Evaluating PCA

Summary

When it comes to evaluating models, there are different factors that come into play when deciding which one is best. Ideally everyone wants the best model that has the greatest accuracy in predicting future events, but this not always the case. Is an increase in accuracy by a few percent worth if the costs are much greater? Or is that extra 2 percent worth the delay in time? For this reason model evaluation sometimes needs to be analyzed on a case by case basis.

A method usually used to speed up the training process of a model is dimensionality reduction. For machine learning problems that use thousands or millions of features, the training and finding a solution can be extremely time consuming. The notion is in many real world problems the training instances are not as far spread out as the dimensions. For this reason it is possible to project or unfold many of the data points onto a lower dimensional space. A reduction in dimensionality to 3 dimensions or less is useful for us to visualize and better understand some of the relations of clustering occurring. In this study we will compare two models, with and without dimensionality reduction.

Research Design

We will be doing a multi class classification on the MNIST dataset used to predict handwritten numbers.

A random forest classifier will be used on all the explanatory variables and will be evaluated using the F1 score, which gives a combined understanding on the precision and recall. We will then perform a similar evaluation on a model that has gone through dimensionality reduction. One of the most common algorithms used to reduce dimensions is the Principal Component Analysis (PCA) which is a form of projection. In both cases we will be varying the dimensionalities but eventually settle on one that preserves 95% of the variance. We will also be recording the time taken for the process in both models.

Data Analysis

In the MNIST dataset there is a total of 70000 observations and 784 explanatory variables. For both models we will run the training set on 60000 observations and leave 10000 as a holdout set to evaluate on.

Model Evaluation

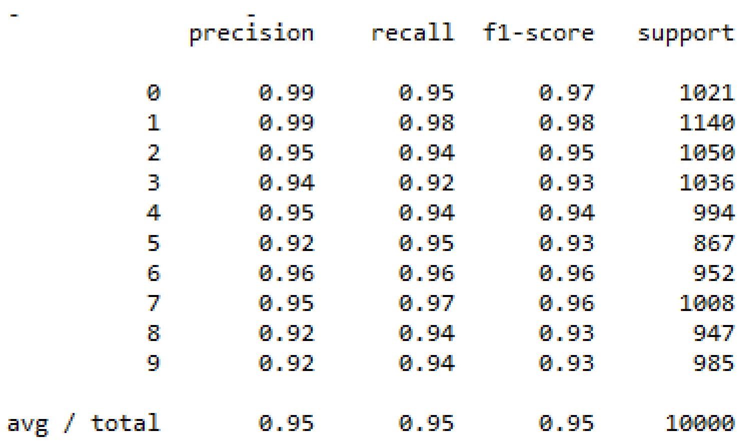
In the first model without PCA we got an average F1 score of 0.95 as seen in the classification report in Appendix A, Figure 1. The predicted vs true instance is further elaborated in the confusion matrix in Appendix A, Figure 2. The time taken to fit and evaluate this model was 4.576 seconds. Overall this seems to be a well performing model.

In order to reduce the number of dimensions we used Scikit Learns PCA function, which has an added benefit as it takes care of centering the data automatically. We started by trying to reduce the number dimensions to 2 as it would be easier to visualize the data. The time taken to run the model was less than the original analysis however the performance was very poor with an F1 score of 0.42, Appendix A Figure 3, 4. Using 12 dimensions we got a much better F1 score of 0.95 however the time taken was much greater at 12.58 seconds, Appendix A, Figure 5. Finally instead of stating the number of dimensions we tried using the minimum number of dimensions required to preserve 95% of the training sets variance. This produced a model with 154 dimensions, the F1 score was favorable at 0.93 however the time taken for the process was not, 30.53 seconds, Appendix A, Figure 6.

Recommendation

As mentioned earlier dimensionality reduction can vary on a case by case basis. Although in most cases it would make sense to reduce the number of dimensions to speed up training, in this case the costs of model development and implementation of PCA did not outweigh the results obtained in the original study without any dimensionality reduction.

