Computational Linear Algebra for Large Scale Problems
Homework 3

An LDA Based Explainable Approach to Breast Cancer Identification

Francesco Capuano matr.295366
Alessandro Licciardi matr.296152
A.A. 2021-22

Abstract

The proposed work aims to introduce the Linear Discriminant Analysis and show its applications in dimensionality reduction. In this work we then show how the KNN classifier is able to reach top performance for breast cancer classifications in the space obtained through LDA. Moreover, we observe how the maximal performance is obtained maximizing the distance between classes rather than utilizing other dimensionality reduction such as PCA. LDA extensively uses concepts such as the SVD decomposition and the Moore-Penrose pseudo-inverse.

1 Introduction

More than two millions of women every year are diagnosed with breast cancer and, of those, more than half a million (six hundred thousands) every year dies from it ([1]).

In Italy, breast cancer alone is the the first cause of death in the female population.

Breast cancer likelihood is reported to be increased by several different risk factors, including genetic factors, conducing a sedentary lifestyle, smoking and being overweight [2].

On the other side, even if new cases are increasing year by year, breast cancer has a five year survival rate of approximately 85%, making it one of the cancers with the highest survival rate [1].

Such an high value of recovery is both related to following the appropriate cure protocols and to possibility, quite common in wealthy countries, of getting tested early and regularly.

In Italy, for example, tests are conducted on average every two years, and women are invited to getting checked for breast cancer once they reach the threshold of the 50 years of age directly by the Minister of Public Health.

Typically, screening exams consists of a preliminary mammography that, when resulting positive, yields to further examinations conducted by health-care providers in the sake of determining the real nature of the lesion that mammography (or breast ultrasound) had eventually detected.

However, these methods are certainly resources demanding. Mammography, in particular, according to the John Opkins University, is a procedure consisting in:

"[...] an X-ray examination of the breast. It is used to detect and diagnose breast disease in women who either have breast problems, such as a lump, pain, or nipple discharge, as well as for women who have no breast complaints. The procedure allows detection of breast cancers , benign tumors, and cysts before they can be detected by palpation (touch). Mammography cannot prove that an abnormal area is cancer, but if it raises a significant suspicion of cancer, tissue will be removed for a biopsy . Tissue may be removed by needle or open surgical biopsy and examined under a microscope to determine if it is cancer [...]"

From this description, we can derive that the process of breast cancer diagnosis requires highly specialized professionals, specialized (and costly) medical imaging machinery and, in the event of concerns-raising mammograms, the possibility of performing surgical procedures and Laboratory analysis on the samples of tissues.

Hence, while mammography is characterized by being mildly invasive, it is clear that this procedure overall is certainly prohibitive to those medium to low income countries in which such resources are not available.

This is also directly testified by data presented in [3]: in their work, the researchers have shown that, while in Europe, out of 450'000 European women diagnosed with breast cancer, approximately one third

(140'000) died from it, out of the 68'000 African women diagnosed with this condition it is possible to count 37'000 deaths: the 55 % percent of the overall ill population.

Researchers have shown that, among many different causes, a reason for such an high imbalance on the survival rate is the lack of early testing, motivated by the shortage of resources typical of certain areas of the African continent.

A far less invasive approach to study the mammary tissue is to use tissue electrical conductivity data, or, for short, *tissue electroscopy*.

This technique is basically based on acquiring data related to electrical properties (one among many, impedance) of the mammary tissue.

To perform such analysis the amount of necessary resources is far lower then the ones needed to do mammograms.

The main problem with using tissue electroscopy is that, to this day, there is a lack of a direct connection between the procedure outcome and the relative diagnose, considered the low interpretability of the readings.

Particularly, in contrast to ultrasounds and mammograms (which produce images), tissue electroscopy produces a series of scalar values whose direct interpretation might not be immediate.

This work revolves around applying Machine Learning techniques along with Linear Discriminant Analysis to develop a fully trained classification model which can distinguish among the different tissues, returning, the class for each input.

For this scope, we have used a public dataset, the *Breast Tissue Dataset* at the UCI Machine Learning repository, [4].

2 Data Exploration

The dataset we have used contains information related to the electrical scan of breast tissue.

It is a collection of 106 records in a nine-dimensional space. In particular, each data point is described with respect to:

- *I0*: Impedivity (ohm) at zero frequency
- *PA500*: Phase angle at 500 KHz
- *HFS*: High-frequency slope of phase angle
- *DA*: Impedance distance between spectral ends
- *AREA*: Area under spectrum
- *A/DA*: Area normalized by DA
- *MAX IP*: Maximum of the spectrum
- *DR*: Distance between I0 and real part of the maximum frequency point
- *P*: Length of the spectral curve

Moreover, each sample is assigned to a *Class*: carcinoma (car, for short), fibro-adenoma (fad), mas mastopathy (mas), glandular (gla), connective (con) and adipose (adi).

The objective of this project is to develop a model which can learn from the data the characteristics of those data points in each class, in order to correctly predict the class label for new unseen records. This can then be expressed as problem of *multi-class classification*, in which the number of possible classes is equal to six.

However, as presented in the general introduction to the dataset, we decided to merge together fibro-adenoma, mastopathy and glandular tissues, considering that those classes are reported to be not accurately discriminated anyway.

This reduced the problem size to a four classes classification. Those four classes can be shortly described as follows:

- **adipose tissue**: adipose tissue is present in the breast to store the excess energy and release it when required by the body. The presence of this kind of tissue is not directly associated with breast cancer, although excess of it might increase the risk of developing this condition.
- **connective tissue**: connective tissue is one of the main components of the breast. Its presence is motivated by the necessity of holding together the different components of the mammals such as the lobules and the ducts. Its presence is not typically associated with breast cancer, although typically the cancer cells' presence has its origin in the lobules and the ducts themselves.
- **fibro-adenoma, mastopathy and glandular issue**: *glandular tissue* is the type of tissue that ensures the typical functionalities to mammals, such as producing milk for the offspring. The presence of this kind of tissue is not associated with breast cancer, although when above a certain threshold it makes the read of an eventual mammogram harder because of the *dense-breast effect*. *mastopathy* and *fibro-adenoma* are two kind of benign tumor which are is not immediately related to a cancer diagnosis but which needs to be accurately monitored to avoid the unnoticed onset of cancer.
- **carcinoma**: carcinoma tissue is tissue directly associated with breast cancer. When this tissue is originating the onset of other cancer cells in other regions from the breast, the cancer to which it is associated is said to have metastasized.

