ISyE 8803 – Topics on High Dimensional Data Analytics

Exam II

- For all questions, you are required to clearly state all assumptions you make and show all necessary details of your solutions.
- You are not allowed to discuss the exam content with your fellow students or receive aid on this exam.
- You are expected to observe the Georgia Tech Honor Code throughout the exam.
- Exam is due on July 26 at 11:59 pm. Late submission is NOT accepted. Please submit your solutions via Canvas.

Question 1. Regularization (33 points)

In this problem, you build a set of different models to classify different forest types based on their spectral characteristics at visible-to-near infrared wavelengths observed by ASTER satellite imagery over a forest area in Ibaraki Prefecture, Japan ($36^{\circ}\ 57\ N, 140^{\circ}\ 38\ E$). The training data can be found in "training.csv" and test data can be found in "testing.csv".

Attribute Information:

4 classes: 's' ('Sugi' forest) ; 'h' ('Hinoki' forest); 'd' ('Mixed deciduous' forest) ; 'o' ('Other' non-forest land)

27 features:

b1 - b9: ASTER image bands containing spectral information in the green, red, and near infrared wavelengths for three days (Sept. 26, 2010; March 19, 2011; May 08, 2011).

pred_minus_obs_S_b1 - pred_minus_obs_S_b9: Predicted spectral values (based on spatial interpolation) minus actual spectral values for the 's' class (b1-b9).

pred_minus_obs_H_b1 - pred_minus_obs_H_b9: Predicted spectral values (based on spatial interpolation) minus actual spectral values for the 'h' class (b1-b9).

1- Run a multinomial logistic regression to classify different forest. Present the coefficients obtained. Present the confusion matrix on the test set.

2- Use ridge multinomial logistic regression to classify different forest. Explain why we use ridge regression. Present the optimal tuning parameter obtained using cross-validation, the coefficients for this parameter and the confusion matrix for the test set.

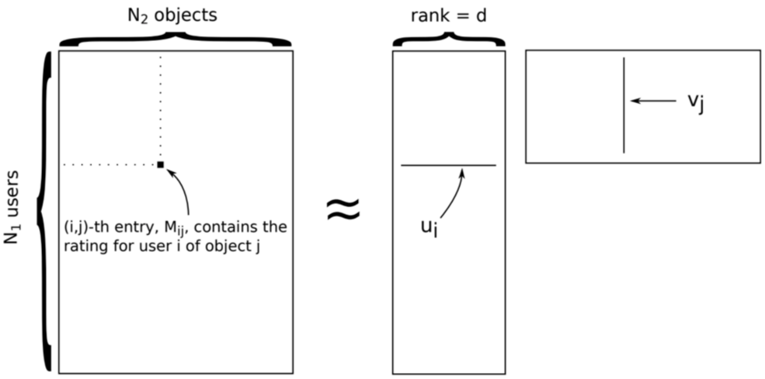
3- Use lasso multinomial logistic regression to classify different forest. Explain why we use lasso. Present the optimal tuning parameter obtained using cross-validation, the coefficients for this parameter and the confusion matrix on the test set.

4- Use adaptive lasso multinomial logistic regression to classify different forest. Explain why we use adaptive lasso regression. Present the optimal tuning parameter obtained using cross-validation, the coefficients for this parameter and the confusion matrix on the test set.

5- Which model do you select? Why?

Question 2. Matrix Factorization with Regularization (33 points)

There is a user-item matrix, M available where nonzero elements of the matrix are ratings that a user has for an item. The objective of matrix factorization is to learn a low-rank factorization of M. It decomposes a large matrix into products of matrices, namely, $M = U \times V$. The picture below demonstrates the idea of low-rank factorization.


N₂ objects
rank = d
N₁ users
(i,j)-th entry, Mij, contains the rating for user i of object j
≈
ui
vj

(1) In order to learn matrix U and V, we will minimize following loss function L. In order to overcome the problem of overfitting, we are going to use matrix factorization with regularization for this problem. We define the loss function L as follows:

$$L = \sum_{i,j} (M_{ij} - U_i V_j^T)^2 + \lambda(\sum_i \|U_i\|^2 + \sum_j \|V_j\|^2)$$

Use the ***Alternating Least Squares*** method to minimize the loss function to learn matrix U and V (Hint: fix U or V one at a time, derive a closed form solution for the other). Write out pseudocode for implementing the algorithm. You should clearly show the closed form solution in the updating step of your pseudocode.