Appendix A

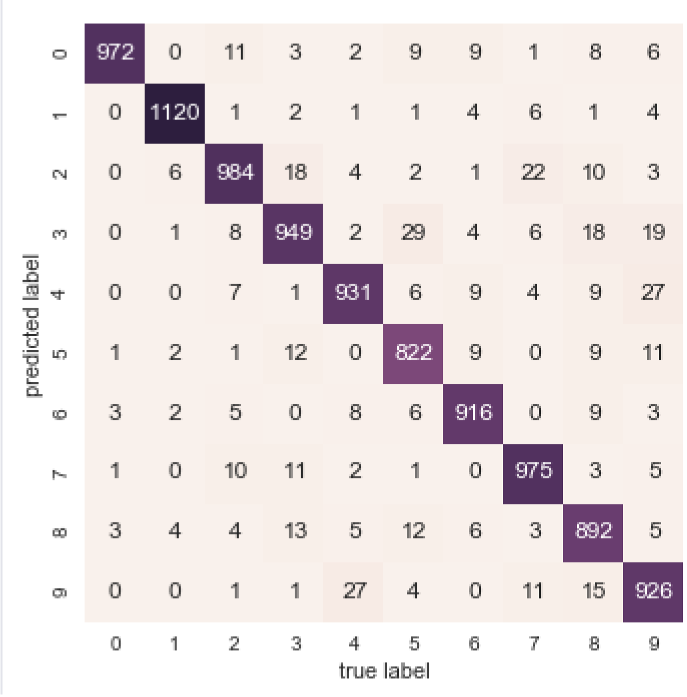


```
             precision    recall  f1-score   support

          0       0.99      0.95      0.97      1021
          1       0.99      0.98      0.98      1140
          2       0.95      0.94      0.95      1050
          3       0.94      0.92      0.93      1036
          4       0.95      0.94      0.94       994
          5       0.92      0.95      0.93       867
          6       0.96      0.96      0.96       952
          7       0.95      0.97      0.96      1008
          8       0.92      0.94      0.93       947
          9       0.92      0.94      0.93       985

avg / total       0.95      0.95      0.95     10000
```

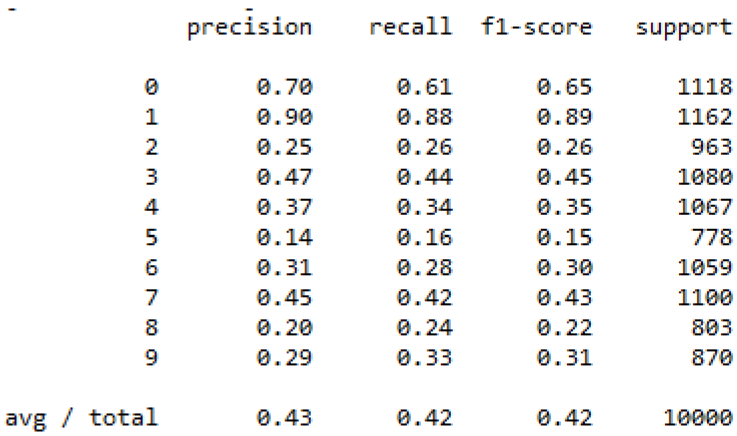
Figure 1 - Classification Report



predicted label \ true label	0	1	2	3	4	5	6	7	8	9
0	972	0	11	3	2	9	9	1	8	6
1	0	1120	1	2	1	1	4	6	1	4
2	0	6	984	18	4	2	1	22	10	3
3	0	1	8	949	2	29	4	6	18	19
4	0	0	7	1	931	6	9	4	9	27
5	1	2	1	12	0	822	9	0	9	11
6	3	2	5	0	8	6	916	0	9	3
7	1	0	10	11	2	1	0	975	3	5
8	3	4	4	13	5	12	6	3	892	5
9	0	0	1	1	27	4	0	11	15	926

Figure 2 - Confusion Matrix

Evaluating PCA

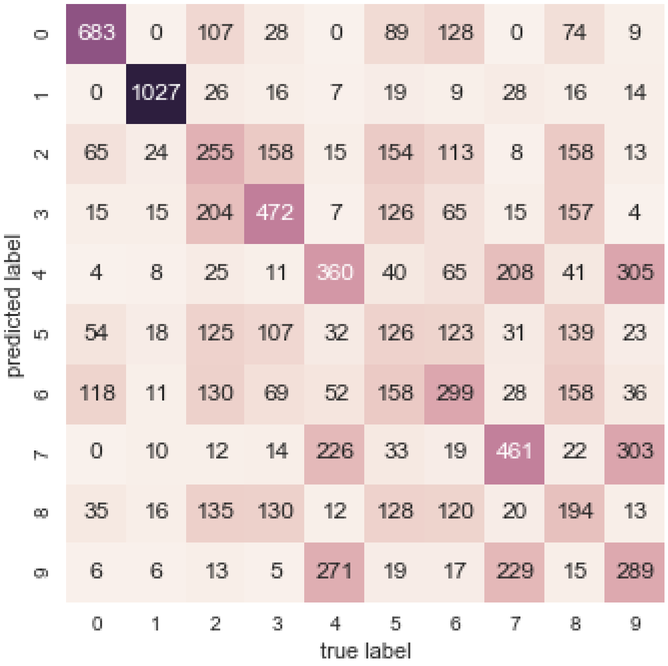


```
             precision    recall  f1-score   support

          0       0.70      0.61      0.65      1118
          1       0.90      0.88      0.89      1162
          2       0.25      0.26      0.26       963
          3       0.47      0.44      0.45      1080
          4       0.37      0.34      0.35      1067
          5       0.14      0.16      0.15       778
          6       0.31      0.28      0.30      1059
          7       0.45      0.42      0.43      1100
          8       0.20      0.24      0.22       803
          9       0.29      0.33      0.31       870

avg / total       0.43      0.42      0.42     10000
```

