ML and PR Project Report – Part 2

Like the other reports, this one pertains to the analysis of the fingerprint database.

This report contains a brief analysis of the distribution of the data processed through PCA and LDA separately, of the performance of classification using thresholds along the LDA direction, with and without PCA preprocessing. The dataset has 6 features and 2 classes, for a total sample size of 6000 samples.

**Analyzing PCA transformation**

We apply PCA without reducing dimensionality, thus keeping 6 directions.

The resulting histograms clearly show that direction 0, the one that maximizes variance, is also the only one that allows for some kind of classification, since for the other ones, samples of different classes overlap almost entirely. This can also be observed by plotting the scatter plot feature by feature: samples of different color (class) overlap for features 1 through 5, while we can identify two just slightly overlapping clusters for feature 0.

That being said, direction 0 seems to be able to offer a great separation of the classes.

Immagine che contiene testo, diagramma, schermata, Diagramma

Descrizione generata automaticamente Immagine che contiene schermata, testo, diagramma, Diagramma

Descrizione generata automaticamente

Immagine che contiene testo, schermata, Policromia

Descrizione generata automaticamenteImmagine che contiene testo, schermata

Descrizione generata automaticamente

(feature and direction are used interchangeably throughout the report)

Plots for features 2 through 5 are omitted as they have pretty much the same behaviour as the ones of feature 1.

**Analyzing LDA transformation**

We apply LDA with the only dimensionality option available, which is reducing to one dimension, as the number of classes is 2.

Through the analysis of the histogram and the scatter plot, where data is distributed along the only LDA axis, the separation of samples of different classes, i.e. the separation of the different classes, seems to be quite good, as the samples only overlap a little, in the center.

Immagine che contiene testo, diagramma, schermata, Diagramma

Descrizione generata automaticamenteImmagine che contiene testo, schermata, Policromia

Descrizione generata automaticamente

If we compare the results with those obtained when applying PCA we can see that the distributions don’t seems to differ much. LDA and PCA directions are probably close to one another.

Immagine che contiene testo, diagramma, schermata, Diagramma

Descrizione generata automaticamenteImmagine che contiene testo, diagramma, schermata, Diagramma

Descrizione generata automaticamente

(PCA) (LDA)

**Using LDA as classifier**

We classify using LDA first by simply computing the threshold as the mean of the class means, i.e. , and then by exploring the performance of other thresholds near the one obtained naïvely. I will anticipate that the minimum threshold(s) don’t optimize the classification that much: they never break through the 0.3% improvement in classification error.

First of all we split the data with a 2:1 ratio between *training* and *validation*, thus applying LDA on the *training* part, obtaining the threshold as explained above, and finally using that value to classify of the *validation* set. We then compare the predicted labels with the actual ones, and find the error rate. The results are as follows:

“186 misses (over 2000 samples) detected using LDA, which is a 9.3% error rate.”

I tried to explore the space near the naïve threshold finding a possibly optimal threshold, which brings down the errors to:

“181 misses (over 2000 samples) detected using LDA, which is a 9.049% error rate.”

The explored space looks like this:

Immagine che contiene testo, diagramma, schermata, Carattere

Descrizione generata automaticamente

**Using LDA with PCA preprocessing as classifier**

We finally try the same approach as in the previous paragraph, this time adding an extra step, which is the application of PCA, variating the number of dimensions of the projection.

The results don’t differ much from the previous approach. Since PCA basically projects the data in a distribution that is hardly different from the one obtained using just LDA, this result seems to suggest that LDA can’t improve the results much more, meaning that PCA was already really good at separating classes.

The errors are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Naive Threshold | | Optimized Threshold | |
| m | Error Rate | m | Error Rate |
| 2 | 9.25% | 2 | 8.95% |
| 3 | 9.5% | 3 | 9.15% |
| 4 | 9.15% | 4 | 9.1% |
| 5 | 9.35% | 5 | 9.049% |
| 6 | 9.25% | 6 | 9.049% |

Immagine che contiene testo, linea, diagramma, Diagramma

Descrizione generata automaticamente

As we can see, for the naïve approach, 3 times out of 5 we get better results, but just slightly, as we never exceed a 0.2% improvement in error rate ***(at least, as far as my limited knowledge extends, I believe that a <0.2% improvement isn’t really much)***. For the optimized approach we beat the LDA only once, and drew twice: the improvement was about 0.1%.

**Conclusion**

LDA allows for a good classification (error rate < 10%), but combining it with PCA doesn’t yield astonishingly better results (for some values of m they’re even worse)