Quiz 2

**Stat. 6620 Statistical Learning in R**  
**Department of Statistics and Biostatistics**  
**CSU East Bay**

May 29, 2018

**Name: Surabhi Asati**

**Class: 6620**

**Section: 1**

**Instructions:** This is a take-home quiz.

Type your answers in this .docx file and upload it in Blackboard before the end of the day Tuesday May 29.

**Academic Honesty:** As a student at CSU East Bay you are held to the standards stated in the Academic Dishonesty Policy. Copying another student’s work or allowing another student to copy your work is academically dishonest. I expect you to be academically honest while taking the test.

1. **Random Forests:**
   1. **What is an ensemble method? Explain.**

Ensemble method is a meta-learning approach of Improving model performance. All ensemble methods are based on the idea that by combining multiple weaker learners, a stronger learner is created. So, ensemble method combines several base models in order to produce one optimal predictive model. By building classifier out of a collection of trees and then performs cross validation.

Advantages:

Improved performance on massive or miniscule datasets

The ability to synthesize of data from distinct domains

Disadvantages:

It may or may not work (depends on the dataset)

Ensemble Methods:

* 1. Bagging or bootstrap aggregating

Bagging generates a number of training datasets by bootstrap sampling the original training data. These datasets are then used to generate a set of models using a single learning algorithm. The models' predictions are combined using voting (for classification) or averaging (for numeric prediction).

* 1. Boosting

The main principle of boosting is to fit a sequence of weak learners that are only slightly better than random guessing, such as small decision trees to weighted versions of the data. More weight is given to examples that were misclassified by earlier rounds.

* 1. Random Forests

This ensemble-based method combines the base principles of bagging with random feature selection (rows and columns both are randomly sampled) to add additional diversity to the decision tree models. After the ensemble of trees (the forest) is generated, the model uses a vote to combine the trees' predictions.

Random forests combine versatility and power into a single machine learning

approach. Because the ensemble uses only a small, random portion of the full feature set, random forests can handle extremely large datasets, where the so-called "curse of dimensionality" might cause other models to fail. At the same time, its error rates for most learning tasks are on par with nearly any other method

* 1. **Explain how Random Forests build decision trees.**

Random Forests build multiple independent decision trees by taking random samples (rows and columns) of the data. Example:

Dataset has 1,000 rows and 30 columns. There are two levels of randomness in this algorithm:

* **At row level**: Each of these decision trees gets a random sample of the training data (say 10%) i.e. each of these trees will be trained independently on 100 randomly chosen rows out of 1,000 rows of data.
* **At column level**: The second level of randomness is introduced at the column level. Not all the columns are passed into training each of the decision trees. Say we want only 10% of columns to be sent to each tree. This means a randomly selected 3 column will be sent to each tree. So for the first decision tree, may be column C1, C2 and C4 were chosen.

The random-forest algorithm brings extra randomness into the model, when it is growing the trees. Instead of searching for the best feature while splitting a node, it searches for the best feature among a random subset of features. This process creates a wide diversity, which generally results in a better model.

1. **Holdout:**
   1. **What is the holdout method? Explain. Draw a diagram.**

The holdout method:

The procedure of partitioning data into training and test datasets that we used in previous chapters is known as the holdout method. The training dataset is used to generate the model, which is then applied to the test dataset to generate predictions for evaluation. Typically, about one-third of the data is held out for testing and two-thirds used for training, but this proportion can vary depending on the amount of data available. Test data should be used only once!

**The holdout method**

**Train**

Split training data to get validation data

**Validate**

Work the model on validation data iterating

and refining the model and hold out the test data

**Test**

Use test data only once as a final step

to report an estimated error rate for future predictions

For the holdout method to result in a truly accurate estimate of future performance, it is better to divide the original data so that in addition to the training and test datasets, a third validation dataset is available. A typical split between training, test, and validation would be 50 percent, 25 percent, and 25 percent respectively.

