Derivatives of 3 different discriminat functions for multivariate Gaussian distribution

Compiled by Karthikeyan, CED16I015
Guided by
Dr Umarani Jayaraman

Department of Computer Science and Engineering
Indian Institute of Information Technology Design and Manufacturing
Kancheepuram

February 22, 2021

Introduction

We started with how this PDF actually influence the structure of the decision surface, because

$$g_{\mathrm{i}}(X) = lnP(X/\omega_{\mathrm{i}}) + lnP(\omega_{\mathrm{i}})$$

The purpose is to find $g_i(X)$ which is the maximum among all possible discriminant function.

$$g_{\mathrm{i}}(X) = lnP(X/\omega_{\mathrm{i}}) + lnP(\omega_{\mathrm{i}})$$

$$P(X/\omega_{\mathrm{i}}) = \frac{1}{(2\pi)^{\mathrm{d}/2}|\Sigma_{\mathrm{i}}|^{1/2}} exp[\frac{-1}{2}(X-\mu_i)^t {\Sigma_i}^{-1}(X-\mu_i)]$$

$$g_{\mathrm{i}}(X) = \frac{-1}{2}[(X-\mu_i)^t {\Sigma_i}^{-1}(X-\mu_i)] - \frac{d}{2}ln(2\pi) - \frac{1}{2}ln(|\Sigma_i|) + lnP(\omega_{\mathrm{i}})$$

Normal Density and Discriminant function

$$g_{\mathrm{i}}(X) = \frac{-1}{2}[(X-\mu_i)^t {\Sigma_i}^{-1}(X-\mu_i)] - \frac{d}{2} ln(2\pi) - \frac{1}{2} ln(|\Sigma_i|) + lnP(\omega_{\mathrm{i}})$$

- This is the discriminant function for the multivariate normal DF
- This classifier can take care of linearly non separable classes.
- When we take a decision boundary between two classes ω_{i} and ω_{j}; the decision surface is quadratic surface.
- It is not a linear surface. However for specific cases, this can be converted into a linear classifier.
- Depending upon the co-variance matrix Σ_{i} we can have different cases of discriminant function (i.e) **Case 1, Case 2 and Case 3**.

Case 1: Normal Density and Discriminant Function

Assumptions:

- In every class, the samples are clustered in hyper spherical of same shape and size
- The covariance matrix is of $\sigma^2 \mathrm{I}$
- The Σ_{i} is same for all classes where $i = 1, 2, .., c$

Case 1: $\Sigma_{\mathrm{i}} = \sigma^2 \mathrm{I}$ [I is Identity matrix]; given this,

- Determinant of

$$|\Sigma_i| = \sigma^{2\mathrm{d}} \Rightarrow \text{d number of diagonal values}$$

- Inverse of Σ_i : $\Sigma_i{}^{-1} = \frac{1}{\sigma^2} I$

Case 1: Normal Density and Discriminant function

- $\boxed{\text{When covariance matrix is same for all different classes } \forall \omega_{\mathrm{i}}}$

$$g_{\mathrm{i}}(X) = \frac{-1}{2}[(X-\mu_i)^t {\Sigma_i}^{-1}(X-\mu_i)] - \frac{d}{2} ln(2\pi) - \frac{1}{2} ln(|\Sigma_i|) + lnP(\omega_{\mathrm{i}})$$

$$-\frac{d}{2} ln(2\pi) \Rightarrow \text{Constant or independent of classes}$$

$$-\frac{1}{2} ln(\Sigma_i) \Rightarrow \text{Remains same for all classes, hence ignored}$$

- By substituting, $\Sigma_i^{-1} = \frac{1}{\sigma^2} I$

$$g_{\mathrm{i}}(X) = \frac{-1}{2}[(X - \mu_i)^t \Sigma_i^{-1}(X - \mu_i)] + lnP(\omega_{\mathrm{i}})$$

$$= \frac{-1}{2\sigma^2}[(X - \mu_i)^t (X - \mu_i)] + lnP(\omega_{\mathrm{i}})$$

$$= \frac{-1}{2\sigma^2}||X - \mu_i||^2 + lnP(\omega_{\mathrm{i}})$$

Case 1: Normal Density and Discriminant functions

$$g_{\mathrm{i}}(X) = \frac{-1}{2\sigma^2}||X - \mu_i||^2 + lnP(\omega_{\mathrm{i}})$$