It is then possible to group those four classes into two groups:

- normal condition group, containing *connective* and *adipose* tissue
- attention-worth group, containing *tumor tissue* and *cancer tissue*.

With respect to the quality of the data we observed how the features' values are in ranges which are completely different.

Moreover, we noticed the absence of missing values in the dataset and the presence of only one single duplicate, which we decided to remove from the dataset itself.

We then explored the distribution of the target variable inside the data. This distribution is presented in Figure [1].

It is clear that the target variable is fairly unbalanced distributed in the data. Since the quantity of the data is certainly not large, we have decided to not perform any down-sampling action to recover some sort of balance. Other than that, since we are interested in fairly valid results by the means of consistency, we also disregarded the option of up-sampling via data generation: we believe that in this way we would certainly have increased the quantity of available data, but we would also have introduced a bias too large to still consider the results reliable.

Anyway, the high imbalance of the target variable is certainly something that affects the overall process, in particular in terms of the choice of the right performance metrics.

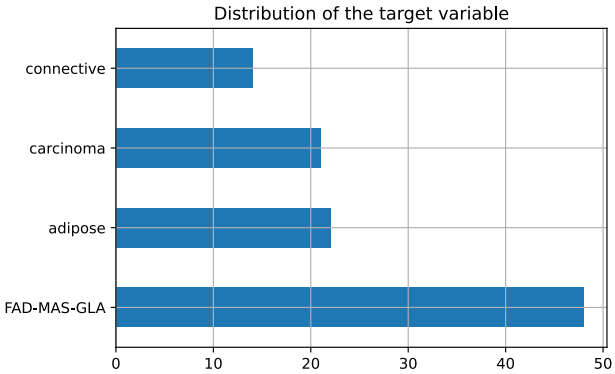

Figure 1: Distribution of breast tissues inside the dataset

3 Classification Models and Metrics

Let us briefly recall what a classification model is.

A classification model can be seen as a function g that depends on a set of given parameters, namely θ, that maps a vector $x \in \mathbb{R}^n$ into a finite set of labels $\{0, \dots, c-1\}$ that are often referred to as *classes*, i.e.

$$\begin{aligned} g(\cdot\,;\theta) : \mathbb{R}^n &\to \{0, \dots, c-1\} \\ x &\mapsto \hat{y} = g(x;\theta) \end{aligned}$$

Classification models aim to predict, as correctly as possible, the class corresponding to the input vector x.

Let us suppose that we are given a dataset $\tau = \{(x_i, y_i)\}_{i=1}^{p}$, where $x_i \in \mathbb{R}^n$ is a feature vector and $y_i \in \{0, \dots, c-1\}$ represents the associated class label. The parameter vector θ is learned from τ, which means that we can find an optimal estimation $\hat{\theta}$ through an optimization problem minimizing the loss function, i.e. a function that measures the distance between the exact label y_i and the predicted one $\hat{y_i}$.

Once one obtains the predicted labels obtained after the estimation of θ, it is important to evaluate the performance of the model with respect to some ground truth observations.

This is typically done via hold-out (i.e. reserving a portion of the original data and the corresponding class labels to perform evaluation) or, alternatively, via k-fold cross validation (i.e., splitting the original data into k different folds of the same size and iterate the hold out process using $k-1$ folds for training the model and 1 fold for performance evaluation).

However, the general outcome of these two procedure is really similar: they both return a partition of the initial dataset into a *training set* and a *test set*. Moreover, the data upon which the general predictions are performed are contained in the *evaluation set*.

The way in which the performance of the model is measured is defined as a function used to evaluate the *goodness* of a prediction obtained with the learned θ.

A first well known metric to evaluate classification performance is *accuracy*.

This metric is a defined as the ratio between the number of correctly assigned objects with respect to the total number of assignations. However, this metric is not robust to class label imbalance. Moreover, it is typically too general to be used as a metric in the different fields in which classification is often performed.

Different metrics from accuracy are the followings. In particular, for each class $j \in \{0, \dots, c-1\}$ we define:

- **recall(j)**: the ratio of correctly classified items in class j over the totality of items belonging to class j (i.e., the capability of the model of correctly classify those objects that actually belong to class j).
- **precision(j)**: the ratio of correctly classified items in class j over the totality of items classified as j (i.e., the capability of the model of avoiding mis-classification of the object belonging to class j).
- **F1_score(j)**: the harmonic mean of precision(j) and recall(j).

While precision and recall might look similar, they actually are at the opposites of classification outcome evaluation.

To further highlight this, recall that for the generic classification problem:

- TP indicates the number of objects belonging to class j and actually assigned to j.
- TN indicates the number of objects not belonging to class j and hence not assigned to j.
- FN indicates the number of objects assigned to class $i \neq j$ while they actually belong to class j.
- FP indicates the number of objects assigned to class j while they actually belong to class $i \neq j$.

We introduced this indices to give a precise formulation for precision and recall. In particular:

$$precision(j) = \frac{TP(j)}{TP(j) + FP(j)} \tag{1}$$

$$recall(j) = \frac{TP(j)}{TP(j) + FN(j)} \tag{2}$$

It should be clear how these two metrics actually capture different aspects of the classification outcome: while *precision* is concerned about avoiding the "pollution" introduced in class j by class-j wrong assignments, *recall* is concerned about "attracting" all the objects actually belonging to class j.

