Investigating Principal Component Analysis With Supervised Machine Learning On Classification

Mingxuan Zhao
University of California, San Diego
mzhao@ucsd.edu

Jiachen Ma
University of California, San Diego
jma@ucsd.edu

1 Introduction

With the increasing needs for classifying and labeling work for large amount of incidents, such as spam filtering and handwriting recognition, labor forces are not always precise and inefficient, or even unable to do such work. This leads to a rapid progress in various classification task in either supervised or unsupervised machine learning. Machine learning on classification has been conducted several decades ago; and now, as deep learning raises up and outperforms the traditional machine learning models, classification task of complex data like high-dimensional images becomes more and more popular. And there are many benchmarks as deep learning develops, for example, CIFAR 10 and ImageNet.
As nowadays more and more concise and complicated datasets emerge, each data sample may have thousands of features (think about high-resolution images). Machine learning models are suffered the curse of dimensionality as the number of features raise up. Therefore, Principal Component Analysis has been widely used in dimensionality reduction. It is an orthogonal linear transformation that projects data to a new coordinate so that the greatest variance by some scalar projection of the data comes to be the first principal components carrying the most amount of information.
In this report, we propose the investigation in principal component analysis with supervised machine learning task classification. Moreover, we introduce deep learning on classification to explore more on the effect of principal component analysis. We also explore and analyze the generalization of PCA on different datasets with large difference in dimensionality.

2 Related Work

Logistic Regression Classic statistical machine learning model for predicting categorical targets by modeling the probability of each target. It generates a prediction of probability between 0 to 1 by applying the logistic function (sigmod function). Besides only predicting binary variables, logistic regression can also predict for multi-class variables.

Multi-layer Perceptron Neural Network Multi-layer Perceptron Neural Network is a simple deep neural network by stacking multiple linear layer with non-linear activation function. For multi-class classification, we need to apply a cross entropy at the output of model to weigh each target's probability (all of them sum up to 1). Here the cross entropy loss function is just the combination of softmax and negative log-liklihood loss function.

Convolutional Neural Network Convolutional neural network is a network with certain regularization on the Multi-layer Perceptron Neural Network. As the layers in MLP are all fully-connected, assume the dimension of input is large enough, the total number of parameters in the MLP model will be extremely large. As a result, on one hand, the overfitting problem is easily occurred; on the other hand, it's computationally expensive as of too many parameters to fit on. For CNN, it has convolutional layers with k by k convoluted kernels for each channel. Instead of capturing all

Preprint. Under review.

	Without PCA	With PCA
Logistic Regression	0.97	0.97
MLP	0.97	1.00

Table 1: Test Accuracy for Models with or without PCA on Iris Dataset

parameters for each neurons, it downsamples the features, pools out the most important information, and saves many parameters and computing time. Nowadays, CNN is a common and powerful deep learning model, especially for image tasks as convolutional layers can capture spatial information from images. Started from 2012, along with the improvement of GPU and TPU, the introduction of AlexNet came with too great performance compared to common traditional machine learning model and defined the power of deep learning [1]. And then in 2015, ResNet dominated the image classification [2].