$\boxed{\text{If } \mathrm{P}(\omega_{\mathrm{i}}) = P(\omega_{\mathrm{j}}) \text{ equal probability } \forall \text{ i,j} = 1,2,\ldots,\mathrm{c}}$

$$g_{\mathrm{i}}(X) = \frac{-1}{2\sigma^2}||X - \mu_i||^2 \;\; \Rightarrow \boxed{\text{squared Euclidean distance}}$$

- By taking negative; $g_{\mathrm{i}}(X)$ becomes maximum
- Regardless of whether the prior probabilities are equal or not; It is not actually necessary to compute distances.

Case 1: Normal Density and Discriminant functions

Expansion of the quadratic form yields:

$$g_{\mathrm{i}}(X) = \frac{-1}{2\sigma^2}[(X-\mu_i)^t(X-\mu_i)] + lnP(\omega_{\mathrm{i}})$$

$$= \frac{-1}{2\sigma^2}[X^tX - X^t\mu_i - \mu_i^tX + \mu_i^t\mu_i] + lnP(\omega_{\mathrm{i}})$$

$\mathrm{X^tX}$ is constant and same for all i $\Rightarrow g_i(X)$

$\boxed{\mathrm{X}^t\mu_i = \mu_i^tX}$ Next slide for explanation

$$= \frac{-1}{2\sigma^2}[-\mu_i^tX - \mu_i^tX + \mu_i^t\mu_i] + lnP(\omega_{\mathrm{i}})$$

$$= \frac{-1}{2\sigma^2}[-2\mu_i^tX + \mu_i^t\mu_i] + lnP(\omega_{\mathrm{i}})$$

Case 1: Normal Density and Discriminant functions

$$X = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} \mu = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

$$X^t \mu = \begin{bmatrix} 2 & 3 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} = [2^2 + 3 + 1] = 8$$

$$\mu^t X = \begin{bmatrix} 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} = [2^2 + 3 + 1] = 8$$

$$\boxed{\text{Hence } \mathrm{X}^t \mu_i = \mu_i^t X}$$

Case 1: Normal Density and Discriminant functions

$$g_i(X) = \frac{-1}{2\sigma^2}[-2{\mu_i}^t X + {\mu_i}^t \mu_i] + lnP(\omega_i)$$

$$= \frac{{\mu_i}^t X}{\sigma^2} - \frac{1}{2\sigma^2}{\mu_i}^t \mu_i + lnP(\omega_i)$$

$$\boxed{W_i = \frac{\mu_i}{\sigma^2}}$$

$\boxed{g_i(X) = {W_i}^t X + W_{i0}}$ $\Rightarrow$ linear equation or linear machine

$$\boxed{W_i = \frac{1}{\sigma^2}\mu_i}$$

$$\boxed{W_{i0} = \frac{-1}{2\sigma^2}\mu_i^t \mu_i + lnP(\omega_i)}$$

Case 1: Normal Density and Discriminant functions

discriminant function for individual class or i^{th} class is given by

$$\boxed{\mathrm{g_i}(\mathrm{X}) = \mathrm{W_i}^t X + W_{\mathrm{i}0}}$$

- If we want to find out the decision boundary between two different classes - ω_i and ω_j then let's understand
- What will be the nature of the decision boundary that separates the two classes ω_i and ω_j?