Throughout this work, we decided to use as performance metric *recall*.

In the medical setting we are considering is indeed more important to minimize the number of False Negatives rather than minimizing False Positives.

We justify this by noting that this work revolves around the possibility of developing early screenings methodologies that shall anyway be followed by deeper analysis: while passing trough different examinations concluding that the subject is fundamentally healthy is certainly not pleasant, we believe that for this specific objective minimizing the number of people who are not alerted of a possibly dangerous situation is certainly more relevant to the final scope of this work.

Moreover, even if we evaluated our models on the recall present in each class, we focused particularly on the *carcinoma* class.

This is motivated by the fact that is the class which is basically more associated with rapidly decay of health and deteration of tissues, as well as being the condition which is, at least in the short time, the most life-threatening among the ones considered.

4 The KNN classifier

Considering the small size of available data, approaches based on Deep-Learning techniques were not feasible. Moreover, considering the task at hand, we believe that good results can also be achieved with classical machine learning algorithms.

We believe that the difference between objects populating one single class is contained and that, more than that, the similarities between reading associated with one particular class label is particularly high. This belief is justified by many different reasons, one among many the fact that it is likely that tissues have electrical properties related to their structure in terms of the cells by which they are composed.

According to this hypothesis and to the small size of the dataset, we decided to use the K-Nearest Neighbor algorithm to perform classification.

This algorithm has been firstly introduced by Evelyn Fix and Joseph Hodges in 1951, and it can be classified as a "lazy" algorithm, since it does not actually learn anything other that the actual data distribution itself.

The overall schema of the algorithm is:

Algorithm 1 KNN algorithm

Let $\tau = \{(x_i, y_i)\}_{i=1}^{p}$ be the dataset
Let k be the *number of neighbors considered*
Let $\delta = \{(z_i)\}_{i=1}^{m}$ be the set of data points to be classified
Let $d : \mathbb{R}^n \mapsto \mathbb{R}$ be the *distance metric* considered
Compute $d(z_i, x_j)\ \forall i, j = 1, \ldots, p : i \neq j$
Let f be a voting schema
Assign z_i to the class y^* which receives the highest number of votes among the k nearest neighbors of z_i.

Typical choices for the value of k are small odd numbers (to prevent the occurrence of draws), whereas when it is not possible to use meaningfully introduce different voting mechanism, a typical choice for f is *uniform* voting, i.e. assigning to each vote a weight equal to $\frac{1}{k}$.

While being quite neat, this algorithm is highly interpretable (in the sense that predictions can be easily explained considering the characteristics of the neighbors of the point classified) and incremental, even if it, as a direct consequence of *the curse of dimensionality* it is not easily scalable.

The necessity of first computing the different distances among each pair of points in the training set and in test set is indeed bounding the application domain of this algorithm to not huge problems. Moreover, since this technique is mainly based on the concept of distance, the performance it reaches decreases as the dimensionality of the dataset increases. One of the main effects of the curse of dimensionality is indeed to reduce the actual meaning of the concept of distance itself.

To deploy this algorithm in high dimensionality problems what is typically done is to use dimensionality reduction techniques that reducing the number of components with respect to which each data point is represented, decrease the effects of the curse of dimensionality.

5 The KNN classifier in the original space

We ran the KNN algorithm in the case $k = 3$ on the original nine-dimensional dataset. We decided to perform a 60/40 hold-out of the dataset to evaluate the results.

While such a split is typically uncommon, we decided to use it considering that in the medical framework training data are typically not abundant, hence we wanted to test our models in the framework in which even less data are available.

Moreover, choosing a different split (like 70/30 or 80/20) would have created the situation in which very few samples per class would have been present in the test set, inevitably biasing the overall evaluation procedure.

Please note that this hold-out structure will rest the same throughout all this work.

5.1 Performance evaluation

With this configuration, we obtained the results presented in Table [1].

Class	***Recall Score***
adipose	1.00
carcinoma	0.875
fibro-adenoma, mastopathy and glandular	1.00
connective	0.833

Table 1: Recall Score using KNN on the original Dataset

We can see that the recall of the adipose and fibro-adenoma, mastopathy and glandular classes is clearly high, whereas the results obtained with the model with respect to carcinoma are not excellent: the 12.5 % of subjects who take the test are sent home when they present a serious condition as carcinoma.

Moreover, the confusion matrix associated with the classification in this space is presented in Figure [2].

We justify these results by noting that in the original nine-dimensional dataset, data are likely not clearly separable and therefore obtaining mis-classification with an algorithm such as KNN is perfectly in line with the expectations.

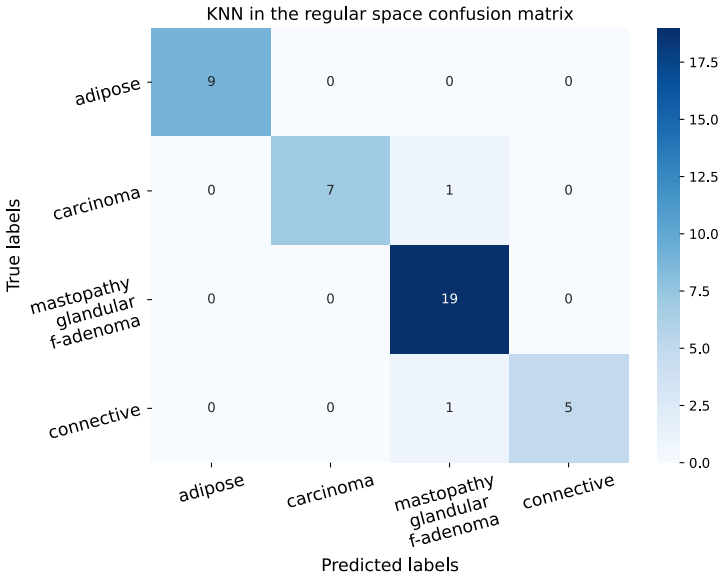

Figure 2: Confusion Matrix for the KNN classifier in the regular space

6 The KNN classifier in the PCA space

Considering the above mentioned effects of the dimensionality on the Classification outcome we resorted to apply Principal Component Analysis to map the original data into a space of reduced dimension.

