Assignment 5

Samuel A, Mikolaj W

March 6, 2020

Problem 2 – CRYSTAL MELTING POINT DATA In their 2005 paper “General Boiling Point Prediction Based on a Diverse Compound Data Set and Artificial Neural Networks”, Karthikeyan, Glen, and Bender examine methods for the prediction of melting points using a number of 2D and 3D descriptors that capture molecular physicochemical and other graph-based properties. The melting point is a fundamental physicochemical property of a molecule that is controlled by both single-molecule properties and intermolecular interactions due to packing in the solid state. Thus, it is difficult to predict, and previously only melting point models for clearly defined and smaller compound sets have been developed. The data frame QSAR.melt contains data for 4401 compounds that can be used to develop a model for melting point. There data are contained in the file QSAR Melting Points (subset).csv on the course website.

Goal: Use PCR and PLS develop models and compare their ability to predict melting point.

Packages:

Data Preperation:

knitr::opts\_chunk$set(include = FALSE)  
QSAR.melt = read.csv("Melting Point QSAR.csv")  
  
#preprocess this data for 203 predictors  
QSAR.bc = preProcess(QSAR.melt, method = "BoxCox")  
QSAR.melt = predict(QSAR.bc, QSAR.melt)  
  
set.seed(1)  
QSAR.melt = QSAR.melt[,-1] # remove Case column which is an ID  
train = sample(nrow(QSAR.melt),3900)  
test = -(train)  
X = QSAR.melt[,-1] # grab all the predictors, Y = MTP is the 1st column  
Xs = scale(X) # scale the predictors  
QSAR = data.frame(MTP=QSAR.melt$MTP,Xs)  
qsar.train = QSAR[train,]  
qsar.test = QSAR[test,]

1. Develop an “optimal” PCR model. Justify your choice using cross-validation. (10 pts.)

Now we will look at a corrplot as well and generate a correlation matrix.

While the NA values do dampen the overall effectiveness, we still are able to seem some correlation bewteen the various variables, we should bode well for our model in terms of variable reduction. Obviously due to our missing values we get uncalculatable correlations, shown as gray in the plot above. We will now fit a pcr model with a bunch of components and use a validation plot to attempt to narrow the number of components used down to a more reasonable number. We will use 40 to start as instructed by the code provided.

Looking at CVs, we see that about 40 components appears to be optimal. Although for variability explained in the melting point, that value first reaches 90 at about 28 components. This is promising, but to be throughout we wanted to look at a wider scale. This lead us to wideing our search to up to 100 components just to see what might be found.

Based on CV values here, we get ever increasing components used, upwards towards our limit, but the returns are ever diminishing. However, in terms of meaningful increases in variation exaplined, nothing past 67 components has a discernable impact, and anything 41 or above is greater then 95%, giving us a decent range of values to investigate further.

To get a better sense of these patterns overall, we also looked at a validation plot.

This plot lines up with our stated results quite nicely, with the late 60s being the last values to truly make a difference while a vast majority of the work is done by roughly 40 of them. Now that we have our candiate values, we can further use cross-validation on the test set to pick the best one.

Using RMSEP as our metric, we can see that the sweet spot appears to be 66 principle components. While that is a lot, compared to the 203 variables we started with that is roughly a 67% reduction in the number or predictors while also dealing with any collineation that was present. Our optimal PCR model uses 66 components.

1. Develop an “optimal” PLS model. Justify your choice using cross-validation. (10 pts.)

Looking at CVs, it seems that by the time we get to 30 components we begin to chase our tails as far as improvements go. This lines up with what we see with percent of variation explained, as we stop seeing any dramatic impacts past around 27 components. Much like with pcr, for the sake of being thorough, we also looked at 100 components, although with pcl especially it is doubtful that anything of greater impact will be found.

The only added conclusion we can make now is that it would take about 53 components to explain 95% of the variation. However, the amount of small increments needed to get there could make one hesitant to use that many components.

Looking at a plot of some internalally generated RMSEPs can help us make a decision as well.

Looking at this plot, we can affirm our understanding based off of the raw numbers, as we begin to start getting diminshing returns around the late 20 values, with anything past it not really gaining anything substantial. From here, we can now to some simple cross-validation on a test set to narrow done to an optimal value.

While not immediatley clear from our plots or numeric measure, based on actual cross-validaion our best metric was returned by a value an amount of components as little as 16. This makes sense given PLS’s nature to return less then pcr, but the ability to reduce the dimensionality by nearly 92% is quite impressive. 16 components will be the basis for our optimal model.

1. Use your “optimal” model for each method (PCR and PLS) to predict the melting point of the test cases. Which modeling method performs the best when predicting the test cases? (10 pts.)

Based on the sums of sqaures of prediction error, we can see that the PLS model performed slightly better, which makes sense given that it had much more variable reduction as well as more heavily focusing on the prediction of y then the organization of the x’compared to the pcr. However, the RMSE was ever so slightly in favor of the pcr. While the difference isn’t major, as shown by the plots, the pls is likely better given its level of reduction. However there is likely still something to be gained from looking at the relevant loadings of the pcr model, as it may help organize the variables which could be useful for unsupervised applications.

1. Use the pcr.cv and pls.cv functions to assess the predictive performance of these two  
   methods. Summarize the results. (10 pts.)

Monte-carlo cross-validation functions accounting for missing values.

Echoing the results of our more rudimentary cross-validation in previous problems, the PLS performs slightyl better over the long term with much more reductive power. However, after 100 samplings, we now can now more confidently state that the pls should perform better as a model on the whole. Between their handling of collinearity and reduction of variables, both methods though certainly earn their stripes and should be kept in mind as an alternative to other methods to handle certain problems.