**GEOG316/PLAN351 Multivariate Statistics**

**Assignment 5: Statistical (machine) learning**

*Submission deadline: Friday Nov 24, 2017 at 11:59pm*

*15% of final grade: This assignment is to be completed individually*

Estimated time for completion: 5-6 hours (please let me know if it takes you significantly longer/shorter than this).

**Introduction**

In this assignment, we apply two different statistical learning methods to tree measurements taken in Roosevelt National Forest in Colorado, USA. The data are found in the file: covtype\_sm\_sample.csv. We wish to determine whether the categorical variable tree type (1-7):

1 - Spruce/Fir   
2 - Lodgepole Pine   
3 - Ponderosa Pine   
4 - Cottonwood/Willow   
5 - Aspen   
6 - Douglas-fir   
7 – Krummholz

can be classified as a function of these continuous measurement variables:

**Elevation**: Elevation in meters   
**Aspect**: Aspect in degrees azimuth   
**Slope**: Slope in degrees   
**Horizontal Distance To Hydrology**: Horz Dist to nearest surface water features (m)

**Vertical Distance To Hydrology**: Vert Dist to nearest surface water features (m)  
**Horizontal Distance To Roadways**: Horz Dist to nearest roadway (m)  
**Hillshade 9am** (0 to 255 index): Hillshade index at 9am, summer solstice   
**Hillshade Noon** (0 to 255 index): Hillshade index at noon, summer solstice   
**Hillshade 3pm** (0 to 255 index): Hillshade index at 3pm, summer solstice   
**Horizontal Distance To Fire Points**: Horz Dist to nearest wildfire ignition points (m)

The data also includes a categorical variable describing which wilderness area (1-4) a particular tree is located inside. See: <https://www.kaggle.com/uciml/forest-cover-type-dataset> for more information about the dataset.

**Tasks [29 marks total]**

* + - 1. 1. Using the starter script provided, use randomForest() to build a model to classify tree Type using all available predictors from the subset of the trees data. Create 5000 trees, and set the parameter *m* for subsetting predictor variables at each split to 2. [3 marks]
         2. What is the accuracy of this model in predicting the training data (i.e. "in the bag" accuracy)? What is the accuracy for "out-of-the-bag" OOB predictions? [2 marks]
         3. Which tree types are generally well predicted, and which are less well predicted? [2 marks]
         4. Fit the model for a second time and note what the OOB error is -- what do you notice, and why does this happen? [1 mark]
         5. This time running randomForest() from within the train() function of the caret package, perform repeated 10-fold cross-validation on a randomForest model to classify Tree type using all predictors. Test different values of the parameter *m* to tune the model by finding the parameter value that produces highest accuracy (lowest error rate). [4 marks]
         6. How does this value of *m* differ from what the textbook would predict? [1 mark]
         7. Create a plot showing variable importance from the cross-validated fit, as a percentage relative to the most important variable. Which variable(s) emerge as most important? Interpret the importance of the first two predictor variables. [4 marks]
         8. Make a table showing the highest accuracy achieved for each method, and fill in the value for randomForest. [2 marks]
         9. Run the svm() function in the e1071 library from within train() to tune the optimal SVM model to classify Type, using a radial kernel and all available predictors from the subset of the trees data. Perform a grid search to tune the cost (C) and gamma (G; referred to as sigma in train()) parameters over the ranges C=10^(-1:5) and G=2^seq(-15,3,2). Report the cross-validated accuracy from the best, and worst, combination of parameters you tried. [5 marks]
         10. Record the highest accuracy value in the table you created in 3. [1 mark]
         11. Compare the accuracy you achieved using the random forest and SVM methods, and comment on which method works best to classify tree type. Making reference to any relevant academic literature sources you can find, offer possible reasons for why one method may perform better than another for this type of problem. [4 marks]

**Submission**

Put all R commands into a script, which should be uploaded along with your written answers. Show all necessary mathematical working in the written document, and include any relevant tables and figures (each one must have a number and caption, and be explicitly referenced in the text).

Prepare your answers to the questions above in a PDF document, 12pt single-spaced. Remember to include your Student ID, name, and date on the first page. All figures/tables must be labelled, including a number and caption, and should appear immediately below the question in which they are first discussed. All figures/tables must be explicitly referred to in the text, otherwise they should not be included.

Declaration: please verify that the following statements are true, and if so include this text on the front page of your submitted report:

1. This assignment was completed by my own efforts and I did not collaborate with any other person for ideas or answers.
2. This is the first time I have submitted this assignment or essay (either partially or entirely) for academic evaluation.

Please upload your final submission to the Dropbox on LEARN, before the submission deadline. There are no late penalties: late work will not be graded, without appropriate supporting documentation (e.g. illness verification).