Homework 4

Due Thursday 3/10

Solutions should include R code and output – try to use `R markdown`. Homework must be submitted on Blackboard as one file.

Zhicheng Ma, Laura Moreno Herrara, Sanjib Panta, and Ruiyang Sun will lead HW discussion.

1. (80 points) *Modeling High COVID-19 Community Levels for Indoor Masking Recommendation.*

   The CDC recently announced a new classification of county COVID-19 risk based on a combination of COVID-19 case rates and COVID-19 hospitalization rates.[1] In conjunction with this county classification, the CDC provides recommendations for risk mitigation at each of the three community levels low, medium, and high. Masks are recommended indoors when a county is classified as high.

   In this question we are interested in building classification models for determining high (masks recommended indoors) vs. low/medium community COVID-19 levels. This homework uses data very similar to previous assignments, but with a new target variable:

   $$high\_community\_level = \begin{cases} 1 \text{ if } covid\_community\_level = \text{high} \\ 0 \text{ otherwise} \end{cases} .$$

   The dataset for this assignment *covid_data_pdb_v3.csv* is posted on Blackboard under HW4.

   Use the following 17 predictors for the classification models:

   - $pct\_URBANIZED\_AREA\_POP\_CEN\_2010$: percent of urbanized land,
   - $pct\_Males\_ACS\_14\_18$: percent males,
   - $pct\_Pop\_under\_5\_ACS\_14\_18$: percent ¡ 5 years old,
   - $pct\_Pop\_5\_17\_ACS\_14\_18$: percent 5-17 years old,
   - $pct\_Pop\_25\_44\_ACS\_14\_18$: percent 15-44 years old,
   - $pct\_Pop\_45\_64\_ACS\_14\_18$: percent 45-64 years old,
   - $pct\_Pop\_65plus\_ACS\_14\_18$: percent 65+ years old,
   - $pct\_Renter\_Occp\_HU\_ACS\_14\_18$: percent in renter-occupied housing,
   - $pct\_Vacant\_Units\_ACS\_14\_18$: percent of housing units that are vacant,
   - $pct\_Mobile\_Homes\_ACS\_14\_18$: percent living in mobile homes,
   - $pct\_HHD\_NoCompDevic\_ACS\_14\_18$: percent without a computing device,
   - $pct\_HHD\_No\_Internet\_ACS\_14\_18$: percent without internet,
   - $pct\_Hispanic\_ACS\_14\_18$: percent Hispanic origin,

[1] See `https://www.cdc.gov/coronavirus/2019-ncov/your-health/covid-by-county.html` for an overview, and `https://www.cdc.gov/coronavirus/2019-ncov/science/community-levels.html` for details on how the CDC determines COVID-19 community level.

- *pct_NH_White_alone_ACS_14_18*: percent non-Hispanic White,
- *pct_NH_Blk_alone_ACS_14_18*: percent non-Hispanic Black,
- *pct_Schl_Enroll_3_4_ACS_14_18*: percent enrolled in school,
- *pct_Prs_Blw_Pov_Lev_ACS_14_18*: percent in poverty.

We are interested in models for classifying geographies into COVID-19 community levels reflecting whether or not masks should be worn indoors, based on geographic and demographic characteristics.

(a) Fit the following models

- linear regression,
- logistic regression,
- LDA,
- QDA,
- support vector classifier (SVC),
- support vector machine with polynomial kernel and
- support vector machine with radial kernel,

and obtain an estimate of test error for each of the four models via cross-validation using 0-1 loss [assuming a probability cutoff of $c = 0.5$ for defining class predictions for logistic, LDA and QDA]. For the SVC and SVMs, select the appropriate tuning parameters C, d [for polynomial kernel], and γ [for radial kernel] via cross-validation and provide plots of the test error curve as a function of tuning parameter values. *Note: for SVMs with the two tuning parameters, it is most appropriate to assess test error evaluated over a 2-dimensional grid of possible values. That is, to jointly assess optimal tuning parameters rather than one at a time (fixing the other).* Make sure that the target variable is defined as a factor variable in R.

(b) Discuss how the test error estimates compare across the models. Why do you think the model with the lowest test error has the lowest test error? Why do you think the model with the largest test error has the largest test error? Are the differences in test error practically significant?

(c) For linear regression, logistic regression and LDA, provide an interpretation for the coefficient associated with *pct_HHD_NoCompDevic_ACS_14_18*. How do they compare to each other?

(d) For LDA, state the estimates for $\hat{\boldsymbol{\pi}}$ and μ_k for $k = 1, \ldots, 17$ and explain what these estimates represent.

(e) For the SVC and two SVMs, compare the number of support vectors and discuss what this tells you about each model fit. What does the number of support vectors tell you about the bias-variance tradeoff?

(f) Plot the ROC curves for logistic regression, LDA, QDA, SVC and SVMs all on the same plot. Discuss what this tells you about the predictive power of each model for predicting high community levels for indoor mask recommendations.

(g) For the logistic regression, LDA, and QDA models, now define class predictions using the cutoff $c = \frac{1}{N}\sum_{i=1}^{N} y_i$ and obtain the estimate of test error. How do the test error

estimates compare across models? How do the test error estimates compare to those obtained using the default cutoff $c = 0.5$?

(h) Calculate the sensitivity and specificity for the logistic regression, LDA, and QDA models for both proposed cutoffs $c = 0.5$ and $c = \frac{1}{N}\sum_{i=1}^{N} y_i$. Discuss how these accuracy measures compare across the models and cutoffs. Is it more important to have a better sensitivity or specificity; that is, do you think the two types of misclassification error should be treated equally? (Put your answer in context of the application.) Do you think any of the models yield good enough sensitivity or specificity?

(i) Which model (and threshold) would you recommend and why?

2. (20 points) *Illustrating Nonlinear Boundaries with Logistic Regression, Suport Vector Classifier and SVM for Simulated Data.* Do problem 5 in Chapter 9 (pg. 399) of ISLR: `https://hastie.su.domains/ISLR2/ISLRv2_website.pdf`.

3. (0 points) *Group Topic Request for Special Topic Lecture*

   Please submit a ranked order list of three topics that your group would like to learn and teach to the class. Refer to the "STAT6289 Project Guidelines" posted on Blackboard for a list of suggested topics. Include the list of first, second and third topic preference in the "comments" section when you submit your HW on Blackboard. Only one group member needs to submit this information.