Assignment #4 — Due Saturday, November 23rd, 2024

*Submit your homework on Canvas.

*No late homework will be accepted for credit.

*Append the codes you used to your submission.

Problem 1: Basics Knowledge (100 pts)

1. Consider the maximal margin classifier $\arg\max_{\beta,\beta_0} \frac{1}{\|\beta\|}\min_{1\le i\le n}\big[y_i(\beta^\top x_i+\beta_0)\big]$. Show that it can be equivalent formulated as

$$\begin{aligned} \text{Maximize}_{\beta,\beta_0;M\ge 0} \quad & M \\ \text{Subject to} \quad & \|\beta\|=1 \\ & y_i\big(\beta^\top x_i+\beta_0\big)\ge M, \quad i=1,\cdots,n \end{aligned}$$

2. Show that the linear support vector classifier can be represented as

$$f(x)=\beta_0+\sum_{i=1}^{n} a_i\langle x_i, x\rangle,$$

where $a_i\neq 0$ only for support vectors and a_i can be computed based on $\langle x_j, x_k\rangle$ and y_j.

3. Explain how the parameter $\gamma>0$ in the Radial kernel affects the bias and variance in SVM.
4. How does PCA handle missing data, and what are the best practices for preprocessing before applying PCA?
5. What is K-means++ method? Discuss its advantage over the K-means algorithm.

Problem 2: Classification on 20newsgroup Data (100 pts)

Dataset: wordlist.txt, documents.txt, newsgroup.txt, groupnames.txt
Description: 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 16,242 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.

Problem 3: Classification on MNIST Data (100 pts)

Dataset: MNIST/train_resized.csv and MNIST/test_resized.csv
Description: train_resized.csv has 30,000 rows and 145 columns, test.csv has 12,000 rows and 145 columns. Each row is corresponding to 1 handwriting digits. The first column label denotes the actual digit that can be $0, 1, \cdots, 9$. The remaining $144 = 12 \times 12$ column are the pixels of one image, so each image is of size 12×12. Some example images are as follows.

Note that the original image is of size 28×28. I have downsized it to 12×12 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.