CS 519: Applied Machine Learning I

HW4: Dimensionality reduction techniques

Submitted By: Md Ishtiaq Ahmed

Aggie ID: 800606216

Report

Dataset Description:

Two datasets are used in the assignment:

1. Iris dataset: The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. There is a total of 4 features. This dataset is available in sklearn library.
2. MNIST Dataset: The MNIST (Modified National Institute of Standards and Technology) dataset consists of 70,000 instances representing images of handwritten digits from 0 to 9. Each image has a size of 28x28 pixels. The dataset has 784 features per image. The dataset is imported using fetch_openml function where the dataset name is "mnist_784"

Performance analysis of the different dimensionality reduction techniques:

I have used the full IRISH dataset to analyze the performance of different dimensionality reduction techniques, but for MNIST dataset a subset is used. As MNIST has a total of 70000 instances, it is memory-consuming to do PCA in our local machines. I used train_test_split method to get 3% of total instances, which is 2100 instances, to analyze the MNIST dataset. Later the dimension-reduced data fed into the decision tree classifier to check the classification result. Different performance metrics are used to check the classification results including test accuracy, precision, recall and F1. Fitting time is also measured.

Three decision tree classifiers are used with the below hyperparameter to check the classification performance:

tree_model1 = DecisionTreeClassifier(criterion='gini', max_depth=4, random_state=1)

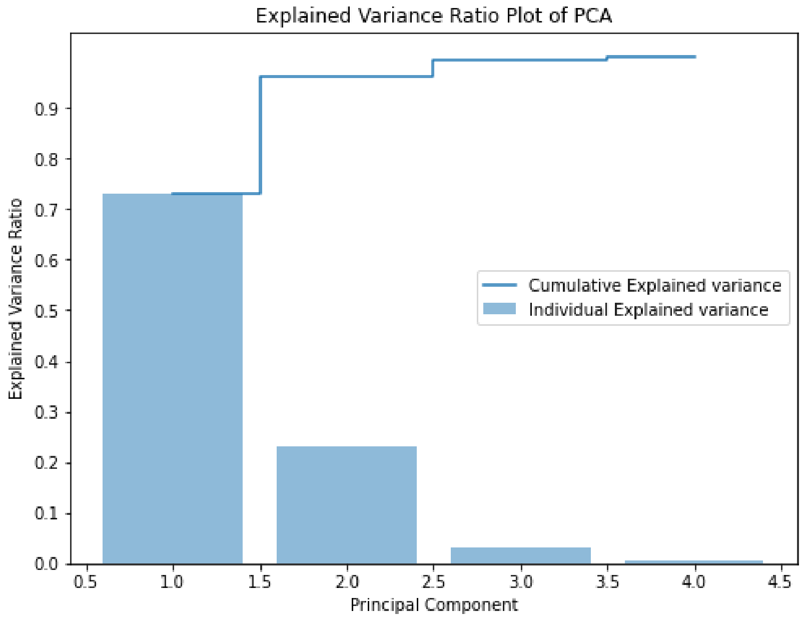
tree_model2 = DecisionTreeClassifier(criterion='gini', max_depth=40, random_state=1)

tree_model3 = DecisionTreeClassifier(criterion='gini', max_depth=400, random_state=1)

Primary Component Analysis:

