Classificador Bayesiano

Advanced Institute for Artificial Intelligence – AI2

https://advancedinstitute.ai

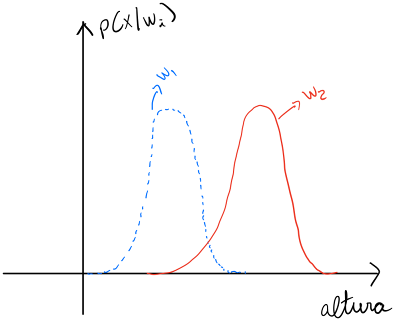
A Teoria de Decisão Bayesiana é um ferramental matemático que nos permite construir classificadores **paramétricos**, ou seja, técnicas que assumem a hipótese de que os dados seguem alguma distribuição (hipótese Gaussiana na grande maioria dos casos).

Definição do problema: seja $\mathcal{X}^1 = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$ um conjunto de dados de treinamento tal que $\boldsymbol{x}_i \in \mathbb{R}^n$ corresponde a uma amostra e $y_i \in \mathcal{Y}$ representa o rótulo dessa amostra, em que $\mathcal{Y} = \{\omega_1, \omega_2, \ldots, \omega_c\}$.

Além disso, temos os seguintes componentes em nosso ferramental:

- $p(\omega_i)$: **probabilidade a priori** da classe ω_i (proporção). Ex: em um problema de classificar indivíduos em jogadores de futebol (ω_1) ou basquete (ω_2), se nós temos 90 jogadores de futebol e 10 de basquete, então $p(\omega_1) = 0.9$ e $p(\omega_2) = 0.1$. Na prática, $p(\omega_1)$ denota a probabilidade de, ao selecionar algum jogador de maneira aleatória, ele ser um jogador de futebol. O mesmo vale para $p(\omega_2)$.

- $p(\boldsymbol{x}|\omega_i)$: **probabilidade condicional** da classe ω_i (verossimilhança). Ela descreve a função de densidade de probabilidade, ou seja, qual o comportamento de $\boldsymbol{x}$ dentro da classe ω_i. Ex: se $\boldsymbol{x}$ corresponde à altura do jogador em metros, $p(\boldsymbol{x}|\omega_i)$ descreve a distribuição das alturas dos jogadores de futebol e $p(\boldsymbol{x}|\omega_2)$ descreve a distribuição das alturas dos jogadores de basquete. Assumindo que temos uma hipótese Gaussiana, podemos representar $p(\boldsymbol{x}|\omega_i)$ como segue:



$$p(\boldsymbol{x}|\omega_1) \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$p(\boldsymbol{x}|\omega_2) \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

- $p(\omega_i|\boldsymbol{x})$: **probabilidade a posteriori** da classe ω_i, isto é, a probabilidade de decidirmos pela classe ω_i dada que observamos a amostra $\boldsymbol{x}$.

Dados esses três componentes, temos que a regra de Bayes é formulada como segue:

$$p(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)p(\omega_i)}{p(\boldsymbol{x})}, \tag{1}$$

em que $p(\boldsymbol{x})$ é uma constante normalizadora que não depende da classe, calculada como segue:

$$p(\boldsymbol{x}) = \sum_{j=1}^{c} p(\boldsymbol{x}|\omega_j)p(\omega_j). \tag{2}$$

A fórmula acima faz com que $p(\omega_i|\boldsymbol{x}) \in [0, 1]$.

Caso Unidimensional

Neste caso, temos que nossa amostra $x \in \mathbb{R}$, ou seja, temos apenas uma única característica. Como em nosso exemplo anterior de classificar um jogador como sendo de basquete ou futebol, assuma que x seja dado pela altura dos indivíduos.

No caso unidimensional, temos que a probabilidade condicional da classe ω_i é dada por:

$$p(x|\omega_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{\frac{-(x-\mu_i)^2}{2\sigma_i^2}\right\}. \tag{3}$$

A equação de uma função de densidade probabilidade Gaussiana possui dois parâmetros, isto é, a média μ_i e a variância σ_i^2, $\forall i = 1, 2, \ldots, c$. Assim sendo, a etapa de treinamento do classificador Bayesiano consiste em estimar esses parâmetros a partir dos dados de treinamento. Dado um problema com c classes, o objetivo é aprender esses parâmetros para cada Gaussiana, isto é, uma para cada classe. Denotamos por $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_c\}$ esse conjunto de parâmetros a ser aprendido, em que $\theta_i = (\mu_i, \sigma_i^2)$.

