Final Project

ITCS 6156 / ITCS 8156 – Spring 2016

For this project, you will address a real-world image classification problem, implement a solution and describe the results of your experimentation. The goal is to classify images from hotels into one of 8 classes: `'Bathroom', 'Guestroom', 'Pool', 'Gym', 'Restaurant', 'Lobby', 'Aerial View', 'Business Center'.` You will be provided training examples with (somewhat noisy) ground truth labels and unlabeled testing images.

Description and Data

The data consists of JPG images (of varying size) from each of the 8 classes and can be downloaded from the link below. *Note: the data can only be downloaded to a machine connected to the UNCC network.*

http://orwell.uncc.edu/6156/project.tar (1.5GB)

All of the training and testing images are named <id>.jpg, where id is the example id used to report results. Testing results will be submitted to Kaggle as a CSV file, similar to train.csv, where row contains the image id and posterior probabilities for each class. For the training data, the correct class is 1 and the rest 0. For your results, this will not likely be the case. The evaluation metric will be multi-class log loss.

Competition

This project will be hosted on Kaggle. Once invited, you should go to https://inclass.kaggle.com/ and find the class project. When you register, make sure your username obviously maps to your real name. Once registered, be sure to pay special attention to the rules of the competition and the format of your results. As the competition progresses, you will be able to see a "public" leaderboard. These results are a ranking based on a randomly selected subset of the solutions that you provide (20% of the testing data). Your *real* score (on 100% of the testing data) is kept on a "private" leaderboard that won't be made available until the end of the competition.

The competition **closes on May 6th**, so be sure you have at least one valid attempt submitted to Kaggle.

The goal of this project is for you to apply machine learning techniques to a real-world problem. It is expected that you will use both material from class and outside resources (e.g., papers, code libraries) as part of your work. Any attempt to obtain labels for the test examples using approaches outside the realm of machine learning (e.g., manual labeling, image-based searching) will treated as an academic integrity violation and dealt with harshly.

Project Milestones

Credit will only be given for milestone submissions submitted **on time via Moodle**. Do not wait until the last minute to make uploads to Moodle, as it will likely be slow at the end of the semester. No exceptions will be made for the deadlines.

prj0: Project Checkpoint [20 pts]

For this milestone, you must download the data for your project, make at least one submission to Kaggle, and submit a short description of your plan for the project. The majority of the credit on this assignment will be based on either evidence of progress on the project (e.g., multiple non-random results submissions) and/or my assessment of the reasonableness of your proposed approach

prj1: Classification Performance [30 pts]

Your score for this portion will be based on the final ranking of your algorithm, based on the private leaderboard at the end of the competition.

prj2: Project Code [30 pts]

You will submit your project code and a README describing how to run the code and recreate the results that you submitted to Kaggle. Your code will be graded based on the Code Submission Guidelines described in the course syllabus and how closely it matches your project report. You are allowed to use external helper functions and libraries (giving appropriate credit, of course). This is a programming assignment, so using a GUI-based tool (e.g., Weka) will not get you much credit.

prj3: Project Report [50 pts]

You will submit a report on your project, which describes your approach, experiments, and results effectively. This will be graded mainly on thoroughness, extensiveness of your experimentation, and clarity. Your score for this report could be negatively affected by describing work different than the code submission or misrepresenting your work and/or results.