**52525 Take-Home Final – Submit by Aug. 10 2020**

For the take-home we will look at a subset of the Israeli PISA Exams in 2018, made available by the OPEN-PISA organization. The PISA exam is given to a representative sample of the classes in each country, once every 3 years. The core exam produces 3 performance scores for each student: reading (PVREAD), math and science. Students are asked to answer many additional questionnaires on topics such as their background, learning environment, attidutes and many others. In addition, aspects of the school and the classes are measured (we will assume that one class is sampled from each school).

We will model the PVREAD value, and selected p=15 explanatory variables. Some are for each student, and some are for identical for the school. The variables are detailed in the variable table attached.

Instructions

* The code for making your report (functions, results and figures) has to run independently and produce similar results. We will try running your code. Please include all libraries on the top of the run. Please also include compiled output.
* Before you begin, make sure you understand what each of the variables in the data-set means, and look at the marginal and pairwise distributions between variables.
* The work is individual. Please mention explicitly any source you used as your help.

Please work alone. Please do not write your name on the work. I will have

* Please ask Boaz and Yuval any question you are unsure of.

1. [20] Find two predictors that have an interesting association with the reading value (PVREAD). Make a figure (1-2 panels) that describes this association. Briefly (2-3 sentences) discuss the **interplay of the** **three** variables as seen in the figure.

(Two unrelated panels with 2 variables in each will not get full score).

1. [60] Write your own prediction function that uses K-nearest neighbors to predict the PVREAD:
   1. Construct a distance function between students based on the student level variables. Construct a second distance function between students based on class-level variables. The distances should not include the PVREAD variable.   
        
      Instructions: *Keep distances simple, but make sure to account for the scale and marginal distributions of the variables and make sure all variables make sense. You will need to account for the missing attributes (NA) in the different variables: you can do this in the distance function, or in the data matrix before estimating the distance function.*

* 1. Write a function that returns a K-nearest-neighbors prediction based on the distance functions. **Do not use pre-existing libraries.**  
     The function should take a value K, the two distance functions, a relative weight for student distance between 0 and 1, training data, and one or more test point. The function should return the KNN predictions. The distance used in the KNN should be the linear combination:

(depending on the structure of and , you may want to square them before combining.)

* 1. Write two functions that estimate out-of-sample prediction for different K’s and different *a*’s using cross validation. The first should use uniform subsampling to the train and test sets. In the second, the split should be according to classes (each class is either in the training data or in the test data). You can identify classes by having identical class variables.

Present a figure of the prediction accuracy as a function of K and *a* for each cross-validation scheme. Does the optimal K change for different a’s?   
Compare the root-mean square error to std-err of PVREAD.

* 1. For the best value of a and K, extract the residual vector. Try to characterize the residuals of the model (when are residuals large? Can you find a predictor associated with the residuals?). Make 1 plot that shows something you have found / have not found. Briefly (2-3 sentences) explain what you show in the plot.

1. Build a prediction model for the estimated reading level (“PVREAD”). You can use any model you like. The goal is to get good predictions. Please include a brief explanation (1-2 paragraphs) of the process of developing your model.

Submit the following three objects:

* 1. A function called my\_prediction\_model that takes a matrix of predictors in their original form and outputs a vector of predicted-values. Make sure it takes the predictors as they are, with no scaling / centering etc. It is ok to use R objects / functions (such as lm.predict()) inside this function.
  2. Numeric “predicted\_accuracy” with your predicted rmse.
  3. A value “my\_id” with your student id number.

Also, please submit a code file (separate from the report) that generates and saves the three objects. We will run the code compare it to the objects that were submitted.

Comments:

* The code file should include all the steps you did to create the model.
* Check that your “my\_prediction\_model” works before submitting!
* We will check your model on held-out data.
* Make sure the function imports any required library explicitly.

General comments about PISA exams and prediction models

* Please remember this is a toy problem. In practice, the PVREAD outcome is quite stochastic and it is not recommended to make inferences for individual students based on this outcome.
* We should **always** be cautious of using models of student performance as priors for the students; we do not want to “punish” student for belonging to backgrounds that hurt performance. Always ask and think about how your model is used.