Case 1: Normal Density and Discriminant functions

- $g_i(X) = g_j(X)$ is the decision boundary
- $g_i(X) = W_i{}^t X + W_{i0}$
- $g_j(X) = W_j{}^t X + W_{j0}$
- $g(X) = g_i(X) - g_j(X) = 0$ is the equation of the decision boundary
- $g(X) = W_i{}^t X + W_{i0} - W_j{}^t X - W_{j0} = 0$

$$\boxed{g(X) = (W_i - W_j)^t X + W_{i0} - W_{j0} = 0}$$

Case 1:Normal Density and Discriminant functions

$$g(X) = (W_{\mathrm{i}} - W_{\mathrm{j}})^{\mathrm{t}} X + W_{\mathrm{i0}} - W_{\mathrm{j0}} = 0$$

$$= \frac{1}{\sigma^2}(\mu_{\mathrm{i}} - \mu_{\mathrm{j}})^t X - \frac{{\mu_{\mathrm{i}}}^t \mu_{\mathrm{i}}}{2\sigma^2} + lnP(\omega_{\mathrm{i}}) + \frac{{\mu_{\mathrm{j}}}^t \mu_{\mathrm{j}}}{2\sigma^2} - lnP(\omega_{\mathrm{j}} = 0$$

$$= \frac{1}{\sigma^2}(\mu_{\mathrm{i}} - \mu_{\mathrm{j}})^t X - \frac{1}{2\sigma^2}({\mu_{\mathrm{i}}}^t \mu_{\mathrm{i}} - {\mu_{\mathrm{j}}}^t \mu_{\mathrm{j}}) + lnP(\omega_{\mathrm{i}}) - lnP(\omega_{\mathrm{j}}) = 0$$

$$= \frac{1}{\sigma^2}(\mu_{\mathrm{i}} - \mu_{\mathrm{j}})^t X - \frac{1}{2\sigma^2}({\mu_{\mathrm{i}}}^t \mu_{\mathrm{i}} - {\mu_{\mathrm{j}}}^t \mu_{\mathrm{j}}) + ln\frac{P(\omega_{\mathrm{i}})}{P(\omega_{\mathrm{j}})} = 0$$

Multiply by σ^2

$$= (\mu_{\mathrm{i}} - \mu_{\mathrm{j}})^t X - \frac{1}{2}[(\mu_{\mathrm{i}} - \mu_{\mathrm{j}})^t(\mu_{\mathrm{i}} + \mu_{\mathrm{j}})] + \sigma^2 ln\frac{P(\omega_{\mathrm{i}})}{P(\omega_{\mathrm{j}})} = 0$$

Case 1:Normal Density and Discriminant functions

$$= (\mu_i - \mu_j)^t X - \frac{1}{2}[(\mu_i - \mu_j)^t(\mu_i + \mu_j)] + \sigma^2 ln\frac{P(\omega_i)}{P(\omega_j)} = 0$$

Take $(\mu_i - \mu_j)^t$ out

$$= (\mu_i - \mu_j)^t[X - \{\frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{(\mu_i - \mu_j)^t(\mu_i - \mu_j)} ln\frac{P(\omega_i)}{P(\omega_j)}.(\mu_i - \mu_j)\}]$$

$$= W^t[X - X_0] = 0$$

where

$$\boxed{W = (\mu_i - \mu_j)}$$

$$\boxed{X_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{||\mu_i - \mu_j||^2} ln\frac{P(\omega_i)}{P(\omega_j)}.(\mu_i - \mu_j)}$$

Case 1:Normal Density and Discriminant functions

$\boxed{\mathrm{W}^t[X - X_0] = 0}$ $\Rightarrow$ Decision boundary between i^{th} and j^{th} class

- W $=$ line joining μ_{i} and μ_{j} where μ_{i} and μ_{j} is vector
- Since $\boxed{W^t[X - X_0] = 0}$, the decision surface is orthogonal to the line joining μ_{i} and μ_{j}
- Since the decision boundary is linear, the surface which separates two classes is nothing but hyperplane.
- If $P(\omega_{\mathrm{i}}) = P(\omega_{\mathrm{j}})$, it turns out to be orthogonal bisector passing through X_0. This is also called as minimum distance classifier.