3 Methods

In our experiments, in order to investigate the power of PCA along with machine learning and deep learning models on classification, we propose the following experiment on two datasets: (1) Iris Flower Dataset and (2) MNIST Handwritten Digit Dataset. These two datasets have large difference in the number of samples (Iris has 150 samples and MNIST has 70,000) and number of features (Iris has 4 features and 3 classes, MNIST has 28x28 features and 10 classes). By conducting experiment on both datasets, we can analyze the generalization of PCA for different datas.
For data processing and preparation of each dataset, we randomly split the data into training set, validation set and test set by ratio of $70 : 10 : 20$. And we standardize data by zero-mean and unit standard deviation, as it's necessary for PCA not being biased due to some large variance by large numerical values. After data normalization, we apply PCA on each dataset. Specifically, for Iris Dataset, we select the first 3 principal components; for MNIST Dataset, we select the first 324 (18x18) principal components.
In experiment, for Iris Dataset, we compare logistic regression with and without PCA, and MLP neural network with and without PCA; for MNIST Dataset, we compare logistic regression with and without PCA, MLP neural network with and without PCA, and CNN with and without PCA. Besides comparison of PCA, we also cross-compare the performance of traditional statistical logsitic regression with deep neural network. In model construction of each dataset, for Iris Dataset, there is no fancy for logistic regression; for MLP neural network, we build a four layers neural network, with an input layer, 2 hidden layers of size 32 and an output layer of size 3 for 3 classes. We use the cross entropy loss function to rescale the model output as the probability of each class. In training, we train models for 100 epochs with a learning rate 0.01, a step learning rate decay 0.001, and early stopping. For MNIST Dataset, we also have logistic regression; for MLP neural network, we build a four layers neural network, with an input layer, 2 hidden layers of size 1024 and an output layer of size 10 for 10 classes. In training, we use the same hyperparameters as the MLP in Iris Dataset; for CNN, the structure follows 2 Conv-MaxPooling-ReLU-Dropout blocks and then connects to 2 fully-connected layers. We use the cross entropy loss function for the same reason as above. In training, we train models for 100 epochs with the same hyperparameters mentioned above.

4 Results

From Table 1, we can observe that logistic regression models with or without PCA all have the same performance. MLP neural network with PCA outperforms all others with a perfect test accuracy, and MLP without PCA has the same test accuracy as logistic regression models. Regard to the fact that there are just 30 test samples, a test accuracy of 0.97 simply means predict 29 out of 30 correctly. Though we cannot really conclude the importance and power of PCA from this table, as the test set is relatively small to reflect true performance, we can still get some senses that MLP deep learning model can achieve a better test accuracy with PCA.

From Table 2, we can observe some interesting and clear results, given the fact that there are 70,000 data samples in total and 14,000 of them for testing. The logistic regression model without PCA has the worst test accuracy among all, that's because 784 (28x28) dimensions of feature are too many

	Without PCA	With PCA
Logistic Regression	0.913	0.922
MLP	0.974	0.977
CNN	0.989	0.962

Table 2: Test Accuracy for Models with or without PCA on MNIST Dataset

for a simple logistic regression and cannot learn the relationship between these features for only one single linear layer. And logistic regression model with PCA performs better than the one without PCA but cannot beat any deep learning models. This strongly proves the point that PCA will improve the fitting for simple linear model and traditional machine learning model cannot reach to the same level of performance as deep learning models. For MLP, the one with PCA performs better than the one without, and that reclaims the power of PCA for this image dataset. Interestingly, for CNN, the CNN without PCA has the best test accuracy and the CNN with PCA performs even worse than MLP without PCA. There is no doubt that CNN is a better model for image tasks. However, for CNN with PCA, I guess that the bad performance comes from the fact that MNIST is an easy dataset and CNN with full 28x28 image information can already learn well. Therefore, with only 324 principal components, the lack of the rest information leads to the failure that CNN cannot learn very well from 18x18 images.

5 Discussion and Future Work

From all methods and experiments above, we can include that PCA will improve the performance of either traditional or deep learning models in supervised machine learning task classification. From the experiments on Iris Dataset and MNIST Dataset, we also observe the power of deep learning models with respect to the traditional machine learning models.
In future work, we think it's meaningful to investigate the bad performance of CNN with PCA. We would start with the previous work and initial guess, and proceed to understand the reason.

6 Reference

[1] ImageNet classification with deep convolutional neural networks. Alex Krizhevsky, NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1December 2012 Pages 1097–1105

@articleDBLP:journals/corr/HeZRS15, author = Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun, title = Deep Residual Learning for Image Recognition, journal = CoRR, volume = abs/1512.03385, year = 2015, url = http://arxiv.org/abs/1512.03385, eprinttype = arXiv, eprint = 1512.03385, timestamp = Wed, 17 Apr 2019 17:23:45 +0200, biburl = https://dblp.org/rec/journals/corr/HeZRS15.bib, bibsource = dblp computer science bibliography, https://dblp.org