1 THE GENERAL 3D CLASSIFICATION PROBLEM

Assuming that the conditional covariance matrices are equal, the 3D classification problem can be formulated as follows:

$$(X_1, X_2, X_3) \sim \begin{cases} N((\mu_1^+, \mu_2^+, \mu_3^+), \Sigma), & \text{if } Y = +1 \\ N((\mu_1^-, \mu_2^-, \mu_3^-), \Sigma), & \text{if } Y = -1 \end{cases}$$

where

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix}$$

We also assume:

$$P(Y = +1) = P(Y = -1) = \frac{1}{2}$$

and:

$$\mu_+ = (1, 1, 1), \quad \mu_- = (-1, -1, -1)$$

Now, suppose the equation of the optimal separation plane is:

$$X_3 = d^* + e^* X_1 + f^* X_2$$

To minimize $P(\text{Error})$ with respect to d^*, e^*, and f^*, we must set their partial derivatives with respect to these coefficients to zero. Then, it can be shown that:

- $d^* = 0$ (due to general symmetry with respect to the origin $(0, 0, 0)$)
- Both e^* and f^* depend on the six parameters that define the matrix Σ, namely: σ_1, σ_2, σ_3, ρ_{12}, ρ_{13}, ρ_{23}.
- The minimum probability of error is $P(\text{Error}) = 1 - \Phi\left|\frac{1-e^*-f^*}{\sqrt{\Delta}}\right|$, where:
  - $\Delta = A^2 + B^2 + \lambda_{33}^2$, with: $A = e^*\lambda_{11} + f^*\lambda_{21} - \lambda_{31}$, $B = f^*\lambda_{22} - \lambda_{32}$
  - $\lambda_{11} = \sigma_1$, $\lambda_{12} = 0$, $\lambda_{13} = 0$
  - $\lambda_{21} = \rho_{12}\sigma_2$, $\lambda_{22} = \sigma_2\sqrt{1 - \rho_{12}^2}$, $\lambda_{23} = 0$
  - $\lambda_{31} = \rho_{13}\sigma_3$, $\lambda_{32} = (\rho_{23} - \rho_{12}\rho_{13})\sigma_3\sqrt{1 - \rho_{12}^2}$, $\lambda_{33} = \sigma_3\sqrt{1 - \rho_{13}^2 - (\rho_{23} - \rho_{12}\rho_{13})^2/(1 - \rho_{12}^2)}$
  - All this comes from the Cholesky decomposition of matrix Σ

2 PRECISION IN ERROR MEASUREMENT AS CARDINALITY INCREASES

Another aspect that should not be overlooked is the accuracy with which we can estimate the error rate of our classification procedure. Moreover, it is also interesting to investigate how each of the input parameters affects this precision. This is what we will discuss now. In particular, we will examine what happens to the accuracy of error rate estimation as the cardinality of the dataset n tends to infinity.

For this purpose, we first need some additional definitions: The VC dimension of the dictionary H is the maximum number of points in $\mathbb{R}^d$ that can be arbitrarily classified by classifiers in H. From now on, let H be a specific dictionary.

Furthermore, suppose we have a dataset D, which is representative of the phenomenon under study and consists of pairs (x_i, y_i), $i = 1, 2, ..., n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$, for all $i = 1, 2, ..., n$.

Then, for each classifier h in H, we can calculate its empirical error rate $\hat{L}(h)$, i.e., its proportion of misclassified observations in the dataset D:

$$\hat{L}(h) = \frac{1}{n} \sum_{i=1}^{n} 1(y_i \neq h(x_i)) \tag{2.1}$$

Let $\hat{h}^{(D)}$ be the best empirical classifier in the dictionary H, i.e., the classifier in H with the minimum empirical error rate. Under these conditions, we can define three error rates:

- $\min_{h \in H} L(h)$ = the smallest theoretically achievable error rate in H
- $L(\hat{h}^{(D)})$ = the theoretical error rate of the classifier $\hat{h}^{(D)}$
- $\hat{L}(\hat{h}^{(D)})$ = the empirical error rate of the classifier $\hat{h}^{(D)}$
- $\hat{L}(\hat{h}^{(D)}, D')$ = the error rate of the classifier $\hat{h}^{(D)}$ using a test dataset D'

It is a well-known fact that these error rates necessarily follow an order:

$$0 \leq \hat{L}(\hat{h}^{(D)}) \leq L(h^*) \leq \min_{h \in H} L(h) \leq L(\hat{h}^{(D)}),$$

where $L(h^*)$ is the Bayes error rate (also known as Bayes risk), i.e., the minimum theoretical loss generally achievable for this particular classification problem.

