Machine Learning CS 550

Programming Assignment 2: Classification

Total Marks: 150 (obtained marks will be divided by 10 for grading)

Note: 50 marks bonus for top 3 solutions (Accuracy and F1 score) for Part B.

Instructions: This assignment must be done by individually. Each student must understand the code and techniques used and be able to answer each question. It is fine to use web-resources and help, provided you reference them and understand the code. **Please don't copy code from the internet resources blindly.**

Due Date: Sep 5, 2022 (End of day): **Hard Deadline**

Late Submissions: There are **ZERO late days** allowed for this assignment. We won't consider late submissions. Submit whatever you have completed by the deadline.
4 out of 7 of your best assignments will be considered in the final grade. That can be used to handle emergencies, sickness etc., so in general no exceptions will be made.

Submission: A Jupyter notebook template containing instructions is being posted along with the assignment.

- **i.** Jupyter Notebook: add cells at appropriate locations & write down your solution. Try to give proper justification of your answer if needed. We may ask you for a demo.
- **ii.** Output of Recommendation for Part-B. Use the format posted on Git-Hub.

Good solutions will be posted on the Course Git-hub.
Write all your references used in completing the assignment at the beginning of your Notebook in the space provided.
If you change the values in the cell manually to generate outputs and try different combinations, we encourage you to show that work in your notebook with appropriate comments. However, irrelevant details should be put at the end as supplementary work.

Note: Colab or link files will not be accepted. Your submission must be one single jupyter notebook (.ipynb) file and the name of the file must be, FirstName_RollNumber.ipynb (Like, for Anirban Haldar, 12110240, the file name should be, Anirban_12110240.ipynb)

Part A: Multi-class Classification [105 marks]

Captchas were invented to prevent bots from attacking websites. But the ML models are breaking captchas. Thus, stronger captchas are being invented. In this part, we will train a ML model to break hcaptcha (https://www.hcaptcha.com/)!!
Please download the following dataset: https://github.com/sarang-iitb/H-captcha-dataset
Here is a sample of the images in the dataset.

1. [20 marks] **Data Cleaning & Visualization**
   Reference: https://www.geeksforgeeks.org/python-grayscaling-of-images-using-opencv/
   a. [1] Check out the labels in the dataset. How many images belong to each class?
   b. [1] How many train and test images are present in your dataset?
   c. [4] Write a function to display a random image and its shape. Find out whether the shape of each image is the same or not. If not then make all images of the same shape.
   d. [4] Do you think removing color channels (R, G, B) from images would lead to poor modeling? How can you justify converting each image to greyscale? What will be the effect of using a colored 3-channel image over a grayscale one on the classification model's performance?
   e. [4] Should you normalize your color channel values? Based on your answers do the steps you think will be best for your model.
   f. [6] Visualize 3 random training images along with the labels for each class. The dataset was manually labeled, do you spot any errors in the labels?

2. [10 marks] **Preparing Balanced Samples for Training using only the Training set provided to you.**
   Reference: https://scikit-learn.org/stable/modules/cross_validation.html
   a. [2] Do you think if you apply cross-validation to the dataset then all the cross-validation folds will be similar? Why or why not?
   b. [8] What can you do to ensure that every fold contains images from each class and no duplicates? Do the needful on your dataset that will ensure similar cross-validation folds.

3. [5 marks] Is Logistic regression a good algorithm for your dataset? Why or why not? Based on the classification algorithms taught in class, which algorithms can be used to solve this problem?

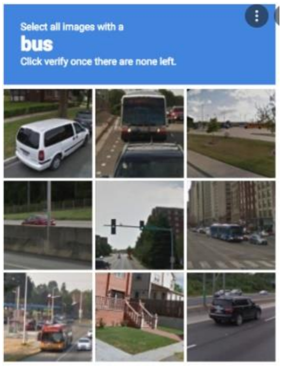
4. [15 marks] Train the KNN classifier algorithm on the training dataset. Don't use the test set at this time.
   Reference: https://docs.scipy.org/doc/scipy/reference/spatial.distance.html

   a. [5] What distance/similarity function should we use for comparing images? Compare at least 3 different metrics on a few random samples to get some understanding of how they work.
   b. [5] What is the impact of increasing K on the speed and accuracy of the algorithm?
   c. [5] Choose a good value of K and the distance metric based on cross-validation.

5. [15 marks] Train SVM classifiers on the dataset. Don't use the test set at this time. Reference: https://scikit-learn.org/stable/modules/svm.html
   a. [5] Which kernel functions can be used for this dataset? Why?
   b. [5] Which of the SVM implementations (`SVC`, SGDClassifier, `NuSVC` and `LinearSVC`) is the best for this dataset? Why?
   c. [5] Finalize your SVM model based on cross-validation

6. [25 marks] Evaluate and compare the classification models on the test set provided to you.
   a. [5] Calculate the classification accuracy of both the models and compare.
   b. [5] Which are the two classes that were most confused by your model? What classification evaluation metric can be used to best visualize it? Plot the visualization for both models.
   c. [2] Do you think you can maximize both Precision and Recall for a model? What do you think will be a better metric to judge a model?
   d. [8] A bridge in Raipur is unstable and we want to avoid fatalities. Trucks and Buses should be stopped while motorbikes and bicycles can be allowed. Which of your 2 models do you think will be better for this task? Explain with the help of Precision/Recall, ROC curve and AUC.
   e. [5] Find the Micro and Macro F1 Score of both of your models and comment on which is better and why.

7. [10 marks] Write a function to create a random 3x3 captcha matrix and ask the user to label the class with maximum number of examples. Estimate how often your best model will be able to solve the captcha problem correctly.
   Example output: Don't focus on graphics, just the concept.



Select all images with a
bus
Click verify once there are none left.

Part B: Ensemble Models [50 marks]

You started a Car Selling business and you are giving recommendations to people for buying cars. We have provided a dataset to help you start your business.
Dataset: https://github.com/sarang-iitb/Car_Condition_evaluation_dataset

Now, your job is to train a robust model and we will test how good you are on the test set.

- [20 marks] **Data Preparation:** Perform necessary transformations on the input dataset to prepare it for ML model training.
- [30 marks] **Ensemble Model:** Train any ensemble model of your choice combining your favorite models. In particular, we would like you to try both bagging and boosting.
  - [15] Bag of models of your choice
  - [15] Adaboost or XGBoost

- **BONUS [50 marks]**

We have released a test set. Top 3 solutions (Accuracy and F1 score) for Part B will receive a bonus of 50 marks.