Um dos métodos mais conhecidos para aprendizado do conjunto de parâmetros $\boldsymbol{\theta}$ é o da máxima verossimilhança, ou seja, queremos maximizar a verossimilhança sobre o conjunto de dados de treinamento. Seja $\mathcal{X}_i^1 \subset \mathcal{X}^1$ o subconjunto dos dados de treinamento que contém apenas amostras da classe ω_i, $\forall i = 1, 2, \ldots, c$. Desta forma, o método da máxima verossimilhança consiste em encontrar θ_i que satisfaz a seguinte formulação:

$$\hat{\mu_i}, \hat{\sigma}_i^2 = \arg\max_{\mu_i, \sigma_i^2} \{p(\mathcal{X}_i^1 | \theta_i)\}. \tag{4}$$

Temos que $p(\mathcal{X}_i^1; \theta_i)$ corresponde à **densidade conjunta** das amostras da classe ω_i em função dos parâmetros:

$$p(\mathcal{X}_i^1|\theta_i) = \prod_{x_j \in \mathcal{X}_i^1} p(x_j|\theta_i), \tag{5}$$

em que $p(x_j|\theta_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{\frac{(x_j - \mu_i)^2}{2\sigma_i^2}\right\}$ possui a mesma formulação da Equação 3.

Para fins de tratabilidade matemática, é comum maximizar o logaritmo da verossimilhança, ou seja:

$$\log p(\mathcal{X}_i^1|\theta_i) = \sum_{x_j \in \mathcal{X}_i^1} \log p(x_j|\theta_i). \quad (6)$$

A equação acima, após algumas derivações matemáticas, resulta em:

$$\log p(\mathcal{X}_i^1|\theta_i) = -\frac{m_i'}{2}\log 2\pi - \frac{m_i'}{2}\log \sigma_i^2 - \frac{1}{2\sigma_i^2}\sum_{x_j \in \mathcal{X}_i^1}(x_j - \mu_i)^2, \quad (7)$$

em que $m_i' = |\mathcal{X}_i^1|$.

Assim sendo, nosso problema passar a ser encontrar o conjunto de parâmetros θ_i que maximiza a seguinte equação:

$$\hat{\mu}_i, \hat{\sigma}_i = \arg\max_{\mu_i \sigma_i^2} \{\log p(\mathcal{X}_i^1 | \theta_i)\}. \tag{8}$$

Note que essa formulação precisa ser realizada para todas as classes, ou seja, $i = 1, 2, \ldots, c$.

Para resolvermos o problema acima, basta calcular a derivada de $\log p(\mathcal{X}_i^1; \theta_i)$ em relação à cada um dos seus parâmetros e igualar à zero, ou seja:

$$\frac{\partial \log p(\mathcal{X}_i^1 | \theta_i)}{\partial \mu_i} \quad \text{e} \quad \frac{\partial \log p(\mathcal{X}_i^1 | \theta_i)}{\partial \sigma_i^2}.$$

Vamos calcular a derivada da função em relação ao parâmetro μ_i, ou seja:

$$
\begin{aligned}
\frac{\partial \log p(\mathcal{X}_i^1|\theta_i)}{\partial \mu_i} &= -\cancelto{0}{\frac{m_i'}{2}\log 2\pi} - \cancelto{0}{\frac{m_i'}{2}\log \sigma_i^2} - \frac{1}{2\sigma_i^2}\sum_{x_j \in \mathcal{X}_i^1}(x_j - \mu_i)^2 \\
&= -\frac{1}{\cancel{2}\sigma_i^2}\sum_{x_j \in \mathcal{X}_i^1}(x_j - \mu_i)^{\cancel{2}} = -\frac{1}{\sigma_i^2}\sum_{x_j \in \mathcal{X}_i^1}(x_j - \mu_i) = 0 \\
&= \Longrightarrow \frac{1}{\sigma_i^2}\sum_{x_j \in \mathcal{X}_i^1}(x_j - \mu_i) = 0.
\end{aligned} \tag{9}
$$

Dividindo ambos termos da Equação 9 por $1/\sigma_i^2$, temos que:

$$\begin{aligned}\sum_{x_j \in \mathcal{X}_i^1} (x_j - \mu_i) = 0 &\implies \sum_{x_j \in \mathcal{X}_i^1} x_j - \sum_{x_j \in \mathcal{X}_i^1} \mu_i = \sum_{x_j \in \mathcal{X}_i^1} x_j - m_i' \mu_i = 0 \\ &\implies -m_i'\mu_i = -\sum_{x_j \in \mathcal{X}_i^1} x_j \implies m_i'\mu_i = \sum_{x_j \in \mathcal{X}_i^1} x_j \\ &\implies \mu_i = \frac{1}{m_i'} \sum_{x_j \in \mathcal{X}_i^1} x_j,\end{aligned} \tag{10}$$

que é, basicamente, a equação da média amostral como conhecemos, ou seja, o melhor estimador possível!

