ISyE 6416 – Computational Statistics - Spring 2016
Project: "Big" Data Analytics
Proposal

Team Member Names: Simon Chow, Pravara Harati

Project Title: What Industry Are You?

Problem Statement

Our project is to build an image classification system. Given the headshot of a person, our classifier will determine which industry the person is likely to belong to. This could have numerous practical applications if a successful classifier is found. For example, based on photos a person posts on social media, it could be possible to determine which industries the person belongs to or is interested in. That person could then be targeted with papers or news articles concerning that industry or advertisements for products typically produced by that industry. Applications designed to match people together could reasonably guess a person's industry from their profile picture, even if they do not choose to list it, and then use that information as part of the criterion for determining the best match.

To build this classifier, we will collect our own data using a subset of the list of the richest people in the world published by Forbes. We will then test various methods of classification, including categorical regression, k-means, Gaussian mixture models, and principle component analysis, and compare their results. Our goal is to determine which classification method is most successful in determining a person's industry and to what extent it performs better than the others.

Data Source

Every year in March, the Forbes Magazine publishes a list of the world's richest people. This list is an estimate in United States dollars and estimates the net worth of a person by counting assets and deducting debts. The list excludes royalty and government figures who acquire wealth from their positions. We scraped the list of the five hundred richest people from the Forbes website, and also extracted the pictures of the five hundred richest people as displayed on the Forbes website. Each of these images is exactly 416 x 416 pixels in size and consists primarily of the person's headshot. We removed people from the list with no picture shown on the website, amounting to 26 people, leaving us with a total of 474 people in our dataset. Forbes classifies each person into one of 18 industries based on how they obtained their wealth. These industries include construction and engineering, metals and mining, automotive, and telecom, among others. Because some of these industries have few people, we will further group them, resulting in ten total industry classifications: healthcare and service, media and entertainment, energy and engineering, diversified, manufacturing, food and beverage, real estate, technology, finance and investments, and fashion and retail. Each of these industries have at least 35 and at most 74 members in our sample, with media and entertainment having the least and fashion and retail having the most.

Methodology

For our initial approach, we will build the 'representative' face for each of the industries we are using for classification. Before we construct the representative face for each industry, we will first have to standardize our data. This will involve checking and resizing all pictures so some common facial features are in roughly the same place for each person. For instance, we want the eyes and mouths to be in roughly the same place in 2D space for each of the people in our dataset. After this is completed, we will need to trim the images so they are of all the same height and width in pixels. After this is complete, we will split the faces into five groups in each industry. These five groups will be used for 5-fold cross validation of our system. We will choose four of the groups for training the classifier, and then use the last group for testing, repeating five times until all groups have been used as the testing data. We will then select the classifier that performed the best on the training and test data as our final classifier.

Our first approach for constructing representative faces and classifying them will be to compute an 'average' face for each industry. We will do this by taking the mean of all the pixels at each position in the image, generating a mean face for each industry. After getting a mean face for each industry, we will then use the average image in each of our classifiers. We will test categorical regression, gaussian mixture models, and k-means to classify our images into categories, using the 5-fold cross validation technique for testing as described before.

The second approach we will use is to the common approach of computing eigenfaces. For each image in a given industry, we will subtract the mean face that we computed from the first step. We then aggregate all images per industry into one large matrix. Then, using this matrix, we calculate the eigenvectors/eigenvalues of the covariance of this matrix of images. After this is complete, we will sort the eigenvalues from largest to smallest, and use the top k eigenvalues to be the representative eigenfaces for the industry. In practice, people have used between 100 and 150 eigenfaces. Similar to the first approach for the classification, we divide the images into 5 groups and apply the 5-fold cross validation to select the best set of eigenfaces.

Our plan for the rest of the semester has three parts. First, we will preprocess the images that we have collected. As stated in the methodology, this will involve resizing and aligning all pictures so they are of similar shape with similar facial feature locations, as well as trimming pictures to be of identical sizes. Second, we will split the testing of the classifier into two parts. Pravara will compute the average face and following the methodology, test categorical regression, gaussian mixture models, and k-means to build the best classifier for the images. Simon will take the second approach and use the common eigenfaces methodology to classify faces. Third, we will compare and contrast the different methods. Using the same training and testing datasets, we will evaluate which overall approach is the best, as well as attempt to quantify why one particular method is better than another.

Expected Results

Due to our relatively small sample size, we do not expect to have extremely high accuracy with any of our models. Most of the industries have fewer than 50 people assigned to them, which may not be enough to determine what is representative of someone in that industry. Another issue is that we are considering only the 500 richest people due to their industries being listed and their conveniently similar pictures being provided. Therefore our results will likely not hold if we expand to consider the entire

population since there will be more variety in people's images. We do, however, expect supervised learning methods to have a higher success rate than unsupervised methods since they will be given an idea of what each industry looks like during training while unsupervised methods may end up categorizing people together based on traits other than industry.