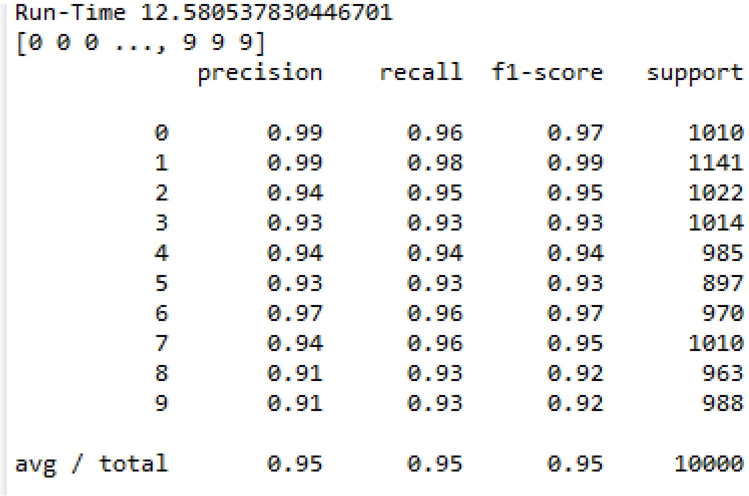
Figure 3 - Classification report with PCA, Number of Dimensions = 2



	0	1	2	3	4	5	6	7	8	9
0	683	0	107	28	0	89	128	0	74	9
1	0	1027	26	16	7	19	9	28	16	14
2	65	24	255	158	15	154	113	8	158	13
3	15	15	204	472	7	126	65	15	157	4
4	4	8	25	11	360	40	65	208	41	305
5	54	18	125	107	32	126	123	31	139	23
6	118	11	130	69	52	158	299	28	158	36
7	0	10	12	14	226	33	19	461	22	303
8	35	16	135	130	12	128	120	20	194	13
9	6	6	13	5	271	19	17	229	15	289

Figure 4 – Confusion Matrix with PCA, Number of Dimensions = 2

Evaluating PCA

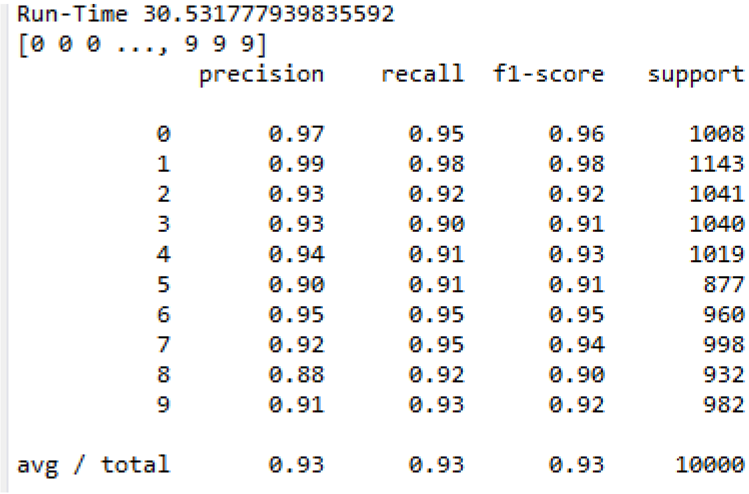


```
Run-Time 12.580537830446701
[0 0 0 ..., 9 9 9]
             precision    recall  f1-score   support

          0       0.99      0.96      0.97      1010
          1       0.99      0.98      0.99      1141
          2       0.94      0.95      0.95      1022
          3       0.93      0.93      0.93      1014
          4       0.94      0.94      0.94       985
          5       0.93      0.93      0.93       897
          6       0.97      0.96      0.97       970
          7       0.94      0.96      0.95      1010
          8       0.91      0.93      0.92       963
          9       0.91      0.93      0.92       988

avg / total       0.95      0.95      0.95     10000
```

Figure 5 - Classification report with PCA, Number of Dimensions = 30



```
Run-Time 30.531777939835592
[0 0 0 ..., 9 9 9]
             precision    recall  f1-score   support

          0       0.97      0.95      0.96      1008
          1       0.99      0.98      0.98      1143
          2       0.93      0.92      0.92      1041
          3       0.93      0.90      0.91      1040
          4       0.94      0.91      0.93      1019
          5       0.90      0.91      0.91       877
          6       0.95      0.95      0.95       960
          7       0.92      0.95      0.94       998
          8       0.88      0.92      0.90       932
          9       0.91      0.93      0.92       982

avg / total       0.93      0.93      0.93     10000
```

Figure 6 - Classification report with PCA, Number of Components = 0.95