3 SIMULATION

Our intention is to compare "x_1, x_2 and x_3 present"with "only x_1 and x_2 present", in relation to the expected performance of the classifier, as the cardinality of the dataset n increases. In particular, we want to observe how the threshold n^* behaves for various different parameter combinations, given a specific scenario.

With x_3 present/absent, we will define a classification problem, characterized by:

1. a scenario,
2. a combination of parameters within that scenario,
3. and a dataset cardinality n.

For this classification problem, we will simulate a large number of replications, always finding the separating plane (or line) through linear SVM.

Given a 3D classifier, defined by the plane: $X_3 = d + eX_1 + fX_2$, the corresponding expected error probability is calculated by:

$$P(\text{Error}) = \frac{1}{2}(1 - \Phi(\text{Distance}(+))) + \frac{1}{2}(1 - \Phi(\text{Distance}(-))) = \\ = \frac{1}{2}\left(1 - \Phi\left(\frac{|d+e+f-1|}{\sqrt{\Delta}}\right)\right) + \frac{1}{2}\left(1 - \Phi\left(\frac{|-d+e+f-1|}{\sqrt{\Delta}}\right)\right). \tag{3.1}$$

Whenever x_3 is absent, we have a 2D classification problem and we proceed analogously

Thus, we will be able to plot two curves (with and without x_3), expressing $P(\text{Error})$ as a function of n. The threshold n^*, present in Table **??** is the abscissa of the point at which the two curves intersect.

For each of the four scenarios, one should:

1. Specify the parameter combinations to be simulated.
2. Run these simulations and obtain the two corresponding performance curves.
3. Interpret the results and connect these interpretations with the geometry of the problem.

4 WHEN DOES X_3 CONTRIBUTE TO INCREASING THE DISCRIMINATORY POWER OF (X_1, X_2)?

Given a classification problem with the features we are analyzing, that is, two trinormal distributions centered respectively at $(1, 1, 1)$ and $(-1, -1, -1)$, both with covariance matrices equal to

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix} \tag{4.1}$$

We know that, depending on the values of the 6 parameters (variances and correlations) that define this matrix Σ, there exists a threshold n^* from which the presence of the feature $X3$ becomes advantageous, given that the other two features $X1$ and $X2$ are already present.

Question: Given a matrix Σ, characterized by a particular combination of the 6 parameters σ_1, σ_2, σ_3, ρ_{12}, ρ_{13}, ρ_{23} that define it, how to assess the potential additional contribution of $X3$ to the discrimination between the two groups, once $X1$ and $X2$ are already present?

To analyze this issue, 4 scenarios were created, in each of which only 3 of the 6 parameters can be chosen freely. In each of these 4 scenarios, given the matrix Σ, the quotient $\frac{Risk2}{Risk3}$ can be calculated, where $Risk3$ and $Risk2$ are the Bayes risks (Bayes error) related, respectively, to the situation where all 3 features are present and to the situation where only $X1$ and $X2$ are present.

Presumably, the larger the quotient $\frac{Risk2}{Risk3}$, the smaller n^* should be.

On the other hand, we know that the level surfaces corresponding to a certain trinormal distribution are ellipsoids centered at the respective centroid (mean vector). Now, if the two centroids are $(1, 1, 1)$ and $(-1, -1, -1)$, the closer the direction of the main axis of these ellipsoids approaches the direction of the line passing through these two centroids, the greater the interchapter between the point clouds related to the two groups will be, that is, the expected error probability of the classifier will be higher, if all 3 features are present. As we can see in figures 1 and 2.

