PROJECT REPORT: C S 487/519 Applied Machine Learning I

Project 6: Dimensionality reduction techniques

Written by

Omid Jafari

Department of Computer Science

New Mexico State University

November 2018

Preprocessing and notes:

For the iris dataset, we are using the last column as the label and converting the categorical values in to numbers.

For the digits dataset, we are using SKlearn to load x and y.

We define accuracy as the number of correctly clustered instance over total number of instances.

datasets are standardized and split into 70% training and 30% testing data.

First, we printout the accuracy of Logistic Regression without doing any dimensionality reduction. After that, we run different dimensionality reduction algorithms and feed their result to Logistic Regression.

Results:

Running time of Iris dataset (ms)		
Regular Logistic Regression	Training	2
	Prediction	8
PCA	5	
LDA	1	
Kernel PCA	19	
Logistic Regression on PCA results	Training	1
	Prediction	0
Logistic Regression on LDA results	Training	1
	Prediction	0
Logistic Regression on Kernel PCA results	Training	1
	Prediction	0

Accuracy of Iris dataset (%)	
Regular Logistic Regression	95.56 %
Logistic Regression on PCA results	84.44 %
Logistic Regression on LDA results	84.44 %
Logistic Regression on Kernel PCA results	33.33 %

Running time of Digits dataset (ms)		
Regular Logistic Regression	Training	273
	Prediction	0
PCA	14	
LDA	2	
Kernel PCA	86	
Logistic Regression on PCA results	Training	7
	Prediction	1
Logistic Regression on LDA results	Training	10
	Prediction	0
Logistic Regression on Kernel PCA results	Training	8
	Prediction	1

Accuracy of Digits dataset (%)	
Regular Logistic Regression	95.74 %
Logistic Regression on PCA results	52.41 %
Logistic Regression on LDA results	52.41 %
Logistic Regression on Kernel PCA results	23.19 %

Analysis:

It seems like Kernel PCA is always giving us bad results. It's worth mentioning that we are always getting worse accuracy because we are using a naïve dimensionality reduction without first analyzing the dataset and tuning the parameters.

Moreover, for the tested datasets, LDA and PCA are giving us same accuracy, but LDA has less running time.