Lab Two: Exploring Image Data

Due Feb 18 by 11:59pm **Points** 10 **Submitting** a file upload

Lab Assignment Two: Exploring Image Data

You are to perform preprocessing and exploratory analysis of a data set: exploring the statistical summaries of the features, visualizing the attributes, and addressing data quality. This report is worth 10% of the final grade. Please upload a report (**one per team**) with all code used, visualizations, and text in a rendered Jupyter notebook. Any visualizations that cannot be embedded in the notebook, please provide screenshots of the output.

Dataset requirements: Choose a dataset that is comprised of image data. The data should be directories of images. That is, the dataset should not yet be pre-processed. The following are required for the dataset:

1. The data includes at least 1000 images
2. The size of the images should be larger than 20x20 pixels
3. The dataset should have a well defined prediction task (i.e., a label for each image)

***A note on grading**: This lab is mostly about visualizing and understanding your dataset. The largest share of the points is from how you interpret the visuals that you make. Making the visuals is not enough to satisfy each of the rubrics below—you should appropriately explain what the implications of the visualizations are. In other words, expect about 20% of the available points for visuals that have no substantive discussion.*

Grading Rubric

- Business Understanding (**20 points total**).
  - [**20 points**] Give an overview of the dataset. Describe the purpose of the data set you selected (i.e., why was this data collected in the first place?). What is the prediction task for your dataset and which third parties would be interested in the results? Why is this data important? Once you begin modeling, how well would your prediction algorithm need to perform to be considered useful to the identified third parties? Be specific and use your own words to describe the aspects of the data.
- Data Preparation (**10 points total**)
  - [**5 points**] Read in your images as numpy arrays. Resize and recolor images as necessary.
  - [**4 points**] Linearize the images to create a table of 1-D image features (each row should be one image).
  - [**1 points**] Visualize several images.
- Data Reduction (**60 points total**)
  - [**5 points**] Perform **linear** dimensionality reduction of the images using principal components analysis. Visualize the explained variance of each component. Analyze how many dimensions are required to adequately represent your image data. Explain your analysis and conclusion.
  - [**5 points**] Perform **non-linear** dimensionality reduction of your image data.

- [**20 points**] Compare the representation using non-linear dimensions to using linear dimensions. The method you choose to compare dimensionality methods should quantitatively explain which method is better at representing the images with fewer components. Be aware that mean-squared error may not be a good measurement for kPCA. Do you prefer one method over another? Why?
- [**10 points**] Perform **feature extraction** upon the images using any feature extraction technique (e.g., gabor filters, ordered gradients, DAISY, *etc*.).
- [**20 points**] Does this feature extraction method show promise for your prediction task? Why? Use visualizations to analyze this questions. For example, visualize the differences **between statistics of extracted features** in each target class. Another option, use a heat map of the pairwise differences (ordered by class) among all extracted features. Another option, build a nearest neighbor classifier to see actual classification performance.

- Exceptional Work (**10 points total**)
  - You have free reign to provide any additional analyses.
  - One idea (**required for 7000 level students**): perform feature extraction upon the images using a feature extractor that requires key point matching (such as SIFT/SURF/ORB or others). Then build a nearest neighbor classifier using a method appropriate for your chosen features. You will need to investigate appropriate methods for comparisons with your chosen feature extraction technique. NOTE: this often requires some type of brute force matching per pair of images, which can be computationally expensive).