STAT3612 Assignment 3: Classification

Date: April 13, 2017

Due date: April 24, 2017

Consider again the Times Higher Education (THE) World University Ranking dataset. As in Test 1, we compare the scores between the top-ranked eastern unversities (China, Hong Kong, Taiwan, Singapore, Janan, South Korea, subject to `world_rank<=200`) and the top universities in the USA (subject to `world_rank<=30`) in the last five years (i.e. 2012–2016).

Step 0. Extract the subdata with conditions described above. The sample sizes for East and USA categories should match the following table.

	2012	2013	2014	2015	2016
East	18	18	17	19	14
USA	21	20	22	22	18

Suppose we take `Region` (East vs. USA) as the target response, and take `research` and `teaching` as two predictors. Perform the following classification tasks.

Step 1. (30%) Fit a logistic regression model for `Region` with predictors `research` and `teaching`. a) Find the fitted decision boundary that separates East and USA observations; b) Visualize the decision boundary together with the training data points on the two-dimensional space; c) What is the misclassification error rate?

Step 2. (20%) Fit a classification tree for `Region` with predictors `research` and `teaching`; a) Plot the fitted tree upon suitable pruning; b) What is the misclassification error rate?

Step 3. (30%) Fit a random forest for `Region` with predictors `research` and `teaching`; a) Visualize the random forest predictions using the 2D image plot together with the training data points; b) Fit a tree bagging model and comare the out-of-bag estimate of error rate to that of the random forest fit.

Step 4. (20%) a) Draw on a single chart the three ROC curves for logistic regression, classification tree and random forest, as fitted in Step 1-3 respectively. b) What are their AUC scores?