Vamos calcular a derivada da função em relação ao parâmetro σ_i^2, ou seja:

$$
\begin{aligned}
\frac{\partial \log p(\mathcal{X}_i^1|\theta_i)}{\partial \sigma_i^2} &= \cancelto{0}{-\frac{m_i'}{2}\log 2\pi} - \cancelto{\frac{\partial \log a}{\partial a} = 1/a}{\frac{m_i'}{2}\log \sigma_i^2} - \frac{1}{2\sigma_i^2}\sum_{x_j \in \mathcal{X}_i^1}(x_j-\mu_i)^2 \\
&= -\frac{m_i'}{2\sigma_i^2} - \cancelto{\frac{\partial -1/a}{\partial a} = 1/a^2}{\frac{1}{2\sigma_i^2}}\sum_{x_j \in \mathcal{X}_i^1}(x_j-\mu_i)^2 = -\frac{m_i'}{2\sigma_i^2} + \frac{1}{2\sigma_i^4}\sum_{x_j \in \mathcal{X}_i^1}(x_j-\mu_i)^2 = 0. \qquad (11)
\end{aligned}
$$

Multiplicando ambos termos da Equação 11 por $2\sigma_i^2$, temos que:

$$
\begin{aligned}
-m_i' + \frac{1}{\sigma_i^2} \sum_{x_j \in \mathcal{X}_i^1} (x_j - \mu_i)^2 = 0 &\implies \frac{1}{\sigma_i^2} \sum_{x_j \in \mathcal{X}_i^1} (x_j - \mu_i)^2 = m_i' \\
&\implies \sigma_i^2 = \frac{1}{m_i'} \sum_{x_j \in \mathcal{X}_i^1} (x_j - \mu_i)^2, \qquad (12)
\end{aligned}
$$

que também denota a equação conhecida da variância das amostras da classe ω_i.

Agora, como calculamos a função de decisão? Seja $d_i(x)$ a função de decisão que define o classificador Bayesiano sob hipótese Gaussiana no caso unidimensional para a classe ω_i. Temos que ela pode ser calculada da seguinte forma:

$$d_i(x) = p(x|\omega_i)p(\omega_i), \tag{13}$$

que é, basicamente, o numerador da Regra de Bayes (Equação 1), ou seja, um índice de pertinência da amostra x para a classe ω_i.

Para fins de tratabilidade matemática, apliquemos a função logarítimica na Equação 13:

$$\begin{aligned} d_i(x) &= \log(p(x|\omega_i)p(\omega_i)) \\ &= \log p(x|\omega_i) + \log p(\omega_i). \end{aligned} \tag{14}$$

Simplificando um pouco mais a Equação 14, temos que:

$$
\begin{aligned}
d_i(x) &= \log p(x|\omega_i) + \log p(\omega_i) \\
&= \log\left[\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left\{-\frac{1}{2\sigma_i^2}(x-\mu_i)^2\right\}\right] + \log p(\omega_i) \\
&= \log\left[\frac{1}{\sqrt{2\pi\sigma_i^2}}\right] + \log\left[\exp\left\{-\frac{1}{2\sigma_i^2}(x-\mu_i)^2\right\}\right] + \log p(\omega_i) \\
&= \log\left(2\pi\sigma_i^2\right)^{-1/2} - \frac{1}{2\sigma_i^2}(x-\mu_i)^2 + \log p(\omega_i) \\
&= -\frac{1}{2}\log\left(2\pi\sigma_i^2\right) - \frac{1}{2\sigma_i^2}(x-\mu_i)^2 + \log p(\omega_i) \\
&= -\frac{1}{2}\left[\log(2\pi) + \log\left(\sigma_i^2\right)\right] - \frac{1}{2\sigma_i^2}(x-\mu_i)^2 + \log p(\omega_i).
\end{aligned} \tag{15}
$$

Finalmente, a Equação 15 torna-se a seguinte:

$$
\begin{aligned}
d_i(x) &= \cancelto{\text{constante}}{-\frac{1}{2}\log(2\pi)} - \frac{1}{2}\log\left(\sigma_i^2\right) - \frac{1}{2\sigma_i^2}(x-\mu_i)^2 + \log p(\omega_i) \\
&= -\frac{1}{2}\log\left(\sigma_i^2\right) - \frac{1}{2\sigma_i^2}(x-\mu_i)^2 + \log p(\omega_i).
\end{aligned} \tag{16}
$$

A equação acima é a função de decisão final do classificador Bayesiano.