The cumulative variance explained via adding principal components to the representation is presented in Figure [3].

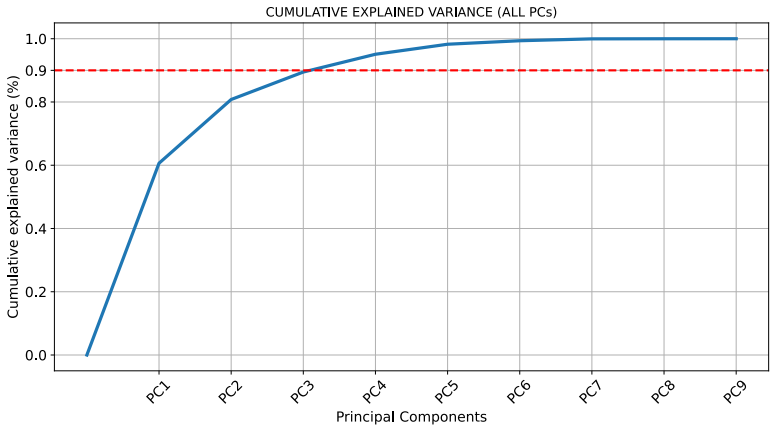

Figure 3: Cumulative variance obtained with PCs

The percentage of variance explained by each PC is presented in Figure [4].

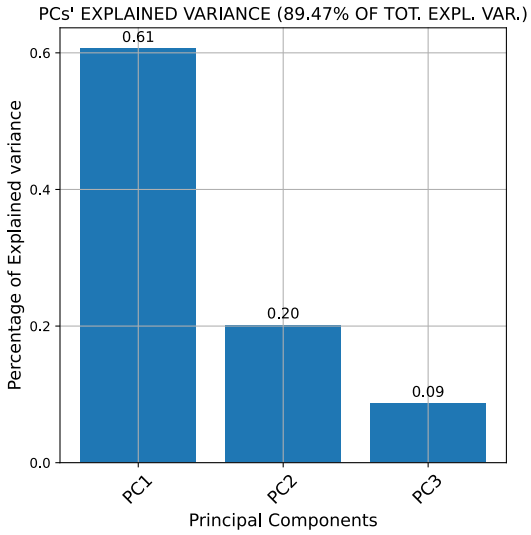

Figure 4: Variance explained by each PC

As it is possible to observe from Figure [3] and Figure [4], almost 90% of the total variance is explainable with three PCs and above 80% with two only.

Considering the high value of explained variance obtained with just two PCs, we plotted the data in the (PC1, PC2) plane. The results are presented in Figure [5].

As it is possible to observe in Figure [5], the data look fairly well separated, especially those records associated with adipose tissue, even if it is possible to observe a region in which the other three types of tissues are close to each other.

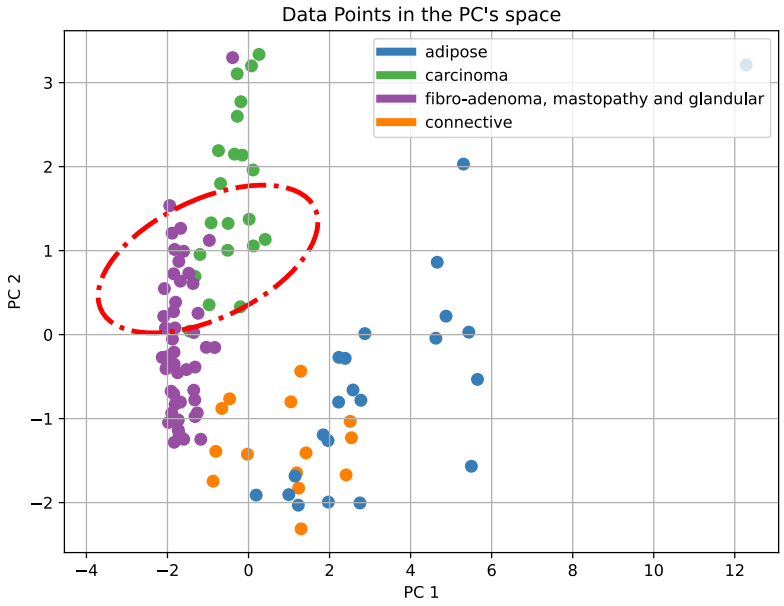

Figure 5: 2D scatter plot of the points in the PC space

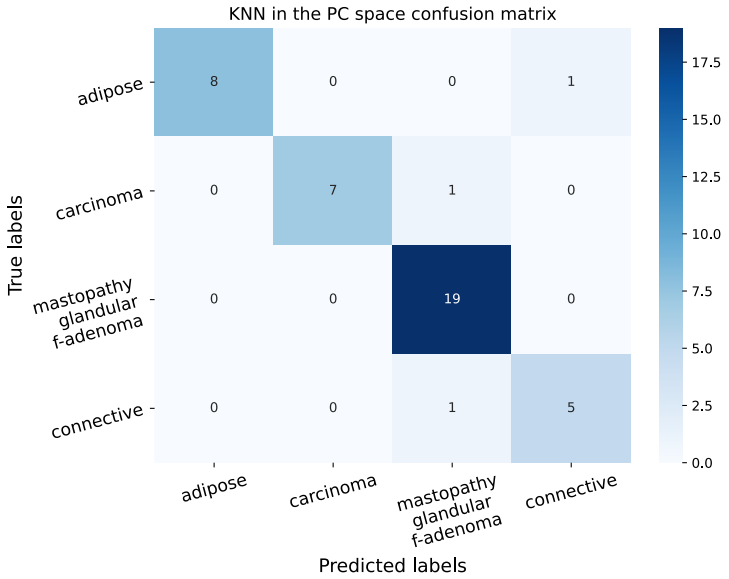

Figure 6: Confusion matrix related to the KNN classifier in the Principal Components space

The red-circled zone represents a probably problematic zone for the KNN algorithm: in that particular region the concentration of carcinoma, fibro-adenoma, mastopathy and glandular data is too homogeneous to avoid mis-classifications with the KNN. Essentially, the points that are just at the bound between the CAR zone and the FAD-MAS-GLA zone have an ambiguous number of neighbors of the different class, hence it is fairly likely that the errors of classification will occur in this region.