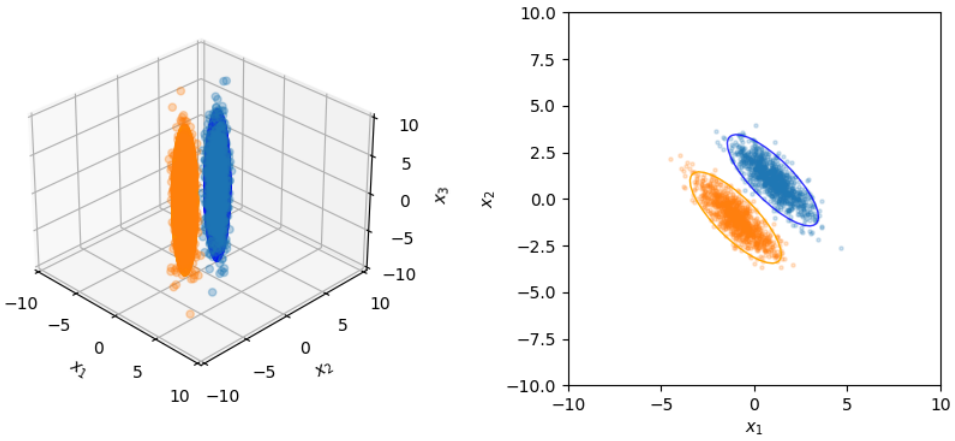

Figura 1 – $\sigma_{\mathbf{3}} = [1, 1, 4], \rho_{\mathbf{3}} = [-0.8, 0, 0]$ e $\sigma_{\mathbf{2}} = [1, 1], \rho_{\mathbf{2}} = [-0.8]$; $n = 1024$

$L_3(\hat{h}^{(D)}) = 0,000616, L_3(h^*) = 0,000756,\ L_3(\hat{h}^{(D)}) = 0,000929, \hat{L}_3(\hat{h}^{(D)}, D') = 0,000884$

$\hat{L}_2(\hat{h}^{(D)}) = 0,000664, L_2(h^*) = 0,000782,\ L_2(\hat{h}^{(D)}) = 0,000895, \hat{L}_2(\hat{h}^{(D)}, D') = 0,000852$

$\frac{Risco2}{Risco3} = 1,034513271, \frac{d_2}{d_3} = 1,003172086$

$n^* = 3155,\ para\ \hat{L}_{2\cap 3}(h^{(D)});\ n^* = 3477,\ para\ \hat{L}_{2\cap 3}(h^{(D)}, D')$

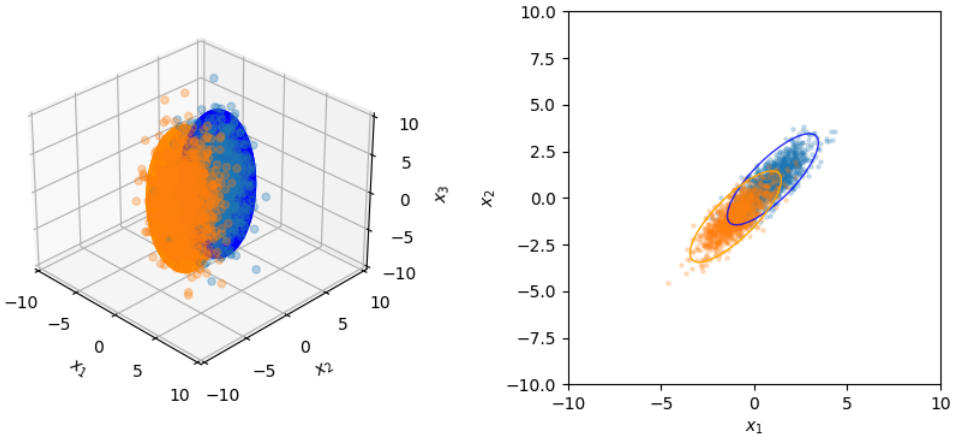

Figura 2 – $\sigma_{\mathbf{3}} = [1, 1, 4],\ \rho_{\mathbf{3}} = [0.8, 0, 0]$ *e* $\sigma_{\mathbf{2}} = [1, 1], \rho_{\mathbf{2}} = [0.8]$; $n = 1024$

$\hat{L}_3(\hat{h}^{(D)}) = 0,138705, L_3(h^*) = 0,139330,\ L_3(\hat{h}^{(D)}) = 0,140103, \hat{L}_3(\hat{h}^{(D)}, D') = 0,140113$

$\hat{L}_2(\hat{h}^{(D)}) = 0,145647, L_2(h^*) = 0,145920,\ L_2(\hat{h}^{(D)}) = 0,146405, \hat{L}_2(\hat{h}^{(D)}, D') = 0,146841$

$\frac{Risco2}{Risco3} = 1,047297879, \frac{d_2}{d_3} = 1,02773264$

$n^* = 51,\ para\ \hat{L}_{2\cap 3}(h^{(D)});\ n^* = 51,\ para\ \hat{L}_{2\cap 3}(h^{(D)}, D')$