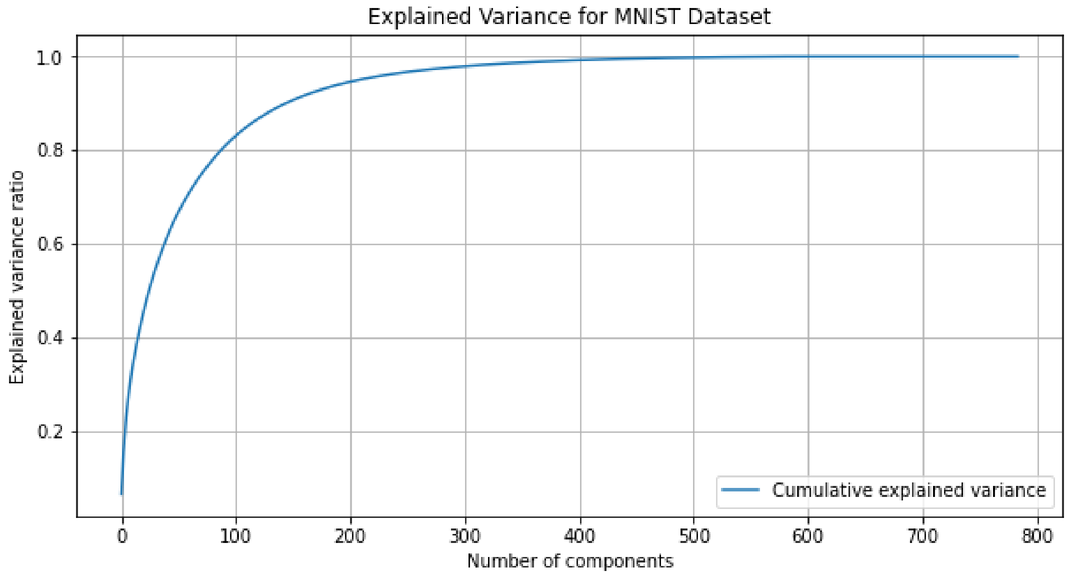
To implement dimensionality reduction techniques, a number of components needs to be selected for PCA, LDA and KPCA. We can get an idea about optimal number of components from the explained variance ratio plot.

Below is the explained variance ratio plot of IRISH dataset using PCA:



It is clear from the plot that selecting 2 primary components will give around 90% variance. So, n_components parameter selected as 2 for all the dimension reduction techniques.

Below is the explained variance ratio plot of MNIST dataset using PCA:



MNIST dataset has 784 feature in total. From above plot we can see that we can select n_components=100 to achieve 80% variance. However, for LDA n_components need to be less that 10 as total class label of this dataset is 10. Hence, for PCA and KPCA n_components =100 and for LDA n_components =8 is selected

Accuracy of Irish Dataset

dim_redu technique	Train Accuracy			Test Accuracy		
	tree_model 1	tree_model 2	tree_model 3	tree_model 1	tree_model 2	tree_model 3
PCA	0.971	1	1	0.822	0.822	0.822
LDA	1	1	1	0.956	0.956	0.956
KPCA (gamma=15, kernel =rbf)	0.79	0.99	0.99	0.422	0.4	0.4
KPCA(gamma=5, kernel =rbf)	0.895	1	1	0.844	0.778	0.778

LDA appears to be the most effective dimensionality reduction technique for this problem, followed by PCA. KPCA with a gamma value of 15 performs poorly on this dataset. It's also visible that different decision tree model has little impact on the accuracy except for KPCA.

Accuracy of MNIST Dataset

dim_redu technique	Train Accuracy			Test Accuracy		
	tree_model 1	tree_model 2	tree_model 3	tree_model 1	tree_model 2	tree_model 3
PCA	0.461	1	1	0.408	0.648	0.648
LDA	0.701	1	1	0.513	0.611	0.611
KPCA(gamma=0.01, kernel =rbf)	0.417	0.972	0.972	0.319	0.538	0.538
KPCA(gamma=0.01, kernel =poly	0.541	1	1	0.484	0.63	0.63

Test accuracies are relatively low across all dimensionality reduction techniques and tree models, indicating that these combinations might not be the best choice for the given problem while training accuracy achieved a perfect score for some cases. This suggests that all the models might be overfitting. Among all the combinations, KPCA with RBF kernel is performing worst.

