MACHINE LEARNING AND PATTERN RECOGNITION REPORT

Fingerprint Spoofing Detection

Antonio Iorio
Who?
Where?
When?

Contents

Introduction The project task consists of a binary classification problem. The goal is to perform fingerprint spoofing detection, i.e. to identify genuine vs counterfeit fingerprint images. The dataset consists of labeled samples corresponding to the genuine (True, label 1) class and the fake (False, label 0) class. The samples are computed by a feature extractor that summarizes high-level characteristics of a fingerprint image. The data is 6-dimensional.

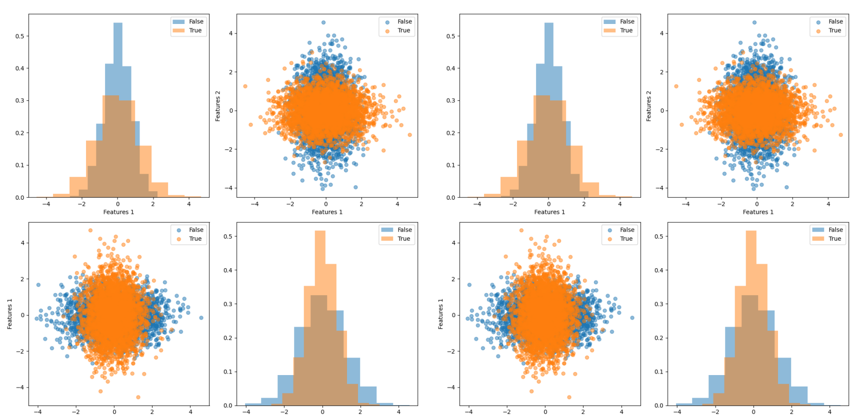
1 Dataset Analysis

In our analysis process, one first begins to represent what are the data related to the various features and how among the various features the data are distributed by making a visual representation in pairs of features.

1. Starting with the analysis of the first two features and creating a histogram and a scatter Figure 1, one can see:

   - Both features overlap
   - Follow a Gaussian distribution
   - Feature 1 has a peak at [-0.213, 0.276] and it is worth 0.541 for the false class, instead for Feature 2 the peak at [-0.402, 0.165] and it is worth 0.516 for the true class.

   Seeing Figure 1 again, it would appear that Figure 1a and Figure 1b appear to be visually the same but in b they are represented centred which as can be seen is quite similar because Feature 1 has $\mu = 0.00170711$ and $\sigma^2 = 1.00134304$, instead Feature 2 has $\mu = 0.00503903$ and $\sigma^2 = 0.9983527$



(a) Without mean (b) With mean

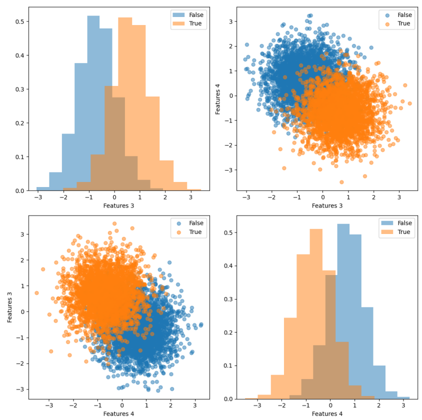
Figure 1: Feature 1 vs Feature 2 - Without centering the data relative to the average (a) and with centering the data relative to the average (b)

2. For features 3 and 4, observed in Figure 2, on the other hand, they have:

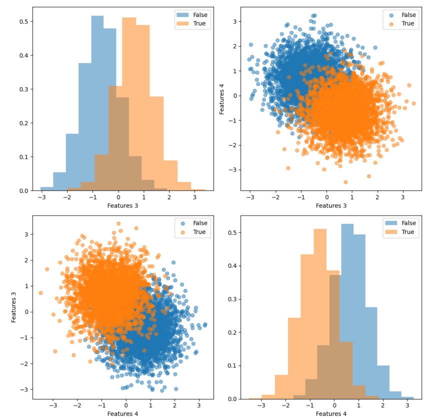
   - do not overlap like the previous two
   - Follow a Gaussian distribution but the true and false labels are centred at different points

- Feature 3 has a peak at [-1.063, -0.568] and it is worth 0.517 for the false class, instead for Feature 4 the peak at [0.290, 0.783] and it is worth 0.525 for the false class.

Seeing Figure 2a and Figure 2b, data are already similar because, the mean calculated with reference to the two classes is close to 0, in fact: Feature 3 has $\mu = -0.00560753$ and $\sigma^2 = 1.0024818$, instead Feature 4 has $\mu = 0.00109537$ and $\sigma^2 = 0.99029389$



(a) Without mean



(b) With mean

Figure 2: Feature 3 vs Feature 4 - Without centering the data relative to the average (a) and with centering the data relative to the average (b)