4.1 A Model to Understand the Behavior of the Bayes Risk in Two Dimensions

Suppose only $X1$ and $X2$ are present. Consider an ellipse in $\mathbb{R}^2$ (say, centered at the origin $(0,0)$) whose equation is

$$x^T \Sigma^{-1} x = 1.$$

We have an ellipse for each class, and for each ellipse we have a centroid. The line that connects the two centroids $(1,1)$ and $(-1,-1)$ is the set of all vectors in $\mathbb{R}^2$ of the type

$$(c, c),$$

whose two coordinates are equal. This line intersects the said ellipse at a point that has two equal and positive coordinates,

$$(c_2, c_2) \cdot \Sigma^{-1} \begin{pmatrix} c_2 \\ c_2 \end{pmatrix} = 1.$$

Let d_2 be the distance from this point to the origin,

$$d_2 = \sqrt{2} c_2.$$

The greater this distance, the less the discriminatory power of the classifier, with X_1 and X_2 present.

4.2 Extending to the Case of Three Dimensions

Extending the discussion, let's consider an ellipsoid in $\mathbb{R}^3$ (say, centered at the origin $(0,0,0)$) whose equation is

$$x^T \Sigma^{-1} x = 1.$$

recall that now $x \in \mathbb{R}^3$.

Recalling, each class of points (for example, orange and blue) is associated with an ellipsoid, which in turn contains a centroid. The line that connects the two centroids is the set of all vectors in $\mathbb{R}^3$ of the type (c, c, c), whose three coordinates are equal. This line intersects the said ellipsoid at a point that has all three coordinates equal and positive.

$$(c_3, c_3, c_3) \cdot \Sigma^{-1} \begin{pmatrix} c_3 \\ c_3 \\ c_3 \end{pmatrix} = 1.$$

Let d_3 be the distance from this point to the origin,

$$d_3 = \sqrt{3} c_3.$$

The greater this distance, the less the discriminatory power of the classifier, with X_1, X_2 and X_3 present.

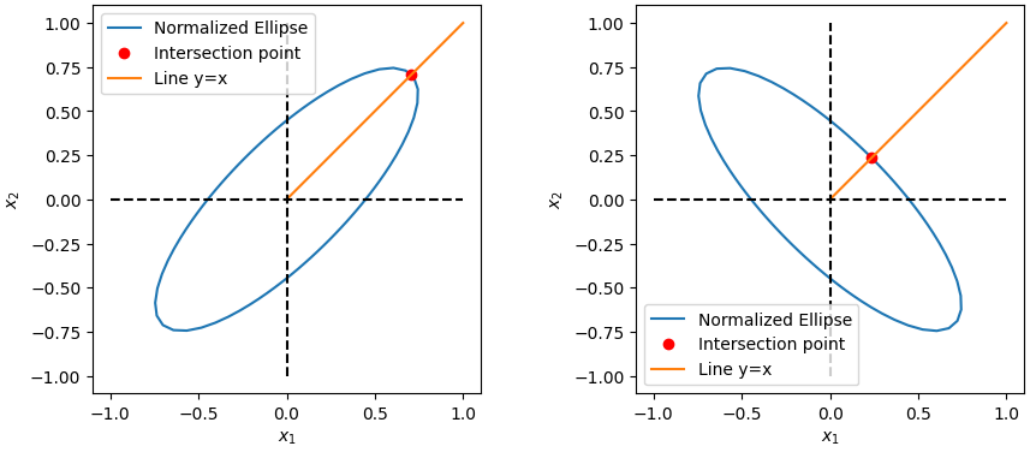

Figura 3 – $\sigma_2 = [1, 1], \rho_2 = [0.8]$ e $\sigma_2 = [1, 1], \rho_2 = [-0.8]$

4.3 Comparing Scenarios with Two and Three Features

How can we use d_3 and d_2 to create, from the matrix Σ, an indicator of the additional discrimination potential of X_3 given that X_1 and X_2 are already present? The indicator we are proposing is d_2/d_3,

$$U_3(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \frac{d_2}{d_3} = \sqrt{\frac{2}{3}}\frac{c_2}{c_3} \tag{4.2}$$

where U_3 is a utility metric of feature 3.

