CSC581 Final Spring 2016

due Tuesday 5/10 @ midnight in Sakai

version 1.0

Name:___________________

Short Answer Problems

1. What is meant by a maximum-margin classifier and why are they preferable over perceptrons?

2. What do we mean by the *feasible region* of an optimization problem?

3. What are the KKT conditions and why are they important?

4. How do a Lagrangian and a Lagrangian dual differ?

5. Why do we consider maximum-margin classifiers with large margin less complex than maximum-margin classifiers with a small margin?

6. Explain the basic idea behind the gradient ascent/descent optimization technique.

7. What is *quadratic programming* and how can it be used in the context of SVMs?

8. Given that both SVMs and MLPs can solve nonlinear decision problems, what are some of the advantages of SVMs?

9. What is meant by the term *kernel trick* in the context of SMVs?

10. What is a *support vector*?

Problems

For the final examination you have a choice:

1. Build and evaluate support vector machine *regression* models for an appropriate regression data set, OR
2. Select a classification data set and build SVM and RandomForest classifiers for this data set and compare them (see Part C below).

Please indicate your choice clearly. For data set selection the same rules apply as for the midterm.

Part A Perform an exploratory data analysis using summary statistics and histograms. Briefly explain your findings.

Part B Build the best model possible for your data set:

1. Document your grid search/model evaluation process carefully, including the type of kernel you are using, the values of its free parameters, and the value of C.
2. For regression use the cross-validated mse (or rmse) in order to determine your best model.
3. Select the two best performing models.

Part C Investigate whether the difference in performance of your top two models is statistically significant or not using the bootstrap. You should use 95% confidence intervals for this investigation.

1. What are the 95% error confidence intervals for your two models?
2. Is the performance difference statistically significant? If yes, which model would you pick? if no, which model would you pick and why?

Write a brief report summarizing your findings from Parts A, B, and C.
NOTE: All the work has to be done in R.
NOTE: All work has to be your own, no collaboration is permitted.