Principle and the Apriori Algorithm.
* Describe and apply hash trees.
* Define and use maximal frequent itemsets and closed itemsets.
* Explain methods for frequent itemset generation.
* Explain how to efficiently generate rules from frequent itemsets.
* Explain and use multiple minimum support thresholds.
* Explain, compute, and critique pattern interestingness measures.
* Describe the effect of support pruning on pattern interestingness measures.
* Define and use the notions of sequence and subsequence.
* Define the sequential pattern mining problem.
* Describe and apply the Generalized Sequential Pattern Mining Approach.
* Explain and use timing constraints for sequential pattern.
* Explain and perform mining for sequential pattern with timing constraints.

# Midterm 1

* 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 asymmetric 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.