Na prática, o classificador Bayesiano funciona da seguinte maneira: supondo o nosso problema de classificação de jogadores de futebol (ω_1) ou de basquete (ω_2), precisamos calcular $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$ via treinamento.

Em seguida, para classificar uma amostra $x^* \in \mathcal{X}^2$, calculamos $d_1(x^*)$ e $d_2(x^*)$. Caso $d_1(x^*) > d_2(x^*)$, então a amostra x^* é atribuída à classe ω_1 (jogador de futebol). Caso contrário, x^* é atribuída à classe ω_2 (jogador de basquete).

Caso Multidimensional

Agora, suponha que cada amostra de nosso conjunto de dados possua mais atributos, isto é, $\boldsymbol{x} \in \mathbb{R}^n$, $n > 1$. A ideia do classificador Bayesiano é a mesma, ou seja, vamos maximizar o logaritmo da máxima verossimilhança para obter as formulações dos parâmetros de nossa função de densidade de probabilidade Gaussiana. No caso de uma Gaussiana multidimensional, a sua formulação é dada por:

$$p(\boldsymbol{x}_j|\theta_i) = \frac{1}{(2\pi)^{m_i'/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)\right\}, \tag{17}$$

em que $\theta_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denota o conjunto de parâmetros da Gaussiana da classe ω_i, em que $\boldsymbol{\mu}_i \in \mathbb{R}^n$ corresponde ao seu vetor de médias e $\boldsymbol{\Sigma}_i \in \mathbb{R}^{n\times n}$ denota a sua matriz de covariância.

Qual a função da matriz de covariância? Ela representa informações sobre as variâncias dos diferentes atributos, como segue:

$$\mathbf{\Sigma}_i = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \dots & \sigma_{2n} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \dots & \sigma_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \sigma_{n3} & \dots & \sigma_n^2 \end{bmatrix}, \tag{18}$$

em que σ_i^2 na diagonal principal corresponde à variância do atributo i, e σ_{ij} denota o valor da covariância entre os atributos i e j, ou seja, como o atributo i se comporta em relação ao valor do atributo j. Temos que a correlação entre elas é **positiva** quando o aumento em uma delas ocasiona aumento na outra. A correlação é **negativa** quando o valor de uma variável aumenta e a outra diminui, e **nula** quando não existe correlação entre elas.

Como mencionado anteriormente, a etapa de treinamento consiste em resolver a seguinte formulação:

$$\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i = \underset{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i}{\arg\max}\{\log p(\mathcal{X}_i^1|\theta_i)\}. \tag{19}$$

Lembrando que a probabilidade conjunta $p(\mathcal{X}_i^1|\theta_i)$ é dada por:

$$p(\mathcal{X}_i^1|\theta_i) = \prod_{\boldsymbol{x}_j \in \mathcal{X}_i^1} p(\boldsymbol{x}_j|\theta_i). \tag{20}$$

Temos, ainda, que:

$$
\begin{aligned}
\log p(\mathcal{X}_i^1|\theta_i) &= \sum_{\boldsymbol{x}_j \in \mathcal{X}_i^1} \log p(\boldsymbol{x}_j|\theta_i) \\
&= \sum_{\boldsymbol{x}_j \in \mathcal{X}_i^1} \log \left[\frac{1}{(2\pi)^{m_i'/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) \right\} \right] \\
&= \underbrace{-\frac{nm_i'}{2}\log(2\pi) - \frac{m_i'}{2}\log\left(|\boldsymbol{\Sigma}_i^{-1}|\right) - \frac{1}{2}\sum_{\boldsymbol{x}_j \in \mathcal{X}_i^1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)}_{\text{função que queremos derivar!}}
\end{aligned}
\tag{21}
$$