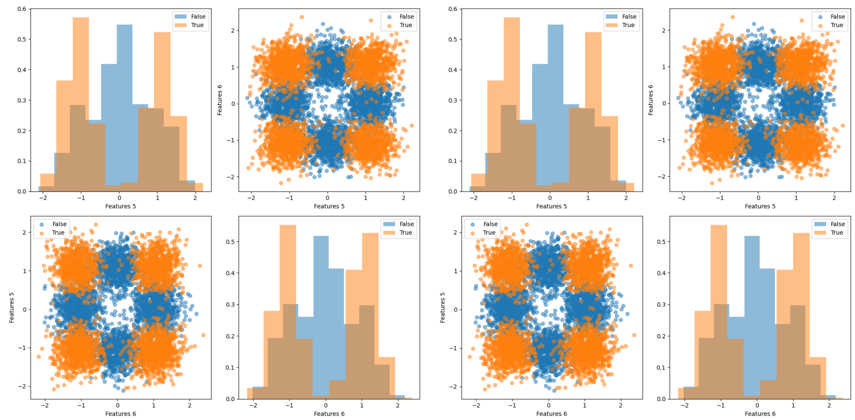
3. For features 5 and 6, observed in Figure 3, one can see:

   - Do not totally overlap
   - For both features, the true labels don't follow a Gaussian distribution as opposed to the false ones, which could be more approximate
   - Feature 5 has a peak at [-1.211, -0.783] and it is worth 0.572 for the true class, instead for Feature 6 has a peak at [-1.273, -0.817] and it is worth 0.553 for the true class.

   Also in this other case, Figure 3a and Figure 3b are similar because again the average is close to 0. Feature 5 has $\mu = -0.00700025$ and Feature 6 has $\mu = 0.00910515$

2 Dimensionality Reduction

Before proceeding with classification, two techniques of dimensionality reduction PCA and LDA can be analysed. The goal is to find a subspace of the feature space that preserves most of the useful information, that is, mapping from the *n–dimensional* feature space to *m–dimensional* space, with $m \ll n$



(a) Without mean (b) With mean

Figure 3: Feature 5 vs Feature 6 - Without centering the data relative to the average (a) and with centering the data relative to the average (b)

2.1 PCA

is an unsupervised technique. Where starting from a dataset $X = \{x_1, \ldots, x_k\}$ and calculated average. It starts with the empirical covariance matrix:

$$C = \frac{1}{K} \sum (x_i - \bar{x})(x_i - \bar{x})^T \tag{1}$$

We compute the eigen-decomposition of $C = U\Sigma U^T$ and project the data in the subspace spanned by the m columns of U corresponding to the m largest eigenvalues.

$$y_i = P^T (x_i - \bar{x}) \tag{2}$$

where P is the matrix of the m columns of U associated to the m highest eigenvalues of C. A cross-validation approach can be used to figure out the optimal value of m to be selected. To evaluate each eigenvalue, one would have to calculate the variance corresponding to the axis. The percentage can be calculated as the rate between the sum of the m eigenvalues and the sum of all of them. In Figure 4 we can see how it changes in the project. A good m, corresponds to that value which allows a percentage greater than 95%, so in our case we would need all 6 features.

2.2 LDA

is a supervised technique. To find a direction that has the best separation between classes, we measure spread between classes in terms of class covariance. The objective is to maximize the $between - class$ variability over $within - class$ variability ratio for the transformed samples:

$$\max_w \frac{w^T S_B w}{w^T S_W w} \tag{3}$$

where:

$$S_B \triangleq \frac{1}{N} \sum_{c=1}^{K} n_c (\mu_c - \mu)(\mu_c - \mu)^T \tag{4}$$

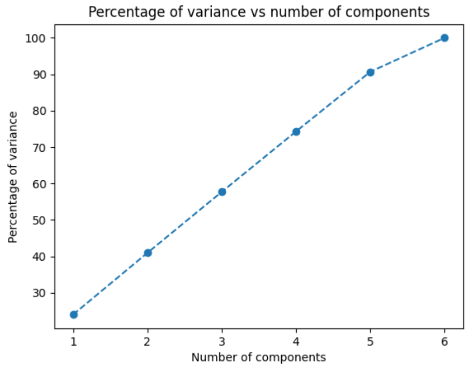


Figure 4: Cross validation for PCA impact evaluation

Method	Num Samples	Error	Error Rate
LDA	2000	186	9.3%

Table 1: Table showing the results of the LDA and PCA + LDA method.

$$S_W \triangleq \frac{1}{N}\sum_{c=1}^{K}\sum_{i=1}^{n_c}(x_{c,i}-\mu_c)(x_{c,i}-\mu_c)^T \tag{5}$$

μ is dataset mean μ_c is class mean

2.3 Our project

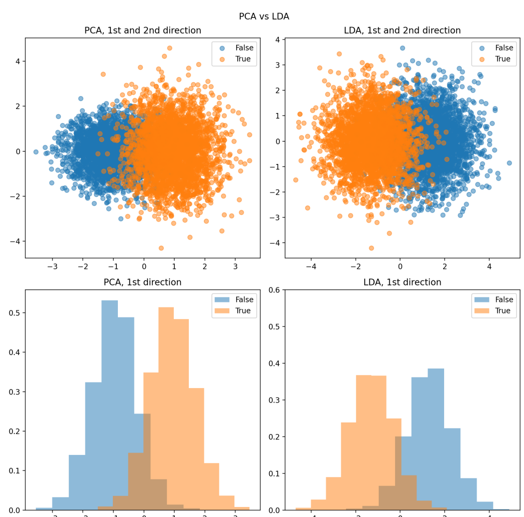


Figure 5: Comparing resulta between PCA and LDA

3 Classification Models Analysis

3.1 Gaussian Models

3.2 Logistic Regression Classifier

3.3 Support Vector Machine Classifier

3.4 Gaussian Mixture Models Classifier