6.1 Performance evaluation

The results obtained using the KNN algorithm in the PC space (with three principal components) are presented in Table [2].

Class	***Recall Score***
adipose	0.889
carcinoma	0.875
fibro-adenoma, mastopathy and glandular	1.00
connective	0.833

Table 2: Recall Score using KNN in the PCA space

It is possible to see how in this case as well, accordingly to what before presented, the KNN classifier is still far from an acceptable outcome with respect to the carcinoma class.

Moreover, the recall for the adipose class has decreased with respect to the KNN in the regular space.

We concluded that while the data presented in the PC space are certainly more separable than they are in the regular space, the different points, even in the PC space, are still too mixed to observe the desired results.

The confusion matrix associated with the entire process of classification is presented in Figure [6].

We can conclude then that in the PC space the KNN classifier does not improve its classification performance with respect to the results obtained in the regular space. In particular, exactly as in the regular space, it is possible to observe the mis-clasification of 1 sample belonging to the carcinoma class.

We then resorted to combine dimensionality reduction with KNN using Liner Discriminant Analysis as dimensionality reduction technique. This methodology is briefly introduced in the next section.

7 A Theoretical Introduction to Linear Discriminant Analysis

As a starting note, please mind that this section is mainly based on what presented in [5].

Let us start recalling the definition of a *Bayesan Classifier*.

7.1 Bayesian Classifiers

Let us consider the following classification function

$$\hat{y} = g(x;\theta) = \underset{y\in\{0,\dots,c-1\}}{\arg\max}\ f(y|x;\theta) \tag{3}$$

where $f(y|x;\theta)$ is the *pdf* of class y conditioned to x, given the parameter θ. This term often takes the name of *posterior probability*.
The idea behind *Bayesian classification* is very simple. In order to estimate, for each class $y \in \{0,\dots,c-1\}$, $f(y|x,\theta)$ we can use the *Bayes' Theorem*, that states

$$f(y|x,\theta) \propto f(y)f(x|y,\theta)$$

where $f(y)$ is called *prior probability* of class y, while $f(x|y,\theta)$ is known as *likelihood* of obtaining feature vector x given class y.

7.2 Linear Discriminant Analysis

Linear Discriminant Analysis, or LDA, is a Bayesian classification model build over two main assumptions:

- for each $y \in \{0,\dots,c-1\}$ the conditioned distribution of the feature vector x, seen as a sample from a r.v. X, is a multivariate gaussian with mean vector μ_y and variance covariance matric Σ_y, namely

  $$X|Y = y \sim \mathcal{N}_n(\mu_y, \Sigma_y)\ .$$

  This assumption specifies that

  $$f(x|y,\mu_y,\Sigma_y) = ((2\pi)^n det(\Sigma_y))^{-1/2} \exp\left(-\frac{(x-\mu_y)^T\Sigma_y^{-1}(x-\mu_y)}{2}\right)$$

- we assume the classes to be homoscedastic, i.e. $\Sigma_y = \Sigma$ for each $y \in \{0,\dots,c-1\}$.

In this case our parameter set θ contains the prior probabilities of the classes, that we will call α_y, the mean vectors μ_y and the variance covariance matrix Σ, namely

$$\theta = \{\alpha_y, \mu_y, \Sigma,\ y = 0,\dots,c-1\}.$$

Let us point out that the prior probabilities $\alpha_y = f(y)$ forms a $c-$dimensional probability vector, hence they cannot take a negative value and their sum must be equal to one. Let us recall that we assign a vector x to class $\hat{y}$ according to the rule (3), i.e. if $\hat{y}$ satisfies

$$\alpha_{\hat{y}}((2\pi)^n det(\Sigma))^{-1/2} \exp\left(-\frac{(x-\mu_{\hat{y}})^T\Sigma^{-1}(x-\mu_{\hat{y}})}{2}\right) \geq \alpha_z((2\pi)^n det(\Sigma))^{-1/2} \exp\left(-\frac{(x-\mu_z)^T\Sigma^{-1}(x-\mu_z)}{2}\right) \tag{4}$$

for all $z \in \{0,\dots,c-1\}$. Taking the logarithmic transformation to each side of inequality (4) and getting rid of the terms that are constant with respect to the classes we obtain the following decision rule: assign x to the class $\hat{y}$ such that

$$\log\alpha_{\hat{y}} - \frac{1}{2}\mu_{\hat{y}}^T\Sigma^{-1}\mu_{\hat{y}} + x^T\Sigma^{-1}\mu_{\hat{y}} \geq \log\alpha_z - \frac{1}{2}\mu_z^T\Sigma^{-1}\mu_z + x^T\Sigma^{-1}\mu_z \tag{5}$$

for all $z \in \{0, \dots, c-1\}$. The function

$$\delta_y(x) = \log \alpha_z - \frac{1}{2}\mu_y^T \Sigma^{-1} \mu_y + x^T \Sigma^{-1} \mu_y$$

is called **linear discriminant function** and it defines the linear boundaries that separate the classes in the feature space $\mathbb{R}^n$.
Given a training set $\tau = \{(x_i, y_i)\}_{i=1}^p$ we can estimate $\{\alpha_y, \mu_y, \Sigma, y = 0, \dots, c-1\}$ via the following estimators:

- $\hat{\alpha}_y = \dfrac{n_y}{n} \quad \in [0,1]$
- $\hat{\mu}_y = n_y^{-1} \sum\limits_{j:y_j=y} x_j \quad \in \mathbb{R}^n$
- $\hat{\Sigma} = \sum_y \hat{\alpha}_y \hat{\Sigma}_y \quad \mathbb{R}^{n \times n}$

where $n_y = |\{j \in \{1, \dots, p\} : y_j = y\}|$ is the number of items within class y and

$$\hat{\Sigma}_y = n_y^{-1} \sum_{j:y_j=y} (x_j - \hat{\mu}_y)(x_j - \hat{\mu}_y)^T$$

is the estimation of the variance covariance matrix for each class.
It is possible to modify the method in order to simplify the classification algorithm, via a *spherification* of the data. Let us consider a non singular squared matrix $B \in \mathbb{R}^{n \times n}$. A suitable choice for B is the *Cholesky Factor* of the variance covariance matrix Σ, that by definition is symmetric positive definite. Recall that in this case B is such that $BB^T = \Sigma$.
Let us consider the linear map induced by the matrix B^{-1}

$$\begin{aligned} B^{-1} : \mathbb{R}^n &\to \mathbb{R}^n \\ x &\mapsto B^{-1}x = x' \ . \end{aligned}$$

Without loss of generality we suppose that the feature vectors x are distributed according to a mixture of multivariate Gaussian distributions with mean μ_y and variance covariance matrix Σ, i.e. the pdf of X is defined by

$$f_X(x;\theta) = \sum_{y \in \{0,\dots,c-1\}} \alpha_y ((2\pi)^n det(\Sigma))^{-1/2} \exp\left(-\frac{(x-\mu_y)^T \Sigma^{-1} (x-\mu_y)}{2} \right) \ . \tag{6}$$

Let us consider the r.v. $X' = B^{-1}X$, let us recall the result from *probability theory* that allows us to compute the pdf of linear transformations of a r.v. :

$$f_{X'}(x') = \frac{f_X(x)}{det(B^{-1})} \ . \tag{7}$$

Applying (7) to (6), and recalling that $\Sigma = BB^T$ it is possible to obtain the pdf of X':

$$\begin{aligned} f_{X'}(x;\theta) &= det(B) \sum_{y \in \{0,\dots,c-1\}} \alpha_y ((2\pi)^n det(B) det(B))^{-1/2} \exp\left(-\frac{(x-\mu_y)^T (BB^T)^{-1} (x-\mu_y)}{2} \right) \\ &= \sum_{y \in \{0,\dots,c-1\}} \frac{\alpha_y}{(2\pi)^{n/2}} \exp\left(-\frac{(B^{-1}(x-\mu_y))^T (B^{-1}(x-\mu_y))}{2} \right) \\ &= \sum_{y \in \{0,\dots,c-1\}} \frac{\alpha_y}{(2\pi)^{n/2}} \exp\left(-\frac{(x'-\mu'_y)^T (x'-\mu'_y)}{2} \right) \\ &= \sum_{y \in \{0,\dots,c-1\}} \frac{\alpha_y}{(2\pi)^{n/2}} \exp\left(-\frac{\|x'-\mu'_y\|^2}{2} \right) \ . \end{aligned}$$

So X' is distributed according to a mixture of multivariate Gaussians which components are mutually independent from each other. Geometrically we can see this as a spherification of each component in the mixture. More over applying the matrix B^{-1} to transform the data we obtain a new classification rule that depends only on the reshaped means and the prior probability of each class: assign x to class $\hat{y}$ whether

$$\hat{y} = \underset{y \in \{0,\dots,c-1\}}{\arg\min} (\|x' - \mu'_y\|^2 - 2\log \alpha_y).$$

7.3 Linear Discriminant Analysis for Data Reduction

More than that, LDA can also be used as a dimensionality reduction technique.
A natural way to perform data reduction can be deducted from the previous construction. Let B^{-1} be the map obtained from the Cholesky decomposition of the variance covariance matrix Σ. The idea is to perform the reduction projecting the data onto the $(c-1)-$dimensional affine space $\mathcal{H}$ spanned by the transformed means, i.e.

$$\mathcal{H} = span\{\mu'_i - \mu'_0, i = 1, \dots, c-1\}.$$

Let $H \in \mathbb{R}^{n \times (c-1)}$ be the matrix whose columns are spanning $\mathcal{H}$, namely we can choose H such that

$$H = [\mu'_1 - \mu'_0, \mu'_2 - \mu'_0, \dots, \mu'_{c-1} - \mu'_0].$$

We know that the orthogonal projector P mapping points in $\mathbb{R}^n$ onto $\mathcal{H}$ may be computed as

$$P = HH^+$$

where $H^+ = (H^T H)^{-1} H^T$ is the Moore-Penrose pseudoinverse of H. Knowing the SVD decomposition of H we can write P as

$$P = (USV^T)VS^+U^T = USS^+U^T = US(S^TS)^{-1}S^TU^T$$

where U, S, V are such that $H = USV^T$.
The final step is to compose the projection with a rotation performed by U^T, hence we get the matrix $\tilde{R} = U^T P$, namely

$$\tilde{R} = U^T US(S^TS)^{-1}S^TU^T = S(S^TS)^{-1}S^TU^T \quad \in \mathbb{R}^{n \times n}.$$

It is easy to prove that the last $n - (c-1)$ rows of $\tilde{R}$ are vanishing, hence we can define the operator R obtained removing the vanishing rows of $\tilde{R}$. Therefore we obtained the matrix $R \in \mathbb{R}^{n-(c-1) \times n}$ that maps each $n-$ dimensional feature vector into a smaller $c-1$ dimensional subspace.

8 KNN in the LDA space

In the previous section we introduced a classification methodology based on (4) and a new data reduction technique.