(2) Implement your proposed algorithm using the data located in the the folder ratings.data from MoviesLens Dataset. This dataset consists of ratings of 1682 movies by 943 users (on a 1-5 scale). The test.csv file is generated by removing 10 of the item ratings from train.csv for each user and assign them to test dataset. Plot both training and test errors in terms of MSE versus number of iterations. Please use rank d =20, λ =0.01 and train it for 100 iterations.
(Hint: the $MSE_{test} = \sum_{(i,j)\in S_{test}} (M_{test_{ij}} - U_i V_j^T)^2$ where U and V are learnt using the train.csv and S_{test} contains the locations of all nonzero ratings in the test.csv.)

Question 3. Sparse Representation for Classification (34 points)

A signal not only might be sparse in an SVD or Fourier basis, but also it might be sparse in an overcomplete dictionary whose columns consist of the training data itself. Wright et al. demonstrated the power of sparse representation in a dictionary of test signals for robust classification of human faces, despite significant noise and occlusions. The so-called sparse representation for classification (SRC) has been widely used in image processing.

In this problem, you need to use two data sets. The first data set is the training set: 30 images are used for each of 20 different people in the Yale B database, resulting in 600 columns. Notice that the first 30 columns are the flattened images of one person, the second 30 columns are the flattened images of another person, and so on. In the test set, there are 4 columns and each column correspond to a flatten noisy and occluded image of 7th person in the training set. We want to use sparse representation to classify the images that are in test sets.

Here is what you should do for SRC in this problem:

1- Use the training data set to build an overcomplete library Θ. To use compressed sensing, we need Θ to be a fat matrix. To do so, you need to reshape each column of the training set to 192*168 image and then resize it to 12*10, so the flattened images are 120-component vectors. You can use a built-in function to resize the images and you should normalize each column of Θ by dividing by its norm. Notice that the method of resizing can have an effect on your result; to get better results use the 'lanczos3' method for resizing.

2- You should repeat the same process to the images in the test set: Reshape them to a 192*168 image and resize them to 12*10. You do not need to normalize the test set.

3- Now that you have library Θ and the test set, you can use L1 norm to find the sparse representation of each image in the test set. To do so, you should solve the following optimization for each image in the test set:

$$Minimize\ \left||s|\right|_1$$

$$s.t. \left||\Theta s - Y|\right|_2 < e$$

Where s is the sparse representation, Y is an image in the test set (one column of the test set). Choose the value of **"e=1"** for this optimization. You need to run this optimization on each test set image separately to get the corresponding **s**.

4- The final classification stage in the algorithm is achieved by computing the L2 reconstruction error using the coefficients in the vector **s** vector corresponding to each of the categories separately. The category that minimizes the L2 reconstruction error is chosen for the test image. To do that, consider a sparse representation of an image in test set (**s**)**,** then for all images corresponding to each 20 persons in the original training set compute the following:

Error for person (j) = || [Test image(i) – (All training images of person(j)×normalized **s**)] $||_2$/ ||Test image(i)$||_2$

All the norms in the above equation are L2 norm. The result of the above equation should be 20 numbers corresponding to the 20 persons in the training set. Normalization of **s** should be based on the normalizing values that you found in part 1. Then you would classify test image(i) as person (j) if it has the minimum error.

Deliverables:

A) Show the 4 images in the test set. You should be able to see that these images are noisy and occluded.

B) Plot the 4 vectors of **s** separately (corresponding to each image in the test set). Ideally the resulting vector of coefficients **s** should be sparse and have large coefficients primarily in the regions of the library corresponding to the correct person in training set. Comment on whether you see large coefficients on specific part or not.

C) Plot 4 bar chart correspond to 4 images in test set. In each of these bar charts, you should show the error for person (j) in training set, so it should have 20 bars in it.

D) All images in the test set are related to the 7th person in the training set. Based on the 4 bar charts in the previous section, determine whether or not this method was successful to classify test images.