**MSDS 6306: Doing Data Science**

**Live session Unit 10/11**

**Due: Monday July 22nd at 11:59pm**

Supplementary Videos:

KNN Classification in R: <https://youtu.be/pBhmsF089vI>

KNN Regression in R: <https://youtu.be/ztYoWV0HYYg>

<https://www.youtube.com/watch?v=ByJW0e-DGKs>

<https://www.youtube.com/watch?v=faGF3AAwUSM>

<https://www.youtube.com/watch?v=VJ0ONXgGi0w>

<https://www.youtube.com/watch?v=13D0GqSfDBk>

<https://www.youtube.com/watch?v=gPGOOpCLuck>

<https://www.youtube.com/watch?v=hTGhy6xIxBM>

**Submission**

**ALL MATERIAL MUST BE KNITTED INTO A SINGLE, LEGIBLE, AND DOCUMENTED HTML DOCUMENT.** Use RMarkdown to create this file. Formatting can be basic, but it should be easily human-readable. Unless otherwise stated, **please enable {r, echo=TRUE} so your code is visible.**

**Questions**

**Background:** Brewmeisters in Colorado and Texas have teamed up to analyze the relationship between ABV and IBU in each of their states. Use the data sets from the project to help them in their analysis. There three main questions of interest are 1) Is there a significant linear relationship between ABV (response) and IBU (explanatory), 2) Is this relationship different between beers in Colorado and Texas and 3) Is there a significant quadratic component in this relationship for either Colorado or Texas or both?

**I. KNN Regression versus Linear Regression**

**A. Clean an prepare the data: (Same as Unit 9 Assignment (nothing new to do).**

1. Create column for brewery ID that is common to both datasets similar to what you did in the project. So we can merge!

2. Merge the beer and brewery data into a single dataframe.

3. Clean the State Column … get rid of extraneous white space.

4. Create One Dataset that has only Colorado and Texas beers and no IBU NAs … name it “beerCOTX”

5. Order beerCOTX by IBU (ascending) ... this will be important later in graphing

**B. Compare two competing models: External Cross Validation**

8. For this assignment ***we will concentrate only on the Texas data***! Create a training and test set from the data (60%/40% split respectively). Print a summary of each new data frame… there should be two: TrainingTX, TestTX.

9. 7. Using the training dat, fit a KNN regression model to predict ABV from IBU. You should use the knnreg function in the caret package. Fit two separate models: one with k = 3 and one with k = 5. (This is 2 models total.)

10. Use the ASE loss function and external cross validation to provide evidence as to which model (k = 3 or k = 5) is more appropriate. Remember your answer should be supported with why you feel a certain model is appropriate. Your analysis should include the average squared error (ASE) for each model from the test set. Your analysis should also include a clear discussion, using the ASEs, as to which model you feel is more appropriate.

Here is the predicted ABV for the ith beer, is the actual ABV of the ith beer and n is the sample size.

11. Now use the ASE loss function and external cross validation to provide evidence as to which model (the linear regression model from last week or the “best” KNN regression model from this week (from question 10)) is more appropriate.

12. Use your “best” KNN regression model to predict the ABV for an IBU of 150, 170 and 190. What issue do you see with using KNN to extrapolate?

**II. KNN Classification**

We would like to be able to use ABV and IBU to classify beers between 3 styles: American IPA and American Pale Ale.

13. Filter the beerCOTX dataframe for only beers that are **from Texas** and are **American IPA and American Pale Ale**.

14. Divide this filtered data set into a training and test set (60/40, training / test split).

15. Use the class package’s knn function to build an KNN classifier with k = 3 that will use ABV and IBU as features (explanatory variables) to classify Texas beers as American IPA or American Pale Ale using the Training data. Use your test set to create a confusion table to estimate the accuracy, sensitivity and specificity of the model.

16. Using the same process as in the last question, find the accuracy, sensitivity and specificity of a KNN model with k = 5. Which is “better”? Why?

BONUS (5 pts total): We did not have a lot data to build and test this classifier. Check out the *class* package’s knn.cv function that will perform leave-one-out cross validation. What is leave-one-out CV (2pts)? Get the accuracy metric for from this function for both the k = 3 and k = 5 KNN classifiers (2pts). Which model is suggested by the leave-one-out CV method (1pt)?

***Unit 11 Questions***

1. Use the most updated code that is zipped with this. It fixes the grep problem by pasting a string with bracketed regular expression. I think someone mentioned this in class. Good call!

2. Use the “snippet” instead of the headline.

3. Look at data from 1989 to 1999

4. To provide external cross validation (50%-50%). Create a training and test set from the total number of articles. Train the classifier on the training set and create your confusion matrix from the test set. Make sure and provide the confusion matrix.

6. Provide accuracy, sensitivity and specificity from the confusion matrix. You may consider News to be the positive.

7. Use your statistics from the last two questions to assess whether the headline or the snippet makes for a better classifier.

**Reminder**

To complete this assignment, please submit **one** RMarkdown and matching HTML file by the deadline. Please submit all files at the same time; only one submission is granted.

Good luck!