Para resolvermos, então, a Equação 18, temos que calcular a sua derivada em relação aos parâmetros e igualar à zero, de maneira similar a que fizemos no caso unidimensional. Começando pela variável $\boldsymbol{\mu_i}$, temos que:

$$
\begin{aligned}
\frac{\partial \log p(\mathcal{X}_i^1|\theta_i)}{\partial \boldsymbol{\mu_i}} &= \cancelto{0}{-\frac{nm'_i}{2}\log(2\pi)} - \cancelto{0}{\frac{m'_i}{2}\log(|\boldsymbol{\Sigma}_i^{-1}|)} - \frac{1}{2}\sum_{\boldsymbol{x}_j \in \mathcal{X}_i^1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) \qquad (22) \\
&= -\frac{1}{2}\sum_{\boldsymbol{x}_j \in \mathcal{X}_i^1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) = 0.
\end{aligned}
$$

Multiplicando-se ambos lados por $-2\boldsymbol{\Sigma_i}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T$, temos que:

$$
\sum_{\boldsymbol{x}_j \in \mathcal{X}_i^1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) = 0.
$$

Distribuindo o somatório, temos que:

$$\sum_{\boldsymbol{x}_j \in \mathcal{X}_i^1} \boldsymbol{x}_j - \sum_{\boldsymbol{x}_j \in \mathcal{X}_i^1} \boldsymbol{\mu}_i = 0.$$

Desta forma, temos que:

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{m_i'} \sum_{\boldsymbol{x}_j \in \mathcal{X}_i^1} \boldsymbol{x}_j, \tag{23}$$

que também corresponde à media dos elementos da classe ω_i.

Precisamos, agora, calcular a derivada da Equação 21 em relação ao parâmetro $\boldsymbol{\Sigma_i}^{-1}$:

$$\frac{\partial \log(a)}{\partial a} = 1/a$$

$$\frac{\partial \log p(\mathcal{X}_i^1|\theta_i)}{\partial|\boldsymbol{\Sigma_i}^{-1}|} = \cancelto{0}{-\frac{nm_i'}{2}\log(2\pi)} - \frac{m_i'}{2}\log\left(|\boldsymbol{\Sigma}_i^{-1}|\right) - \frac{1}{2}\sum_{\boldsymbol{x}_j\in\mathcal{X}_i^1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i) \tag{24}$$

$$= \frac{m_i'\boldsymbol{\Sigma}_i}{2} - \frac{1}{2}\sum_{\boldsymbol{x}_j\in\mathcal{X}_i^1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^T(\boldsymbol{x}_j-\boldsymbol{\mu}_i) = 0.$$

Isolando $\boldsymbol{\Sigma}_i$, temos que:

$$\boldsymbol{\Sigma}_i = \frac{1}{m_i'}\sum_{\boldsymbol{x}_j\in\mathcal{X}_i^1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^T(\boldsymbol{x}_j-\boldsymbol{\mu}_i), \tag{25}$$

que é a expressão conhecida da matriz de covariância.

Temos que o termo exponencial da Gaussiana multivariada, isto é, $(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma_i}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)$, é conhecido como **Distância de Mahalanobis**. No caso da função de decisão, temos que:

$$
\begin{aligned}
d_i(\boldsymbol{x}) &= \log(p(\boldsymbol{x}|\omega_i)p(\omega_i)) \\
&= \log p(\boldsymbol{x}|\omega_i) + \log p(\omega_i) \\
&= \log\left[\frac{1}{(2\pi)^{m_i'/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)\right\}\right] + \log p(\omega_i) \\
&= -\frac{1}{2}\log(|\boldsymbol{\Sigma_i}|) - \frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma_i}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i) + \log p(\omega_i).
\end{aligned}
\tag{26}
$$

No entanto, dependendo da forma com a qual estimamos as matrizes de covariância, o comportamento do classificador Bayesiano se altera. Temos, basicamente, três casos:

- <u>Caso 1:</u> quando as matrizes de covariância das classes são iguais e diagonais, ou seja, utilizamos a mesma matriz de covariância para todas as classes.

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

Esse caso é bom quando temos um pequeno número de amostras por classe, pois podemos calcular a matriz utilizando todas as amostras de treinamento. **Note que temos uma mesma variância para todos os atributos.**

Neste caso, temos que $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, em que $\boldsymbol{I} \in \mathbb{R}^{n \times n}$ corresponde à matriz identidade. Para esta situação, a função de decisão dada pela Equação 26 pode ser escrita como segue:

$$
\begin{aligned}
d_i(\boldsymbol{x}) &= -\frac{1}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i) + \log p(\omega_i) \\
&= -\frac{1}{2}\log\left((\sigma^2)^n\right) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \frac{1}{\sigma^2}(\boldsymbol{x} - \boldsymbol{\mu}_i) + \log p(\omega_i) \\
&= \underbrace{\cancelto{0}{-\frac{n}{2}\log(\sigma^2)}}_{\text{constante em } i} - \frac{1}{2\sigma^2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T(\boldsymbol{x} - \boldsymbol{\mu}_i) + \log p(\omega_i) \\
&= -\frac{1}{2\sigma^2}\Big(\underbrace{\cancelto{0}{\boldsymbol{x}^T\boldsymbol{x}}}_{\text{constante em } i} - 2\boldsymbol{x}^T\boldsymbol{\mu_i} + \boldsymbol{\mu_i}^T\boldsymbol{\mu_i}\Big) + \log p(\omega_i) \\
&= \frac{1}{\sigma^2}\boldsymbol{x}^T\boldsymbol{\mu_i} - \frac{1}{2\sigma^2}\boldsymbol{\mu_i}^T\boldsymbol{\mu_i} + \log p(\omega_i).
\end{aligned}
\tag{27}
$$

A Equação 27 é, na verdade, **linear** em $\boldsymbol{x}$, ou seja:

$$d_i(\boldsymbol{x}) = w_{i1}^T\boldsymbol{x} + w_{i0}, \tag{28}$$

em que $w_{i1} = \frac{1}{\sigma^2}\boldsymbol{\mu_i}$ e $w_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu_i}^T\boldsymbol{\mu_i} + \log p(\omega_i)$ atua como sendo um bias.

Denotamos que a equação acima representa uma **função discriminante linear**. Assim sendo, podemos dizer que o classificador Bayesiano sob hipótese Gaussiana com uma única matriz de covariância diagonal é um classificador linear!

Suponha um problema com duas classes, ou seja, temos que calcular $d_1(\boldsymbol{x})$ e $d_2(\boldsymbol{x})$. Para obtermos a decisão de separação, basta igualarmos $d_1(\boldsymbol{x})$ e $d_2(\boldsymbol{x})$ e encontrarmos esse hiperplano analiticamente, como segue:

$$
\begin{aligned}
d_1(\boldsymbol{x}) = d_2(\boldsymbol{x}) &\implies w_{11}^T\boldsymbol{x} + w_{10} = w_{21}^T\boldsymbol{x} + w_{20} \\
&\implies \frac{1}{\sigma^2}\boldsymbol{\mu_1}\boldsymbol{x} - \frac{1}{2\sigma^2}\boldsymbol{\mu_1}^T\boldsymbol{\mu_1} + \log p(\omega_1) = \frac{1}{\sigma^2}\boldsymbol{\mu_2}\boldsymbol{x} - \frac{1}{2\sigma^2}\boldsymbol{\mu_2}^T\boldsymbol{\mu_2} + \log p(\omega_2) \\
&\overset{\times\sigma^2}{\implies} \boldsymbol{\mu_1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu_1}^T\boldsymbol{\mu_1} + \sigma^2\log p(\omega_1) = \boldsymbol{\mu_2}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu_2}^T\boldsymbol{\mu_2} + \sigma^2\log p(\omega_2) \\
&\overset{\text{rearranjando}}{\implies} \boldsymbol{\mu_1}^T\boldsymbol{x} - \boldsymbol{\mu_2}^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu_1}^T\boldsymbol{\mu_1} + \frac{1}{2}\boldsymbol{\mu_2}^T\boldsymbol{\mu_2} + \sigma^2\log p(\omega_1) - \sigma^2\log p(\omega_2) = 0 \\
&\overset{\text{evidência}}{\implies} \boldsymbol{x}(\boldsymbol{\mu_1}^T - \boldsymbol{\mu_2}^T) - \frac{1}{2}(\boldsymbol{\mu_1}^T\boldsymbol{\mu_1} - \boldsymbol{\mu_2}^T\boldsymbol{\mu_2}) + \sigma^2(\log p(\omega_1) - \log p(\omega_2)) = 0 \\
&\implies \boldsymbol{x}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T - \frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) + \sigma^2\left(\log\frac{p(\omega_1)}{p(\omega_2)}\right) = 0
\end{aligned}
\tag{29}
$$

Colocando $(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T$ em evidência na Equação 34, temos que:

$$\boldsymbol{x}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T - \frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) + \sigma^2\left(\log\frac{p(\omega_1)}{p(\omega_2)}\right) = 0 \tag{30}$$

$$(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T\left[\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) + \frac{\sigma^2}{(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T}\log\frac{p(\omega_1)}{p(\omega_2)}\right].$$

