Machine Learning and Pattern Recognition: final report

Claudio Macaluso, s317149

November 11, 2024

1 Preliminary hypothesis: a visual inspection of data

!! add something about unimodal/multimodal !!
!! add something about correlation?? (2-3 seems correlated) !!

All the analyses conducted in this section consist of graphical evaluations made by observing the scatter plots and histograms of the data (Figure 1), therefore, they are not measurable and quantitative assessments, but only qualitative results extracted 'by eye'.

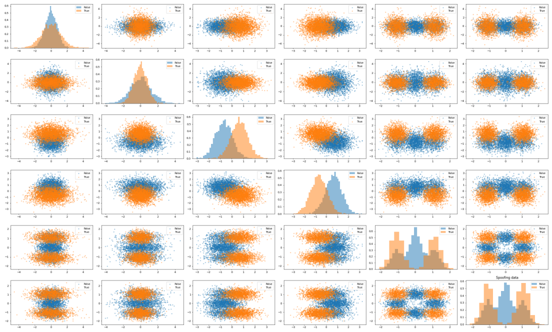

Figure 1: Spoofing dataset scatter matrix

About the first two features, they are first of all unimodal: this can be useful if we use some model that assume a distribution that is inherintly unimodal (e.g. Gaussian models). Unfortunately, as we can see they show a significant overlap, and for these, their discriminative power can be less than other features. Despite this, these two features combined with the other ones can anyway contribute positively to the classification task. Another thing that have to be taken into account, is that the variances are different for the two classes and this can lead to a decrease in performance if we consider some model that assume the same variance for all the classes (e.g. Tied Covariance Gaussian Model).

Moving on the last two, they seem less overlapped compared to the previous ones, but in this case they're multimodal: multimodality can be challenging, specifically if we use some model that assume an underlying unimodal distribution (e.g. Gaussian Models). So an extrimely simple model in the specific case of these two features could perform poorly. So, taking into account mean and variance of the classes might not be significative anymore, since mean doesn't represent any actual cluster of data points and, since variance represent spread around mean but mean isn't representative of any specific cluster points, we can't rely on the variance as a good parameter of spread of data. (review previous) In this case, looking at the scatter plots, we can imagine that a quadratic and more complex separation surface can lead to better results, since data cluster are multiple and seem to be not separable from a simpler, linear surface (e.g. Tied Covariance Gaussian Model, Logistic Regression and Support Vector Machine using standard kernel).

The third and the fourth features instead are unimodal and have a lower overlap, so they could probably perform well using models that assumes underlying unimodal distribution and be, on their own, the most discriminative features over all. Furthermore, since these features have similar class variance, it could be suitable to apply some models that make the assumption of same covariance among classes (e.g. Tied Covariance Gaussian Model). Applying

LDA moreover, we can notice that the found directions highly correspond to the directions of the third and the fourth features, since LDA tend to select linear combination of the features in a way that make classes more distinguishable in the projected data, confirming the hypothesis for which these two features lead to a more discriminative separation of classes.

```
D, L, label_dict = load("project/data/trainData.txt")
W, _ = lda(D, L, m=1)
print(W)

# Output:
#[[-0.01063821]
# [ 0.0134172 ]
# [-0.96566604]
# [ 0.96774246]
# [ 0.0217633 ]
# [-0.03289391]]
```

Last thing about the third and fourth features, as we can see from scatter plots, the two are correlated, so using models assuming non correlated features (e.g. Naive Bayes Gaussian Models) can lead to a decrease in performance for these two features, so we will need to take care about this in the further analysis.

2 Dimensionality reduction techniques

Now, let's evaluate qualitatively, as we've done in the previous section, the impact of dimensionality reduction in our dataset visually.

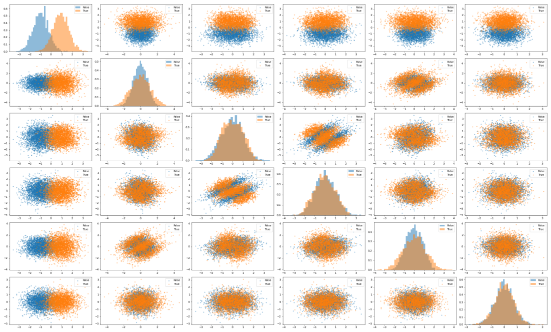

Figure 2: Spoofing dataset scatter matrix projected according to PCA directions

As we expected, since PCA is a non supervised dimensionality reduction technique (i.e. class agnostic), the found directions seem to not separate well, at least in some of the features, the classes to which the data belong. From a first analysis to the scatter plots of the PCA projected data, we can see that, except for the first component, the others seem to show high overlap between classes. Calculating the means and variances of classes in the projected space, we can see that the first component is the only one with a consistent difference in the mean of the classes. For the variances of the classes instead, we can notice that second and fifth features seem those with the bigger variance, so they still could be useful to distinguish between one or the other class. A possible benefit of PCA projection in this case could be that, from the histograms, we can notice that all the new features have unimodal distribution, so some models could improve their performance due to these new characteristics of the projected data. The last one point is a coincidence, since PCA objective isn't inherintly searching for a projection space in which features are unimodal, but still can be taken into account as an improvement in the new features space.

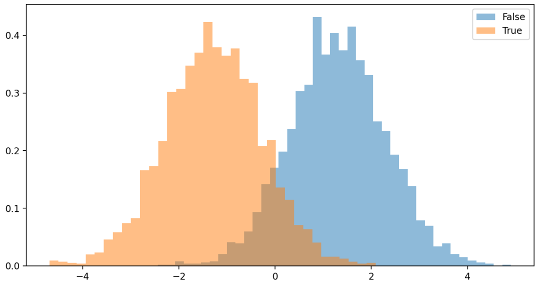

Figure 3: Spoofing dataset scatter matrix projected according to LDA directions

Clearly, a better separation can be achieved by applying LDA dimensionality reduction technique, that, unlike the previous one, is a supervised technique that aims to find the direction in which to project data for which the separation between classes is maximized.

Setting up a simple classification rule for 1-dimensional data that uses the mean of the means of the classes as a threshold to classify each sample, we can begin to check the impact of dimensionality reduction technique on classification (Table 1), and possibly use accuracy metrics for this simple classification as a baseline to compare to the next models performance.

Dimensions	**Error Rate**	**Explained Variance**
6	9.30%	100.00%
5	9.25%	90.68%
4	9.40%	74.29%
3	9.25%	57.69%
2	**9.05**%	40.94%
1	9.25%	23.99%

Table 1: Table showing the error rate using m Principal Components and classifying with a simple LDA, mean based classifier

3 Models evaluation

3.1 Gaussian Models

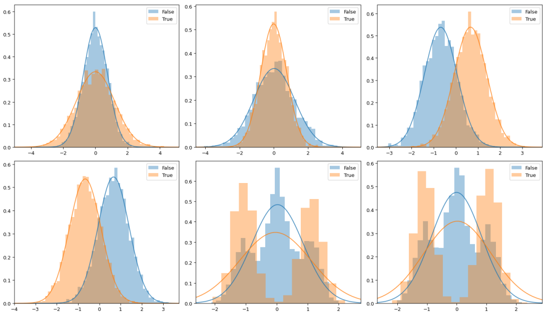

Figure 4: Spoofing dataset comparison of per class/feature histograms with Gaussian distribution using data mean and variance