MUSA650_Spring2022_HW2

March 1, 2022

HW2 - Satellite image classification

Include your code in the relevant cells below. Subparts labeled as questions (Q1.1, Q1.2, etc.) should have their answers filled in place or plots placed prominently, as appropriate.

Please make sure to remove irrelevant code or outputs, and to include descriptive comments with all of your code.

0.0.1 S1:

- Load the Planes in Satellite Imagery dataset: https://www.kaggle.com/rhammell/planesnet.

Q1.1: Visualize a few of the images for different labels.

Q1.2: What is the total number of images in this dataset: _____

Q1.3: What is the number of labels in this dataset: _____

Q1.4: What is the dimensionality of each image in this dataset: _____

0.0.2 S2:

- Create data matrices X and y as follows:
  - Extract the color channels from each image and flatten them to a feature matrix X (*Hint: use the included JSON file to do this easily*).
  - Create the labels y (with binary labels 0 and 1) for each image.
- Using X and y, create a split dataset with 70% training and 30% testing data with similar distributions for the two classes.

Q2.1: What is the size of X (before splitting): _____

Q2.2: What is the size of y (before splitting): _____

0.0.3 S3:

- Train a SVM classifier using the Sigmoid kernel (with default values for other parameters) on the training data and use it to predict labels of the testing data.

Q3.1: What is the training accuracy: _____

Q3.2: What is the testing accuracy: _____

Q3.3: Show the confusion matrix for the classification of testing samples.

Q3.4: What is the AUC (area under the curve) for the classification of testing samples : _____

0.0.4 S4:

- Train a new SVM classifier using the RBF kernel (leaving other parameters as their default values), and use it to predict labels of the testing data.

Q4.1: What is the training accuracy: _____

Q4.2: What is the testing accuracy: _____

Q3.3: Show the confusion matrix: _____

Q3.4: What is the AUC (area under the curve) for the classification: ______

Q4.5: Using the metrics of accuracy and AUC, which of your models is better? Is there anything else to consider?

0.0.5 S5:

- You will now try to find the best value for the regularization parameter C from among the values [0.1, 1, 10].
- Train a SVM classifier with Sigmoid kernel with different values for C using leave-10%-out cross-validation within your training data.
- Train a SVM classifier with RBF kernel with different values for C using leave-10%-out cross-validation within your training data.
- Select the best model parameters (from the 6 models: 2 SVM kernels x 3 parameters) based on highest cross-validated accuracy. Train the selected model on the complete training set and apply on the testing set.

Q5.1: What is the best choice of C for the linear kernel: _____

Q5.2: What is the best choice of C for the RBF kernel: _____

Q5.3: What is the accuracy of best model on testing data: _____

0.0.6 S6:

Unsupervised learning: Clustering

- Apply k-Means clustering with k=2 on the complete set using vectorized imaging features

Q6.1: What is the distribution of plane vs non-plane images into the two classes? Show it with a 2x2 matrix: _____

Q6.2: Does the clustering (without using actual labels in learning) work for detecting the two target classes? Why or why not?

0.0.7 S7:

Visualization

- Show the average of all airplane images
- Show the average of all images in your first cluster in S6
- Show the average of all images in your second cluster in S6

Q7.1: Based on what you see in these average images, what would you suggest to improve your classifier for differentiating these two classes?

0.0.8 S8 (Bonus):

Extract only 5 features from each image [1]. Train a linear classifier with default parameters on the training data using only these 5 features and apply it on the test data.

[1] Feature extraction should be done without using the class labels

Q8.1: Describe how you extracted your 5 features. How did you choose them?

Q8.2: What is the accuracy of your classifier on the test data: _____