* 1. **What is repeated holdout? Explain.**

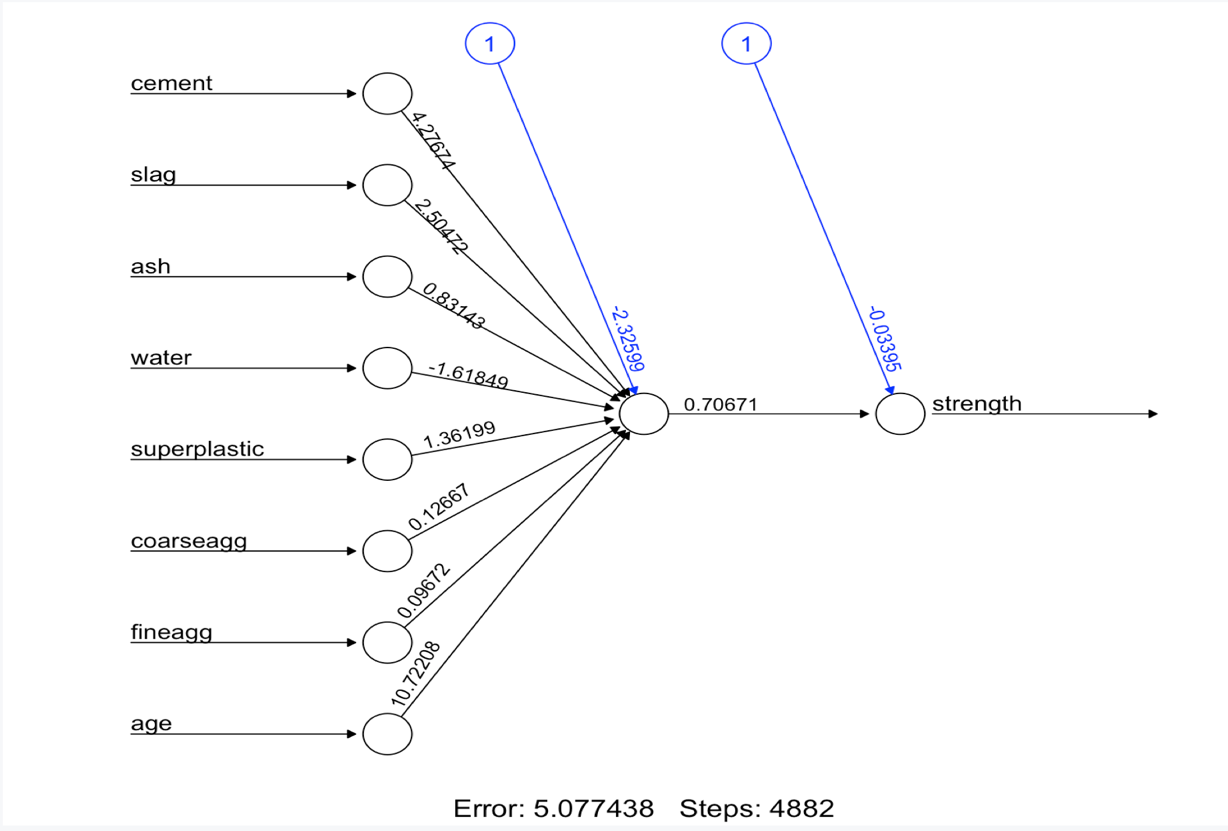
A technique called repeated holdout is sometimes used to mitigate the problems of randomly composed training datasets. The repeated holdout method is a special case of the holdout method that uses the average result from several random holdout samples to evaluate a model's performance. As multiple holdout samples are used, it is less likely that the model is trained or tested on non-representative data.

* 1. **Give the name of two repeated holdout methods.**
     1. Cross Validation
     2. Bootstrap Sampling

1. **Neural Networks:**
   1. **Sketch a diagram of a multilayer feedforward neural network with one hidden layer.**

Example:

# simple ANN with only a single hidden neuron  
set.seed(12345) # to guarantee repeatable results  
concrete\_model <- neuralnet(formula = strength ~ cement + slag +  
 ash + water + superplastic +   
 coarseagg + fineagg + age,  
 data = concrete\_train)  
# visualize the network topology  
plot(concrete\_model)



Input Nodes: 8

Hidden Nodes: 1

Output Nodes: 1

* 1. **(Stat. 6620) Why are continuous activation functions used with neural networks?**

With **continuous activation functions, t**he output signal is no longer binary; output values can fall anywhere in the range from 0 to 1. Additionally, the sigmoid is differentiable, which means that it is possible to calculate the derivative across the entire range of inputs, this feature is crucial for creating efficient ANN optimization algorithms.

1. **Market Basket Analysis:**

**What is transactional data?**

Transactional datasets are typically extremely large, both in terms of the number of transactions as well as the number of features (that is, items) that are monitored. Adding difficulty is the fact that the number of potential itemsets grows exponentially with the number of features; given k items that can appear or not appear in a set, there are on the order of 2^k possible itemsets that must be searched for rules. A retailer that sells only 100 different items could have about 2^100 = 1e+30 itemsets that a learner would have to evaluate - a seemingly impossible task.

* 1. **Why are item matrices sparse?**

To allows the data structure to be more memory efficient than an equivalently sized matrix or data frame. A conventional data structure will quickly become too large to fit into memory for every additional item .Even with the relatively small transactional dataset, the matrix may contains nearly over a million cells, most of which would contain zeros (hence the name "sparse" matrix). Since there is no benefit to storing all these zero values, sparse matrix does not actually store the full matrix in memory; it only stores the cells that are occupied by an item.

* 1. **Explain the Apriori property that is used in the Apriori algorithm?**

Apriori Property: “ All subsets of a frequent itemset must also be frequent ”

For example, the set {motor oil, lipstick} can only be frequent if both {motor oil} and {lipstick} occur frequently as well. Consequently, if either motor oil or lipstick is infrequent, then any set containing these items can be excluded from the search.