In particular, differently from what is happening with the PCA, we noted how the transformation of the data through B^{-1} has the effect of maximising the separability between the different classes, whereas the transformation of the data trough PCA has the main effect of capturing those directions that presents an higher variance.

Moreover, since the 4 classes/ 9 dimensional original problem has been mapped to a 3 dimensional space, the data points are also possible to visualize.

Before doing so, note the results presented in Figure [7].

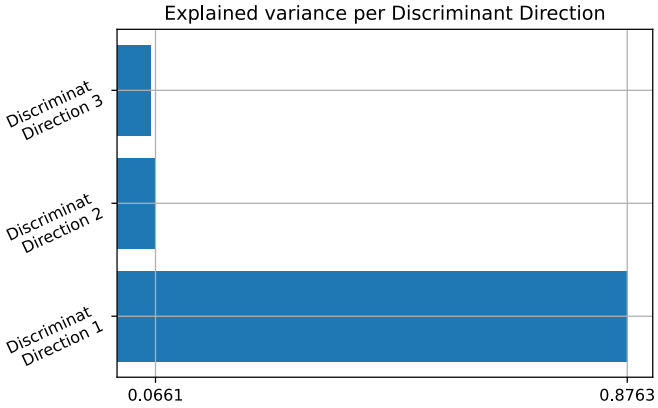

Figure 7: Explained variance per Discriminant Direction

Considering that above 90% of the total variance in obtained with the first two Discriminant Directions (DD, for short) we decided to visualize the projection of the overall three dimensional dataset in the LDA space onto the (DD1, DD2) plane. The results are presented in Figure [8].

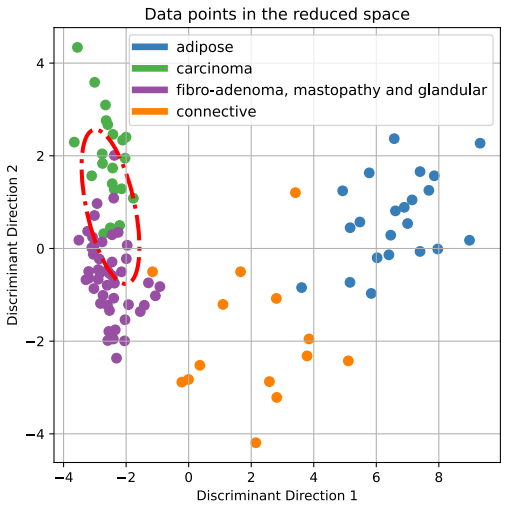

Figure 8: 2D scatter plot of the points in the DD space

Exactly as we did in the PC space, we highlighted what we thought could be a problematic region for the KNN algorithm. However, contrarily to what observed in the PC space, in the DD space this region is characterized by points which are belonging to different classes, but the number of neighbors belonging to the correct class in fairly more stable for the carcinoma, connective and adipose class.

It is anyway worth noting that error might occur in the classification of the FAD-MAS-GLA points.

Once we transformed the dataset in the LDA space we used once again the KNN algorithm on this new data points, finally reaching a valid and encouraging result with respect to recall.

8.1 Performance evaluation

The results obtained in terms of recall for the KNN on the LDA data are presented in Table [3].

Class	***Recall Score***
adipose	0.889
carcinoma	1.00
fibro-adenoma, mastopathy and glandular	1.00
connective	1.00

Table 3: Recall Score using KNN in the LDA space

To better explain these results it might be worth it to introduce also the confusion matrix for the overall classification process, presented in Figure [9].

In this case the elements on the diagonal are certainly more relevant than the ones outside of the region of the True Positives.

In particular, note how all the points belonging to the carcinoma class have been correctly classified, although being close to the region of the FAD-MAS-GLA, which is fairly more populated.

We justify this by recalling the scope of LDA itself. Maximizing the separability between the objects in the different classes, we obtained a separability that even simpler techniques as KNN could exploit to reach top performance.

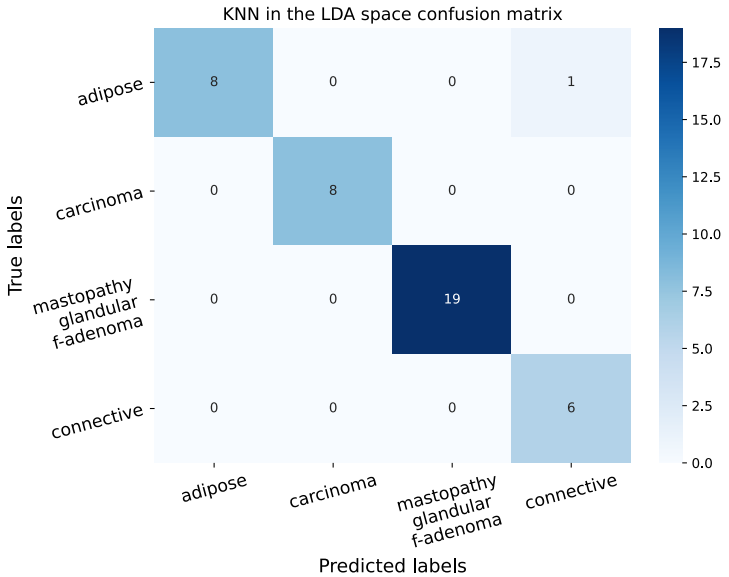

Figure 9: Confusion matrix related to the KNN classifier in the Discriminant Directions space

9 Our implementation of the LDA Class

The results we have presented in the previous section were obtained using the LDA class implementation present in scikit-learn [6].

In this section we present the results obtained implementing ourselves the LDA class. The results obtained mapping the points in the new DD space through our implementation, compared to the results obtained using scikit's one, are presented in Figure [10].