Let R_2 be the Bayes risk with 2 features, and R_3 the corresponding risk with 3 features. Empirically, we have found that the ratio of risks is a smooth and increasing function of the above utility function, i.e.,

$$\frac{R_2}{R_3} = \varphi(U_3) \tag{4.3}$$

where the function $\varphi(\cdot)$ is a smooth and increasing function, to be experimentally obtained in the following sections.

5 SCENARIOS

To simplify the mathematics, we will consider four specific scenarios. In each scenario, the idea is to reduce the number of free parameters, allowing us to obtain analytical expressions for the optimal coefficients of the plane, e^* and f^*, and the Bayes risk.

1. The three features are pairwise conditionally independent, given the label, i.e., $\rho_{12} = \rho_{13} = \rho_{23} = 0$.
   In this scenario, it can be demonstrated that:
   $$e^* = -\frac{\sigma_3^2}{2\sigma_1^2} \; ; \; f^* = -\frac{\sigma_3^2}{2\sigma_2^2} \; ; \text{ Bayes Risk } = 1 - \phi\left(\frac{\sqrt{1-e^*-f^*}}{\sigma_3}\right)$$

2. X_1 and X_2 are conditionally correlated, and (X_1, X_2) and X_3 are conditionally independent.

   - The features x_1 and x_2 are conditionally correlated, i.e., $\rho_{12} = \rho$ (which can be nonzero).
   - The features x_1 and x_3 are conditionally independent, i.e., $\rho_{13} = 0$.
   - The features x_2 and x_3 are conditionally independent, i.e., $\rho_{23} = 0$.
   - $\sigma_1 = \sigma_2 = \sigma$.

   In this scenario, it can be demonstrated that:
   $$e^* = f^* = -\frac{\sigma_3^2}{2\sigma^2(1+\rho)} \; ; \; P(Error) = 1 - \phi\left(\sqrt{1 - 2e^*\frac{1}{\sigma_3}}\right)$$

3. X_1 and X_2 are conditionally independent, and (X_1, X_2) and X_3 are conditionally correlated.

   - The features x_1 and x_2 are conditionally independent, i.e., $\rho_{12} = 0$.
   - The features x_1 and x_3 are conditionally correlated, i.e., ρ_{13} can be nonzero.
   - The features x_2 and x_3 are conditionally correlated, i.e., ρ_{23} can be nonzero.
   - $\rho_{13} = \rho_{23} = r$ (with restriction: $-\frac{\sqrt{2}}{2} < r < \frac{\sqrt{2}}{2}$).
   - $\sigma_1 = \sigma_2 = \sigma$.

   In this scenario, it can be demonstrated that:
   $$e^* = f^* = \frac{\sigma_3}{(\sigma_3 - r\sigma)} \cdot \frac{\sigma}{(2r\sigma_3 - \sigma)} \; ; \; P(\text{Error}) = 1 - \phi\left(\frac{1-2e^*}{\sqrt{\Delta}}\right), \text{ where}$$
   $$\Delta = 2(e\sigma - r\sigma_3)^2 + \sigma_3^2(1 - 2r^2).$$

4. X_1 and X_2 are conditionally correlated, and (X_1, X_2) and X_3 are conditionally correlated.

   - All features are equally dispersed ($\sigma_1 = \sigma_2 = \sigma_3 = \sigma$).
   - The features x_1 and x_2 are conditionally correlated, i.e., $\rho_{12} = \rho$ (which can be nonzero).
   - The features x_1 and x_3 are conditionally correlated, i.e., ρ_{13} can be nonzero.
   - The features x_2 and x_3 are conditionally correlated, i.e., ρ_{23} can be nonzero.
   - $\rho_{13} = \rho_{23} = r$ (with restriction: $-\sqrt{\frac{1+\rho}{2}} \leq r \leq \sqrt{\frac{1+\rho}{2}}$).

   In this scenario, it can be demonstrated that:

   $e^* = f^* = \frac{1-r}{2r-(1+\rho)}$; $P(\text{Error}) = 1 - \phi\left(\frac{1-2e^*}{\sqrt{\Delta}}\right)$, where
   $\Delta = A^2 + B^2 + \lambda_{33}^2$;
   $A = \sigma[e^*(1+\rho) - r]$;
   $B = \sigma\left[\sqrt{1-\rho^2}e^* - \frac{r(1-\rho)}{\sqrt{1-\rho^2}}\right]$;
   $\lambda_{33} = \sigma\sqrt{1 - 2r^2/(1+\rho)}$.