Data Science 2: Midterm Exam

March, 2018

Please include at the top of your exam whether you are a 109b/121b student, or a 209b student

This exam involves exploring a data set on red wine quality, and how quality relates to physio-chemical features of wine. A description of the study can be downloaded from the **publisher's web site.** The following questions will focus on a subset of data consisting of 1599 red wines. The data are contained in the file `winequality-red.csv`. For each bottle of wine, the following features are measured.

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. pH
10. sulphates
11. alcohol
12. quality (score between 0 and 10)

The main goal of this analysis is to predict red wine quality from the physio-chemical features.

Problem 1 [30 points]

Fit an additive model of quality on the physio-chemical variables on all 1599 wines in the data set. Use smoothing splines to fit each predictor variable. No need to explicitly perform cross-validation - please use the default smoothing selections.

(a) [5 points] Plot the smooth of each predictor variable with standard error bands. Which variables seem to have a non-linear contribution to mean quality?

(b) [10 points] Is the overall non-linearity evidenced in the variable-specific smooths statistically significant? Justify your answer with a likelihood ratio test comparing the additive model to a model that includes the features linearly.

(c) [10 points] We now want to investigate how to produce the best expected wine quality based on the physio-chemical content.

- [109b/121b students only] Based on the additive model fit, how might you **approximately** optimize the physio-chemical composition to produce the highest expected wine quality? Use the results from part (a) to answer this question. What is the resulting estimated wine quality? *Hint: For the latter part, use the* `predict` *function.*
- [209b students only] Based on the GAM fit, determine **numerically** the optimal combination of physio-chemical features to produce the highest expected wine quality. What are the values of the physio-chemical features corresponding to the optimal wine, and what is the quality rating for the optimized wine? *Hint: The* `preplot` *function applied to a GAM fit produces a list containing as many components as there are smoothed predictors, and each component is a list itself containing the* `x` *and* `y` *values that produce the variable-specific smooths.* It may help that the intercept is the sample mean of the quality scores. Or you can use the values as part of the `predict` function to obtain the estimated value of the optimized wine.

(d) [5 points] What might be a concern or limitation with optimizing the physio-chemical composition of wine based on the additive model fit? Your answer should be connected to the assumptions underlying additive models.

Problem 2 [40 points]

Rather than fit a single model to all of the wines, we will fit different models to different subsets of the data (in Problem 3). In preparation, this problem will involve partitioning the data into different clusters/subsets.

(a) [5 points] Explain a reason we might expect different relationships between quality and physio-chemical wine composition by different subsets of the data identified in Problem 1.

(b) [5 points] Prior to performing clustering, you will center each column, and also scale each column so that each transformed feature has a standard deviation of 1.0. Briefly justify the decision to scale the data in this manner. Be specific to the context of this problem.

(c) [10 points] Suppose we decide to perform partitioning-around-medoids clustering of the observations based only on the physio-chemical features but not using quality. To determine the best number of clusters, optimize based on the gap statistic in the following manner:

1. Set the random number seed to 123 (`set.seed(123)`). Now select a random sample of 200 wines (*hint: use the* `sample` *function*).
2. Set the random number seed to 321 (`set.seed(321)`). Optimize the gap statistic using the method described by Tibshirani (2001) based on the standard error rule, using `d.power=2`.

   Explain how 1 cluster is the optimal number of clusters according to Tibshirani's rule, even though 6 clusters would be chosen if one were to use the maximum gap statistic.

(d) [10 points] Partition the full data into six clusters via partitioning-around-medoids on the scaled version of the data. Save the cluster identifiers as a new column in the original data frame (*Hint: the* `clustering` *component of the resulting cluster object contains the IDs*). Plot the first two principal components of the scaled data and visually show the cluster memberships. Show that the proportion of variance in the original data represented by the principal component plot is 45.7%. Use the output of `prcomp` to demonstrate this.

(e) [10 points] Create a side-by-side boxplot of quality scores by cluster (*Hint: If using* `ggplot` *you should use the* `geom_boxplot` *function – do not forget to make the cluster ID variable a factor in R*). Does the distribution of quality scores differ visually by the clusters you determined? Would you have expected the distribution of quality scores to differ?

Problem 3 [30 points]

We will now fit a normal hierarchical linear model for quality scores against the physio-chemical predictors nested in the formed clusters from the previous problem.

(a) [10 points] Implement a normal hierarchical linear model in Stan (called from R) to fit the model. Make sure you let all the linear model coefficients vary by cluster. *Hint: You may find the Stan code supplied with the lecture notes helpful.* You may assume that the intercepts across the six clusters have a normal prior distribution with a mean which is the average of the quality scores across the whole data set, and with an unknown standard deviation. The 11 physio-chemical coefficients across the six clusters can be assumed to be normally distributed centered at 0 with different standard deviations. Finally, you may assume that all the standard deviation parameters have a prior uniform distribution with a minimum of 0 and maximum of 100 (which is sufficiently large).

(b) [10 points] Briefly report on the details of your model implementation (number of iterations of burn-in and the number of iterations of saved parameters, number of parallel samplers, and any assurances that the sampler converged). (*Hint: If you saved the Stan fit of your model in the R object* `wine.fit`, *you can access the matrix of model summaries from* `summary(wine.fit)$summary`.) Do not be concerned about warnings of divergent transitions after warm-up if you have evidence that the sampler converged for the feature coefficients.

(c) [10 points] Create a visualization that demonstrates the variation of coefficients across clusters. One natural way would be to display side-by-side boxplots of the posterior simulated draws for the relevant coefficients. (*Hint: Use the* `extract` *function applied to the fitted Stan model to obtain simulated coefficient values.*) Based on these results, do you think that the hierarchical model by formed clusters was helpful in explaining the variation in quality scores? Briefly justify.