Running Time (Time is in ms)

dim_redu technique	IRISH Dataset KPCA_1 (gamma=15, kernel =rbf) KPCA_1 (gamma=5, kernel =rbf)			MNIST Dataset KPCA_1 (gamma=0.01, kernel =rbf) KPCA_1 (gamma=0.01, kernel =poly)		
	tree_model 1	tree_model 2	tree_model 3	tree_model 1	tree_model 2	tree_model 3
PCA	1.998	2	2.998	128.925	209.88	179.897
LDA	1.999	2.997	4.004	9.993	15.99	17.989
KPCA_1	2	3.998	3	86.946	217.881	180.896
KPCA_2	1.998	1.997	2.996	107.931	174.902	180.894

The table shows that the choice of dimensionality reduction technique and decision tree model (choice of hyperparameters) can impact the fitting time. For the IRISH data set difference is not that visible, which might be due to a low number of instances. But when the number of instances increased, like in the MNIST dataset, running time varies widely. For the MNIST Dataset, LDA yields the fastest fitting times across all three decision tree models. The KPCA techniques have varying fitting times and the performance of KPCA is highly dependent on the kernel and parameter choices.

Precision

dim_redu technique	IRISH Dataset KPCA_1 (gamma=15, kernel =rbf) KPCA_1 (gamma=5, kernel =rbf)			MNIST Dataset KPCA_1 (gamma=0.01, kernel =rbf) KPCA_1 (gamma=0.01, kernel =poly)		
	tree_model 1	tree_model 2	tree_model 3	tree_model 1	tree_model 2	tree_model 3
PCA	0.825	0.822	0.822	0.464	0.65	0.65
LDA	0.961	0.961	0.961	0.544	0.623	0.623
KPCA_1	0.628	0.569	0.569	0.387	0.603	0.603
KPCA_2	0.846	0.776	0.778	0.612	0.636	0.636

For the IRISH Dataset, LDA consistently yields the highest precision across all three decision tree models (0.961). For the MNIST Dataset, the precision scores across all dimensionality reduction techniques and decision tree models are generally lower compared to the IRISH Dataset. This could be due to the increased complexity and size of the MNIST Dataset. Tree_model 1 gives the lowest precision for all dimension reduction technique in MNIST dataset. PCA achieves the highest precision scores for tree_model 2 and tree_model 3 (0.65)

Recall

dim_redu technique	IRISH Dataset KPCA_1 (gamma=15, kernel =rbf) KPCA_1 (gamma=5, kernel =rbf)			MNIST Dataset KPCA_1 (gamma=0.01, kernel =rbf) KPCA_1 (gamma=0.01, kernel =poly)		
	tree_model 1	tree_model 2	tree_model 3	tree_model 1	tree_model 2	tree_model 3
PCA	0.822	0.822	0.822	0.399	0.642	0.642
LDA	0.956	0.956	0.956	0.506	0.608	0.608
KPCA_1	0.422	0.4	0.4	0.308	0.532	0.532
KPCA_2	0.844	0.778	0.778	0.478	0.628	0.628

For IRISH dataset, like the precision score, recall is also highest for all decision tree models when LDA is used. In the MNIST Dataset, PCA achieves the highest recall scores for tree_model 2 and tree_model 3 (0.642) which is also like the precision score pattern.

F1

dim_redu technique	IRISH Dataset KPCA_1 (gamma=15, kernel =rbf) KPCA_1 (gamma=5, kernel =rbf)			MNIST Dataset KPCA_1 (gamma=0.01, kernel =rbf) KPCA_1 (gamma=0.01, kernel =poly)		
	tree_model 1	tree_model 2	tree_model 3	tree_model 1	tree_model 2	tree_model 3
PCA	0.821	0.822	0.822	0.353	0.644	0.644
LDA	0.955	0.955	0.955	0.48	0.609	0.609
KPCA_1	0.334	0.297	0.297	0.281	0.533	0.533
KPCA_2	0.841	0.77	0.77	0.485	0.628	0.628

F1 scores are also following a similar pattern as precision and recall scores.