ISE 5103 Intelligent Data Analytics
Homework #4

Instructor: Charles Nicholson

Due: Monday, October 31, 2016 at 11:59 pm

Learning objective: Perform predictive modeling using regression techniques.
Submission notes:

1. You will submit **multiple** files for this assignment: a written document, your complete R script, and various image files.
2. The write-up is your primary submission.
   - Clearly identify each problem (e.g. Problem 1a, Problem 2b, etc.).
   - You **may** use "R Markdown" to *help* with your submission. Edit the final submission to *clearly and concisely* respond to the questions.
   - Failure to submit this file will result in a grade penalty ($\geq 70\%$).
   - You will be graded primarily on your write-up.
3. You will also submit your R code.
   - *Provide comments* on what your code is doing. Keep it clean and clear!
   - Include `library` commands to load *all* packages that are used in the completion of the assignment at the beginning of your code.
   - Submissions without R code will incur a grade penalty up to 30%.
4. Do not zip your files for submission. Name the files "LastName-HW1" with the appropriate file extension (that is, .R, .pdf, or .docx) – no HTML files. The required image filenames are provided in the problem description.
5. Extra-credit will be awarded for submissions that exceed my high expectations. Please make sure you highlight anything you deem worthy of possible extra-credit. You may receive extra-credit up to 15% of the total assignment points for this assignment.

1 Face Recognition – well, sort of...

Instead of performing facial recognition for this problem, you will need to use those same techniques in order to perform handwritten digit recognition. You will be using handwritten digit files since, they are much smaller files, and the computations will be less likely to eat up all your RAM on your computer. The logic process however is the same as discussed in class regarding facial recognition.

Load the digit file "ClassDigits.csv" from the course website. This file contains data for 30,000 images, each corresponding to a 28×28 B&W image of a handwritten digits, 0, ..., 9. The data file provided for you is a comma separated values (CSV) text file. The first column of the raw data is the label associated with the digit (i.e., if the handwritten digit is a 0, then the label equals 0; if the handwritten digit is a 1, then the label equals 1; etc.) The remaining data are pixel values ranging between 0 and 255.

Your overall goal is to compute the eigenvectors of the digit data and use these eigenvectors to create a significantly lower dimensional representation of the handwritten images in order to perform digit recognition. Note: do *not* scale the image data during the PCA analysis. The image recognition testing will be applied to the 7 observations in the “Class7Test.csv” file.

Tasks:

(a) Compute the eigenvectors of the digit data. *Note: for output, I only need to see a part of the eigenvector data, not the entire set of eigenvectors!*

(b) Create a JPG image of the mean digit. Name this file `meanDigit.jpg`

(c) Reconstruct two training images (image #15 and #100) based on $k = 5, 20$, and 100 principal components. Name these files `image15-5.jpg, image15-20.jpg, image15-100, image100-5`, …”.

(d) Choose a value for $k \ll 784$ based on the PCA summary or a screeplot. Using this value of k, for each of the 7 observations in the *test* data, determine the average mahalanobis distance from “digit-space”. Describe the results.

(e) For the *test* images, 4, 5, and 6, determine the lowest value of k principal components that you need to correctly identify the 10 digits. The value of k may be different for each test image.

Some notes: If you are wanting to create a JPG image of raw numeric data, there are a couple of required steps. First, install and load the `jpeg` library in R.

Next, if you have a numeric vector X of 28×28=784 values, you need to force this into a 28×28 matrix like this: `digitMatrix <- matrix(X,28,28,byrow=TRUE)`

Finally, you will create the JPG file: `writeJPEG(digitMatrix,target="FileName.jpg")` where you should replace “FileName.jpg” with whatever file name you want. I should mention that the JPG data expects the numeric values to be between 0 and 1.

Also note that matrix multiplication in R is performed using `%*%`. The results of the operation is dependent on the the matrices being conformable!

2 Predicting house prices

The `housingData.csv` file in the course website is real data associated with 1,000 residential homes sold in Ames, Iowa between 2006 and 2010. The data set includes over 70 explanatory variables – many of which are factors with several levels. The file `housingVariables.pdf` provides a concise explanation of the variables and the factor levels in the data.

For this problem, you are challenged to make the best possible predictions of the final price of each home. Due to the wide range of variables, you should spend some time exploring the data, and deciding on some type of feature construction and selection strategy.

In this assignment you will build OLS, PLS, and LASSO models to predict the natural log of the sale price, i.e., `log(SalePrice)`.

Given that this is the first time I am giving this problem to students, you may use up to **15 pages** for this problem alone! Yay!

(a) Explore the data, construct new features, and/or collapse (reconstruct) factor levels as you see fit. *Note: w.r.t. grading, I am looking for a logical and thoughtful process which should be supported some visualizations and/or tables.*

(b) OLS Model

   i. Create a hold-out validation set using the first 100 observations in the data. Based on your findings from part (a), build a linear model using `lm` for the remaining 900 observations. You may use a stepwise regression technique if you like or build a model based on hypothesis. For your best model, report the variables, the coefficient estimates, p-values, adjusted R^2, AIC, BIC, VIF, and RMSE.

ii. Provide an analysis of the residuals. Please provide the visualizations that you choose to best depict the residuals as well as any of the metrics we discussed in class that you prefer. Is there anything interesting in the residual pattern? How might this residual pattern influence you to change your model?

(c) Using all 1,000 observations, create a PLS model to predict the log of the sale price. Use hyper-parameter tuning to determine the number of components with RMSE as the error metric (show your chart!). Report the number of components and the CV RMSE estimate for the final model you choose.

(d) Using all 1,000 observations, create a LASSO models to predict the log of the sale price. Use hyper-parameter tuning to determine the penalty value with RMSE as the error metric (show your chart!). For the final model of your choosing, report the variables with non-zero coefficients (and the coefficient values) as well as the CV RMSE estimate.

(e) Using any regression technique or combination of techniques you prefer, predict the final sale price of the data in the file `housingTest.csv` in the course website. You will submit a CSV file with two columns (Id and SalePrice) based on your predictions, e.g.,

```
Id, SalePrice
1, 169000.1
2, 187724.1233
3, 175221
etc...
```

Your predictions will be compared to the true sale price values of the homes described in the test set and your predicted accuracy will be measured. The evaluation will be based on the RMSE of the log of the predicted sale price with the log of the true sale price.