2018FA_MSDS_422-DL_SEC58

Practical Machine Learning Assignment 5

Prabhat Thakur

1. Problem Definition

In this assignment we worked on the MNIST dataset, which is a set of 70,000 small images of digits handwritten by high school students and employees of the US Census Bureau. Each image is labeled with the digit it represents. For this assignment we will develop a classifier that may be used to predict which of the ten digits is being written.

Our task is to evaluate the performance of two models 1st: random forest classifier model and 2nd: random forest classifier model based reduced dimensionality achieved by principal components analysis (PCA).

2. Research Design and methods

There are 70,000 images, and each image has 784 features. This is because each image is 28×28 pixels, and each feature simply represents one pixel's intensity, from 0 (white) to 255 (black).

As per the assignment instructions, following design is used to create two models:

1) Fit a Random forest classifier using full set of 784 explanatory variables and training set of 60,000 observations.
2) Execute principal components analysis (PCA) on full set of 70,000 observations, generating principal components that represent 95 percent of the variability in the explanatory variables. Record the time it takes to identify the principal components. Now using the identified

principal components, use 60,000 observations (training set) to build another random forest classifier.

Record the time it takes to fit both models and evaluate the models on the test data. Assess classification performance of both models using the F1-score.

3. Implementation and Programming:

Python programming language and different packages like numpy, pandas and matplotlib for data preparation and visualization are used. Classes and methods from sklearn package for building RandomForestClassifier (RFC) model, PCA analysis and model performance evaluation is used.

The first step is to read the MNIST data from mnist-original.mat file into python dictionary object and create separate array objects so that it can be used in sklearn package. Feature scaling is not performed as all features have same scale (0 to 255 pixel intensity) and also for classifier algorithms feature scaling is not very useful. I divided the data into training set (60000 images) and test set (10000 images).

Now I fit RFC on training data set. Next, I generated principal components that represent 95 percent of the variability in the explanatory variables and using identified principal components build another RFC model on training set. For both models I recorded execution time and evaluate performance measurement using F1-score on test set of 10000 images.

4. Findings and recommendations

Following are the findings from the experiment:

Model 1: RFC model using all 784 features on training set of 60000 images.

Model 2: Principal components analysis (PCA) on the full set of 70,000, generating principal components that represent 95 percent of the variability in the explanatory variables. It reduced dimensions to **154** variables from 784 variables. RFC model on identified principal components (154 variables).

Models	**Avg Time of 10 runs in sec**	**F1-Score**
Model 1 (RFC)	12	76.51%
Model 2 (PCA + RFC using Principal components)	13 + 18 = 31	77.72%
Model 3 (PCA on training dataset + RFC using Principal components on training dataset)	11 + 18 = 29	78.16%

I build another model, **Model 3** in which PCA was performed on the training set instead of entire dataset. We can see the performance of model 3 is better than model 2.

I believe, **choosing entire dataset for PCA could be the flow in the design.** Another important thing to notice is after dimensionality reduction, the training set takes up much less space. So while most of the variance is preserved, the dataset is now less than 20% of its original size. It may not be helping speed up RFC algorithm but this can speed up SVM classifier tremendously.

Another flow could be not exploring SVM classifier or any other classifier.

Recommendation: Considering 1.6% F1-score improvement and less memory need, management should use PCA analysis for dimension reduction.