Please Click https://canvas.ust.hk and SFQ on the left panel To Fill out End-of-Term Course Survey.

Thanks for your attention!

Final Project of MSDM5054, Spring 2021

You need to submit a report and also your source code

Part I. Classification on 20newsgroup Data

Data info: the goal is to classify the types of postings based on their context. The dataset is a tiny version of the 20newsgroups data, with binary occurrence data for 100 key words across 16242 postings. The file "wordlist.txt" lists the 100 key words. The file "documents.txt" is essentially a 16242x100 occurrence matrix where each row is corresponding to 1 posting and each column is corresponding to 1 keyword. The occurrence matrix has binary entries where the (i,j)-th entry is 1 if and only if the i-th posting contains the j-th keyword. Since the occurrence matrix is extremely sparse, the "documents.txt" is a sparse representation of the occurrence matrix. Basically, each line in "documents.txt" represents 1 non-zero entry of the occurrence matrix. For instance, the first line of "documents.txt" is "1 23 1" which means that the entry (1,23) of the occurrence matrix is 1, i.e., the 1st posting contains the 23th keyword.
The file "newsgroup.txt" has 16242 lines where i-th line stands for the group labels of i-th posting. There are 4 different groups which means "comp.", "rec.", "sci." and "talk." respectively. The goal is predict the type, i.e. 4 different group, of the posting based on the words in this posting.

1. Build a random forest for this dataset and report the 5-fold cross validation value of the misclassification error. Note that you need to train the model by yourself, i.e., how many predictors are chosen in each tree and how many trees are used. There is no benchmark. Stop tuning when you feel appropriate. Report the best CV error, the corresponding confusion matrix and tuning parameters. What are the ten most important keywords based on variable importance?
2. Build a boosting tree for this dataset and report the 5-fold cross validation value of the misclassification error. Similarly, report the best CV error, the corresponding confusion matrix and tuning parameters. Note that the R example in the textbook only considers binary classification. But the library 'gbm' can deal with multi-class case by setting 'distribution=multinomial'.
3. Compare the results from random forest and boosting trees.
4. Build a multi-class LDA classifier. Report the 5-fold CV error of misclassification and the confusion matrix.
5. Build a multi-class QDA classifier. Report the 5-fold CV error of misclassification and the confusion matrix.
6. Compare the performances of all above methods and give your comments.

Part II. Spectral Clustering on 20newsgroup Data

1. Apply PCA on the binary occurrence matrix and apply K-means clustering. Basically, take the top 4 left singular vectors of the occurrence matrix (of size 16242x100) and apply K-means on the rows of these singular vectors with K=4. Report the mis-clustering error rate.
2. Now take the top 5 left singular vectors of the occurrence matrix and apply K-means on the rows of these singular vectors with K=4. Report the mis-clustering error rate.
3. Compare with the performances from part I.

Part III. Classification on MNIST Data

Dataset: MNIST/train_resized.csv, MNIST/test_resized.csv
Description: train_resized.csv has 30000 rows and 145 columns, test.csv has 12000 rows and 145 columns. Each row is corresponding to 1 handwriting digits. The first column label denotes the actual digit that can be 0,1,…,9. The remaining 144=12*12 column are the pixels of one image, so each image is of size 12x12. Some example images are as follows.

Note that the original image is of size 28x28. I have downsized it to 12x12 to make your computation faster. As a result, the image pixel values are not 0 or 1 anymore.

1. Use only the digit images of 3 and 6 from train_resized.csv and test_resized.csv to build an SVM classifier for binary classification. More specifically, use a linear kernel and choose the best cost (the data size is large so a large cost value is suitable) parameter (called budget in our course) by 5 fold cross validation. Apply your model on the test data and report the misclassification error, confusion matrix. Also report the time cost of training your model.
2. Use only the digit images of 3 and 6 from train_resized.csv and test_resized.csv to build an SVM classifier for binary classification. More specifically, use a radial kernel and choose the best cost parameter, gamma parameter by 5 fold cross validation. Apply your model on the test data and report the misclassification error, confusion matrix. Also report the time cost of training your model.
3. Compare the results of the above two models and report your comments.
4. Use only the digit images of 1,2,5 and 8 from train_resized.csv and test_resized.csv to build an SVM classifier for multi-class classification. More specifically, use a linear kernel and choose the best cost parameter (called budget in our course) by 5 fold cross validation. Apply your model on the test data and report the misclassification error, confusion matrix. Also report the time cost of training your model.
5. Use the complete dataset of train_resized.csv and test_resized.csv to build an SVM classifier for classifying all 10 classes. You can use any SVM model and tune the parameters by yourself. Report the best test performance (misclassification error) you can get, the model you used and the time cost of training your model.

Part IV. Additional Bonus

Try any other machine learning methods you know, not necessarily limited to this course, on this MNIST dataset. Report the best test performance you can get, the model and the time cost of training your model. How does it compare to part III?