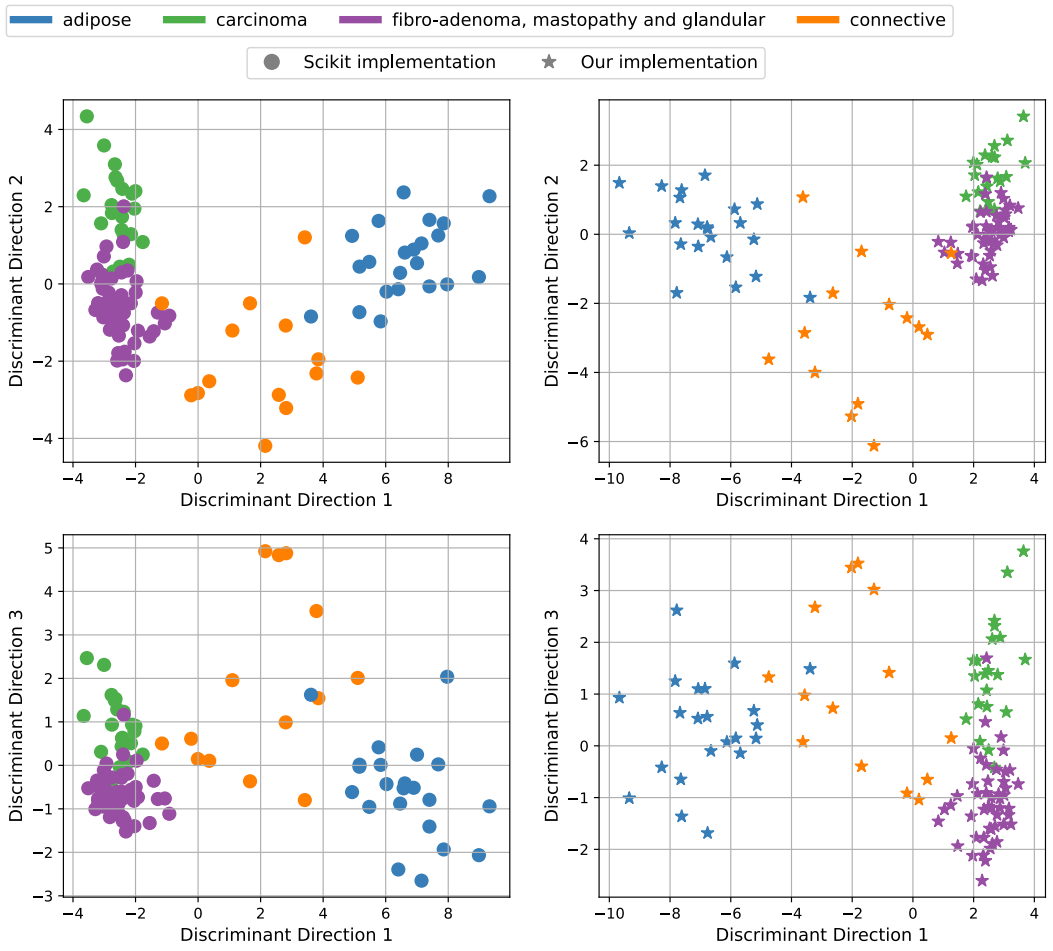

Figure 10: The points in the LDA space. Scikit implementation versus our implementation

The results are fairly similar in terms of structure, although they look different in terms of the rotation applied on the points. However, the overall concentration of class points in the different regions is perfectly comparable. Considering this similarity, the outcome of the classification process is expected to be comparable as well.

9.1 Performance evaluation

The results obtained with our implementation are presented in Table [4].

Note how, even if different rotations have been applied to the data, the results obtained are exactly the same.

We justify this by noting that it is the concentration of target variable values in each region to determine the outcome of the KNN classification process rather than the actual location of the region itself.

Class	***Recall Score***
adipose	0.889
carcinoma	1.00
fibro-adenoma, mastopathy and glandular	1.00
connective	1.00

Table 4: Recall Score using KNN in the LDA space obtained using our implementation

10 Conclusions

In this work we have presented the results obtained with the usage of a traditional algorithm such as KNN for a multi-label classification task.

Moreover, we presented the differences in terms of performance (measured by means of recall score) using KNN in the original space, in the space obtained using PCA and in the space obtained using LDA.

In particular, we noted that, as a direct consequence of the high separability among classes, the KNN algorithm reached the top performance in the Discriminant Direction space. The results, in terms of the weighted recall (i.e., the weighted average of recall, using as weights the concentration of each class label with respect to the total) and recall of che Carcinoma class are presented in Table [5].

	Weighted Recall	Carcinoma Recall
Regular KNN	0.952381	0.875
KNN in the PCA space	0.928571	0.875
KNN in the LDA space	0.976190	1.000
KNN in our LDA space	0.976190	1.000

Table 5: Overall results

We obtained this result specifically using 60% of the data for training. This shows once more that techniques based on Linear Algebra concepts (such as LDA) do not extremely suffer from the lack of enormous quantity of data.

Surely enough, the larger the quantity of available data is, the better, but as shown in this work, these techniques can cope, reaching outstanding results, also with shortage of information.

This fact is particularly relevant with respect to the medical field related applications considering that, as today, the quantity of information related to healthcare problems is no way even near to the giant quantity of data available for other topics, coming from other sources, such as data coming from IoT applications or user generated data.

Significant steps are anyway taken everyday in the direction of improving health through Machine Learning techniques.

The main improvement in this field come from the introduction in the market of devices which collect huge quantities of data related to the status of those who wear them (such as smartwatches or smart sleep monitors).

References

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] "Breast cancer risk factors." https://www.cdc.gov/cancer/breast/basic$_info/what - is - breast - cancer.htm$.

[3] "Breast cancer statistics." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3368191/B1.

[4] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[5] D. Kroese, Z. Botev, T. Taimre, and R. Vaisman, *Data Science and Machine Learning: Mathematical and Statistical Methods*. Chapman & Hall/CRC machine learning & pattern recognition, Boca Raton: CRC Press, 2019.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.