Case 1:Normal Density and Discriminant functions

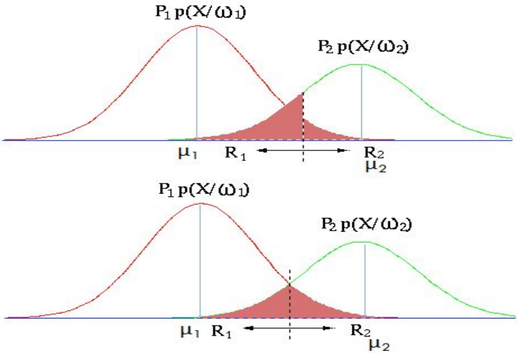

Figure: Decision Boundary

- If $P(\omega_1) == P(\omega_2)$ It is on the point X_0
- If $P(\omega_1) > P(\omega_2)$ The decision surface is away from μ_1
- If $P(\omega_2) > P(\omega_1)$ The decision surface is away from μ_2

Case 1: Summary

- $\Sigma_i = \sigma^2 I$; i = 1,2,...,c; All covariance matrix of type $\sigma^2 I$
- $\Sigma^{-1}{}_{i} = \frac{1}{\sigma^2}$
- Σ_i is of hyper sphere of same shape and size.
- $\boxed{W^t[X - X_0] = 0}$ is the decision surface and it is linear.
- It is Euclidean minimum distance classifier
- In 2d, it turns out to be $\Sigma_1 = \Sigma_2 = \begin{bmatrix} {\sigma_1}^2 & 0 \\ 0 & {\sigma_2}^2 \end{bmatrix} = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix}$
- x_1 and x_2 are independent

Case 2: Normal Density and Discriminant functions

Case 2 Assumption:

1. $\Sigma_i = \Sigma$

   Σ is arbitrary $\Rightarrow \Sigma_1 = \Sigma_2 = \begin{bmatrix} {\sigma_1}^2 & \sigma_{12} \\ \sigma_{21} & {\sigma_2}^2 \end{bmatrix}$
2. x_1 and x_2 are not necessarily independent
3. Σ_i is the same for all different classes
4. The samples are clustered in hyper ellipsoidal of same shape and size
5. $\sigma_{12} = \sigma_{21}$, hence symmetry

Case 2: Normal Density and Discriminant functions

$$g_{\mathrm{i}}(X) = \frac{-1}{2}[(X-\mu_i)^t{\Sigma_i}^{\text{-1}}(X-\mu_i)] - \frac{d}{2}ln(2\pi) - \frac{1}{2}ln(|\Sigma_i|) + lnP(\omega_{\mathrm{i}})$$

After ignoring constant $-\frac{d}{2}ln(2\pi)$ and $-\frac{1}{2}ln(|\Sigma_i|)$

$$g_{\mathrm{i}}(X) = \frac{-1}{2}[(X-\mu_i)^t{\Sigma_i}^{\text{-1}}(X-\mu_i)] + lnP(\omega_{\mathrm{i}})$$

- If all the classes are equal probable then
- Minimum distance classifier for
- **Case 1** - Squared Euclidean Distance
- **Case 2** - Squared Mahalanobis Distance

Case 2: Normal Density and Discriminant functions

Expansion of the quadratic form yields:

$$g_{\mathrm{i}}(X) = \frac{-1}{2}[(X-\mu_i)^t {\Sigma_i}^{-1}(X-\mu_i)] + lnP(\omega_{\mathrm{i}})$$

$$= \frac{-1}{2}[(X^t-\mu_i^t){\Sigma_i}^{-1}(X-\mu_i)] + lnP(\omega_{\mathrm{i}})$$

$$= \frac{-1}{2}[(X^t{\Sigma_i}^{-1}-\mu_i^t{\Sigma_i}^{-1})(X-\mu_i)] + lnP(\omega_{\mathrm{i}})$$

$$= \frac{-1}{2}[X^t{\Sigma_i}^{-1}X-\mu_i^t{\Sigma_i}^{-1}X-X^t{\Sigma_i}^{-1}\mu_i+\mu_i^t{\Sigma_i}^{-1}\mu_i] + lnP(\omega_{\mathrm{i}}) \Rightarrow (1)$$

