Homework 2 Programming Write-up

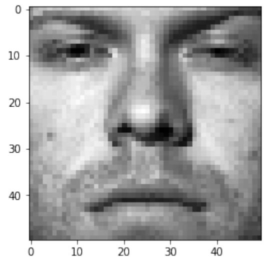
Jingxuan Sun, js3422
Yue Wang, yw986

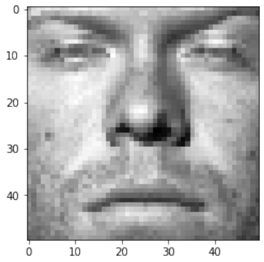
November 6, 2018

1 Eigenface for face recognition

1.1 Observing Normalizing Data

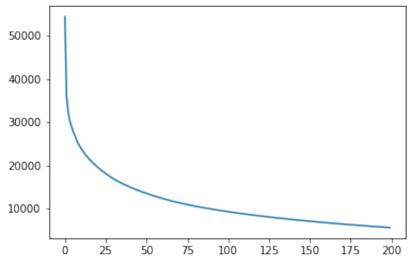
Training set and testing set are in the same format: image path and label. There are 10 labels and 540 images for train set and 100 images for test set. Each image is 50*50 pixels and read as 2500-dimensional vector.
For normalization, first calculate the average face for both train set and test set. And then do the mean subtraction to all dataset.





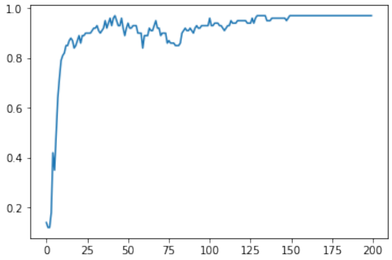
1.2 SVD and Low Rank Approximation

Perform singular value decomposition on normalized train data.$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, where D presents the eigenvalues is descending order, and V has the same dimension as image, which is the eigenface.
In order to reduce dimensions of data (which is 2500 right now), we would only choose the first r eigenvalues in D, and the approximated image is $\mathbf{X} = \mathbf{U}[:,:\mathbf{r}]\mathbf{D}[:\mathbf{r}]\mathbf{V}[:\mathbf{r},:]^\top$. This is called rank-r approximation. To visualize how ”well” the rank-r tells the original image, we use approximation error.



1.3 Logistics Regression

As mentioned in section 1.2, we processed our original images into approximated one, and the question now is how to choose a proper r for test set. Thus, we conduct logistics regression on different r from 1 to 200, and plot their accuracy.



2 What's cooking

2.1 Observing Data

Training set and test set are both in json format. There are in total 39774 training samples and 9944 test samples. For cuisines, we loaded the ”cuisine” entry of the training samples into a set to eliminate the duplicate, so there turned out to be 20 cuisines/classes. For ingredients, we also created a set, but we loaded the ingredients from both training and test samples. Because in further steps, we wanted to encode the feature vectors into a binary representation, one ingredient as one feature. We thought models should observe a global set of features, although some of them might not appear in training data. Thus, there are in total 7137 unique ingredients.

2.2 Feature Encoding

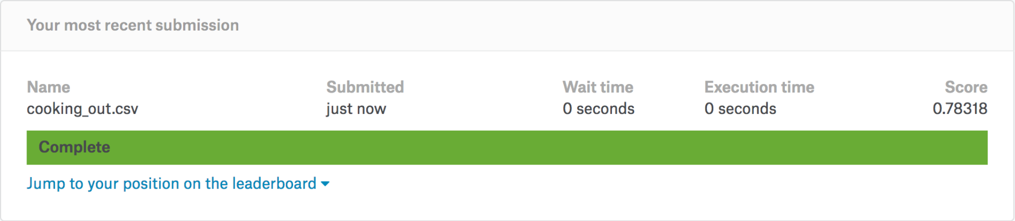
As mentioned in Section 2.1, we encoded binary features with dimensional of 1*d, where d equals to the number of unique ingredients 7131, where xi = 1 if dish contains ingredient i and xi = 0 otherwise. All row vectors were stacked together to form the feature matrix with dimension of n*d, where n equals to the number of training samples 39774. More specifically, we initiated a n*d zero matrix and updated corresponding entry to 1 for those included ingredients for each dish.
Actually, we tried to print out the unique ingredients set, and found that many of them are too specific and similar to each other. We think a preprocessing step might be useful to group similar unique ingredients together so that dimension reduction could be achieved and future calculation might be simplified.

2.3 Cross Validation on Naive Bayes Classifier

Both Gaussian and Bernoulli Naive Bayes classifiers and cross validation was directly performed through sklearn library. The average accuracy of Gaussian distribution was 0.38 while that of Bernoulli distribution was 0.68. That Bernoulli prior distribution performed much better was reasonably anticipated because the features were literally encoded in such fashion, and the underlying distribution was not necessarily falling in a Gaussian.

2.4 Cross Validation on Logistic Regression

The average accuracy achieved was 0.78. After training on logistic regression with full set of training samples, the test accuracy was 0.78. Naive Bayes might still be too naive because for a certain type of cuisine, dishes might use a similar set of ingredients so that each feature has certain degree of correlation.



Your most recent submission

Name	Submitted	Wait time	Execution time	Score
cooking_out.csv	just now	0 seconds	0 seconds	0.78318

Complete