Data Mining and Exploration
Interim Report

Thorvaldur Helgason (s1237131)
Daniel Stanoescu (STUDENT NUMBER)
Maria Alexandra Alecu (STUDENT NUMBER)

February 27, 2013

What we have done so far:

- Pre-processing:
  - Replaced missing values with both zeros and mean values.
  - Converted the dataset to a binary bag-of-features.
- Familiarized ourselves with Naive Bayes, SVMs, and Decision Trees.

What we plan on doing:

- See which pre-processing techniques and features work best for different classifiers.
- Familiarize ourselves with more classifiers.
- Compare the performance of the classifiers with the criteria described below.

What comparisons we want to run:

- Split data up randomly: 80% of instances will be the training set and 20% the test set.
- Perform 5-fold cross-validation on the training set and do final evaluation on the test set.
- For each classifier we store their accuracy, confusion matrix, ROC curve and AUC.