Case 2: Normal Density and Discriminant functions

$X^t \Sigma_i^{-1} X$ is same for all classes and hence ignored

$$= \frac{-1}{2}[-2\mu_i^t \Sigma_i^{-1} X + \mu_i^t \Sigma_i^{-1} \mu_i] + lnP(\omega_{\mathrm{i}})$$

$$\boxed{\mu_i^t \Sigma_i^{-1} X == X^t \Sigma_i^{-1} \mu_i}$$

$$= \mu_i^t \Sigma_i^{-1} X - \frac{1}{2}[\mu_i^t \Sigma_i^{-1} \mu_i] + lnP(\omega_{\mathrm{i}})$$

$$\boxed{\mathrm{g_i(X) = W_i{}^t X + W_{i0}}} \Rightarrow \text{linear equation/ machine}$$

where
$W_{\mathrm{i}} = \mu_i \Sigma_i^{-1}$
$W_{\mathrm{i0}} = -\frac{1}{2}[\mu_i^t \Sigma_i^{-1} \mu_i] + lnP(\omega_{\mathrm{i}})$

Case 2: Normal Density and Discriminant functions

What will be the nature of the decision boundary that separates the two classes ω_i and ω_j?

$$g_i(X) - g_j(X) = 0$$

By deriving as like previous, it turned to

$$W^t(X - X_0) = 0$$

where,

$W = \Sigma^{-1}(\mu_i - \mu_j)$

$X_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{1}{(\mu_i - \mu_j)^t \Sigma^{-1}(\mu_i - \mu_j)} ln \frac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j)$

Case 3: Normal Density and Discriminant functions

Case 3: It is more general case

1. Σ_i is arbitrary; different classes have different covariance matrix; $\Sigma_i \neq \Sigma_j$
2. The decision surface is hyper quadratic in nature
3. Covariance matrix is arbitrary

From (1), We cant ignore anything here because of Σ_i is arbitrary in nature

$$= \frac{-1}{2}[X^t \Sigma_i^{-1} X - \mu_i^t \Sigma_i^{-1} X - X^t \Sigma_i^{-1} \mu_i + \mu_i^t \Sigma_i^{-1} \mu_i] + lnP(\omega_i) - \frac{1}{2} ln|\Sigma_i|$$

$$= \frac{-1}{2}[X^t \Sigma_i^{-1} X - 2\mu_i^t \Sigma_i^{-1} X + \mu_i^t \Sigma_i^{-1} \mu_i] + lnP(\omega_i) - \frac{1}{2} ln|\Sigma_i|$$

Case 3: Normal Density and Discriminant functions

$$\boxed{\mathrm{g_i}(\mathrm{X}) = \mathrm{X}^t A_\mathrm{i} X + B_\mathrm{i}{}^t X + C_\mathrm{i0}}$$

where

$$A_\mathrm{i} = \frac{-1}{2}\Sigma_i{}^{\text{-1}}$$

$$B_\mathrm{i} = \Sigma_i{}^{\text{-1}}\mu_i$$

$$C_\mathrm{i0} = \frac{-1}{2}\mu_i^t\Sigma_i{}^{\text{-1}}\mu_i - \frac{1}{2}ln|\Sigma_\mathrm{i}| + lnP(\omega_\mathrm{i})$$

The decision surface is quadratic hyperplane

Summary of all 3 cases

Multivariate case:

Case 1: $\Sigma_\mathrm{i} = \sigma^2\mathrm{I}$; Same for all class
Case 2: $\Sigma_\mathrm{i} = \Sigma$; Same for all class
Case 3: $\Sigma_\mathrm{i} \neq \Sigma_\mathrm{j}$; Different for different class

Bivariate case:

Case 1: ${\sigma_1}^2 = {\sigma_2}^2$; $\Sigma = \begin{bmatrix} {\sigma_1}^2 & 0 \\ 0 & {\sigma_2}^2 \end{bmatrix}$

Case 2: ${\sigma_1}^2 > {\sigma_2}^2$; $\Sigma = \begin{bmatrix} {\sigma_1}^2 & 0 \\ 0 & {\sigma_2}^2 \end{bmatrix}$

Case 3: $\Sigma = \begin{bmatrix} {\sigma_1}^2 & \sigma_{12} \\ \sigma_{21} & {\sigma_2}^2 \end{bmatrix}$