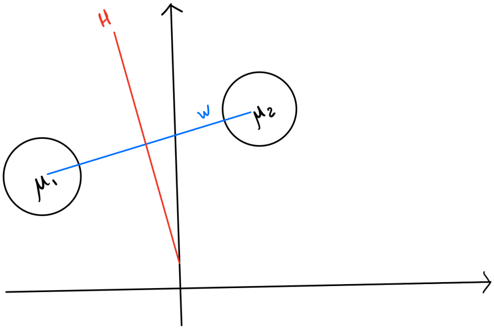
Assumindo que $\boldsymbol{w} = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$ e $\boldsymbol{b} = -\frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) + \frac{\sigma^2}{(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T}\log\frac{p(\omega_1)}{p(\omega_2)}$, temos a equação de um hiperplano:

$$\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{b}) = 0. \tag{31}$$

A Equação 26 define um **hiperplano separador** que passa sobre o bias $\boldsymbol{b}$, em que $\boldsymbol{w}$ denota a diferença entre as médias. Temos, então, que a solução da Equação 26 corresponde à todos os valores de $\boldsymbol{x}$ que são ortogonais à $\boldsymbol{w}$ e deslocados do bias $\boldsymbol{b}$.

Assumindo, por exemplo, que as classes são equiprováveis, ou seja, $p(\omega_1) = p(\omega_2)$, temos que $\log \dfrac{p(\omega_1)}{p\omega_2} = 0$ na Equação 35, resultando em $b = -\dfrac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$. Neste caso, temos que o hiperplano separador é ortogonal ao vetor $\boldsymbol{w}$, ou seja, ortogonal à diferença das médias e corta w no seu ponto médio (mediatriz).

A figura abaixo ilustra esta situação.



- <u>Caso 2:</u> quando as matrizes de covariância das classes são iguais mas não diagonais. Considerando, novamente, a função de decisão da Equação 26, temos que:

$$
\begin{aligned}
d_i(\boldsymbol{x}) &= \underbrace{\cancelto{0}{-\frac{1}{2}\log(|\boldsymbol{\Sigma}|)}}_{\text{constante em } i} - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + \log p(\omega_i) \\
&= \underbrace{\cancelto{0}{-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}}}_{\text{constante em } i} + \boldsymbol{\mu_i}^T\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu_i}^T\Sigma^{-1}\boldsymbol{\mu_i} + \log p(\omega_i) \\
&= \boldsymbol{\mu_i}^T\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu_i}^T\Sigma^{-1}\boldsymbol{\mu_i} + \log p(\omega_i).
\end{aligned} \tag{32}
$$

De maneira similar ao caso 1, a Equação 29 é, na verdade, **linear** em $\boldsymbol{x}$, ou seja:

$$d_i(\boldsymbol{x}) = w_{i1}^T \boldsymbol{x} + w_{i0}, \tag{33}$$

em que $w_{i1} = \boldsymbol{\mu_i}^T \boldsymbol{\Sigma}^{-1}$ e $w_{i0} = -\frac{1}{2}\boldsymbol{\mu_i}^T \Sigma^{-1} \boldsymbol{\mu_i} + \log p(\omega_i)$.

Novamente, temos uma **função discriminante linear**. Assim sendo, podemos dizer que o classificador Bayesiano sob hipótese Gaussiana com uma única matriz de covariância é um classificador linear!

Novamente, suponha um problema de classificação em duas classes, ou seja, temos que calcular $d_1(\boldsymbol{x})$ e $d_2(\boldsymbol{x})$. Para obtermos a função de decisão, basta igualarmos ambos termos, isto é, $d_1(\boldsymbol{x}) = d_2(\boldsymbol{x})$ e calcularmos o hiperplano separador analiticamente. Neste caso, temos que:

$$d_1(\boldsymbol{x}) = d_2(\boldsymbol{x}).$$

Expandindo a equação anterior, temos:

$$
\begin{aligned}
&\Longrightarrow \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^T\Sigma^{-1}\boldsymbol{\mu_1} + \log p(\omega_1) = \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_2^T\Sigma^{-1}\boldsymbol{\mu_2} + \log p(\omega_2) \\
&\Longrightarrow \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^T\Sigma^{-1}\boldsymbol{\mu_1} + \log p(\omega_1) - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_2} - \log p(\omega_2) = 0 \quad (34) \\
&\overset{\text{rearranjando}}{\Longrightarrow} \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^T\Sigma^{-1}\boldsymbol{\mu_1} + \frac{1}{2}\boldsymbol{\mu}_2^T\Sigma^{-1}\boldsymbol{\mu_2} + \log\left(\frac{p(\omega_1)}{p(\omega_2)}\right) = 0 \\
&\overset{\text{evidência}}{\Longrightarrow} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu}_2)\boldsymbol{x} - \frac{1}{2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1^T\boldsymbol{\mu_1} - \boldsymbol{\mu}_2^T\boldsymbol{\mu_2}) + \log\left(\frac{p(\omega_1)}{p(\omega_2)}\right) = 0 \\
&\overset{\text{evidência}}{\Longrightarrow} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu}_2)\left[\boldsymbol{x} - \frac{1}{2}\frac{(\boldsymbol{\mu}_1^T\boldsymbol{\mu_1} - \boldsymbol{\mu}_2^T\boldsymbol{\mu_2})}{(\boldsymbol{\mu_1} - \boldsymbol{\mu}_2)} + \frac{\log\left(\frac{p(\omega_1)}{p(\omega_2)}\right)}{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu}_2)}\right] = 0
\end{aligned}
$$

.

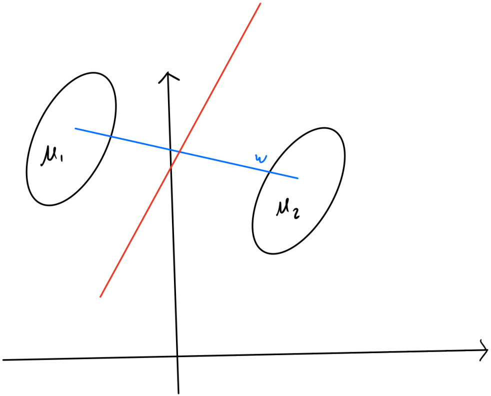
A Equação 37 pode ser escrita da seguinte forma:

$$\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{b}) = 0,$$

em que $w = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu}_2)$ e $b = \frac{1}{2}\frac{(\boldsymbol{\mu}_1^T\boldsymbol{\mu_1} - \boldsymbol{\mu}_2^T\boldsymbol{\mu_2})}{(\boldsymbol{\mu_1} - \boldsymbol{\mu}_2)} - \frac{\log\left(\frac{p(\omega_1)}{p(\omega_2)}\right)}{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu}_2)}$.

Neste caso, temos que o hiperplano separador não é ortogonal à diferença entre as médias, pois ele está rotacionado pela matriz de covariância $\boldsymbol{\Sigma}^{-1}$. **O hiperplano fica alinhado de acordo com a matriz de covariância.**

A figura abaixo ilustra esta situação.



- <u>Caso 3:</u> quando temos uma matriz de covariância para cada classe. Considerando, novamente, a função de decisão da Equação 26, temos que:

$$
\begin{aligned}
d_i(\boldsymbol{x}) &= -\frac{1}{2}\log(|\boldsymbol{\Sigma_i}|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma_i}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + \log p(\omega_i)\\
&= -\frac{1}{2}\log(|\boldsymbol{\Sigma_i}|) \underbrace{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma_i}^{-1}\boldsymbol{x}}_{\text{termo quadrático em } \boldsymbol{x}} + \boldsymbol{\mu}_i^T\boldsymbol{\Sigma_i}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma_i}^{-1}\boldsymbol{\mu_i} + \log p(\omega_i).
\end{aligned} \tag{35}
$$

Temos, agora, uma função discriminante quadrática, ou seja:

$$
d_i(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{W}_{i2}\boldsymbol{x} + \boldsymbol{W}_{i1}\boldsymbol{x} + \boldsymbol{W}_{i0}, \tag{36}
$$

onde $\boldsymbol{W}_{i2} = -\frac{1}{2}\boldsymbol{\Sigma_i}^{-1}$, $\boldsymbol{W}_{i1} = \boldsymbol{\mu_i}^T\boldsymbol{\Sigma_i}^{-1}$ e $\boldsymbol{W}_{i0} = -\frac{1}{2}\log(|\boldsymbol{\Sigma_i}|) - \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma_i}^{-1}\boldsymbol{\mu_i} + \log p(\omega_i)$.

Novamente, suponha um problema de classificação em duas classes, ou seja, temos que calcular $d_1(\boldsymbol{x})$ e $d_2(\boldsymbol{x})$. Para obtermos a função de decisão, basta igualarmos ambos termos, isto é, $d_1(\boldsymbol{x}) = d_2(\boldsymbol{x})$ e calcularmos o hiperplano separador analiticamente. **Essa é, geralmente, a versão mais utilizada do classificador Bayesiano pelo seu melhor poder de generalização. A figura abaixo ilustra esta situação.**