Critério MAP e Classificadores Bayesianos Gaussianos

Prof. Dr. Guilherme de Alencar Barreto

12 de fevereiro de 2021

Departmento de Engenharia de Teleinformática
Programa de Pós-Graduação em Engenharia Elétrica (PPGEE)
Universidade Federal do Ceará (UFC), Fortaleza-CE

gbarreto@ufc.br

1 Formulação

Nesta seção introduz-se o critério de decisão ótima, conhecido como critério MAP (*Maximum a posteriori*), além de quatro classificadores gaussianos obtidos a partir da suposição de que os exmepos de uma dada classe seguem uma lei de distribuição de probabilidades normal.

Para começar, assume-se que se está de posse de um conjunto de N pares $\{\mathbf{x}_n, \omega_n\}_{n=1}^{N}$, em que $\mathbf{x}_n \in \mathbb{R}^p$ representa o n-ésimo padrão[1] de entrada e ω_n é o rótulo da classe à qual pertence $\mathbf{x}_n$. Assume-se ainda que se tem um número finito e pré-definido de K classes ($K \ll N$), i.e. $\omega_n \in \{\omega_1, \omega_2, \ldots, \omega_K\}$. Por fim, seja n_i o número de exemplos da i-ésima classe (i.e. ω_i). Assim, $N = n_1 + n_2 + \cdots + n_K = \sum_{i=1}^{K} n_i$.

Primeiramente, seja $p(\omega_i)$ a probabilidade *a priori* da i-ésima classe. Esta é probabilidade de a classe ω_i ser selecionada *antes* do experimento ser realizado, sendo o experimento o ato de observar e classificar um certo padrão. Perceba que este é um experimento aleatório, visto que não sabemos de antemão a que classe o padrão será atribuído. Logo, uma modelagem probabilística é plenamente justificável.

O modelo probabilístico mais simples para $p(\omega_i)$ é a densidade de probabilidade uniforme, ou seja, assume-se que todos os padrões da i-ésima classe são equiprováveis, i.e. tem a mesma probabilidade de ser selecionado aleatoriamente. Assim, pode-se estimar $p(\omega_i)$ como

$$p(\omega_i) = \frac{n_i}{N}, \tag{1}$$

em que n_i o número de exemplos da i-ésima classe, conforme definido no parágrafo anterior.

Agora vamos olhar apenas para os dados da classe ω_i, ou seja, ao subconjunto de padrões $\mathbf{x}_n$ cujos rótulos são iguais a ω_i. Um modelo probabilístico comum para estes dados é a densidade normal multivariada, denotada por $p(\mathbf{x}_n|\omega_i)$, de vetor-médio $\mathbf{m}_i$ e matriz de covariância $\mathbf{\Sigma}_i$. Matematicamente, este modelo é dado pela seguinte expressão:

$$p(\mathbf{x}_n|\omega_i) = \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}_n - \mathbf{m}_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}_n - \mathbf{m}_i)\right\}, \tag{2}$$

[1]Por padrão de entrada, entende-se um vetor de atributos descrevendo o objeto a ser classificado.

em que $|\mathbf{\Sigma}_i|$ denota o determinante da matriz de covariância $\mathbf{\Sigma}_i$ e $\mathbf{\Sigma}_i^{-1}$ denota a inversa desta matriz.

A densidade $p(\mathbf{x}_n|\omega_i)$, no contexto de classificação de padrões, também é chamada de *função de verossimilhança*[2] da classe ω_i. A função de verossimilhança da classe ω_i pode ser entendida como o modelo probabilístico que tenta explicar (ou seja, modela) como os dados estão organizados (i.e. distribuídos) nesta classe.

Supõe-se agora que um novo padrão $\mathbf{x}_n$ é observado. Pergunta-se então qual é a probabilidade de que este padrão pertença à classe ω_i? Em outras palavras, dado $\mathbf{x}_n$, qual a probabilidade de ocorrer ω_i? Esta informação pode ser modelada através da função densidade *a posteriori* da classe, $p(\omega_i|\mathbf{x}_n)$.

Através do Teorema da Probabilidade de Bayes, a densidade a posteriori $p(\mathbf{x}_n|\omega_i)$ pode ser relacionada com a densidade a priori $p(\omega_i)$ e a função de verossimilhança $p(\mathbf{x}_n|\omega_i)$ por meio da seguinte expressão:

$$p(\omega_i|\mathbf{x}_n) = \frac{p(\omega_i)p(\mathbf{x}_n|\omega_i)}{p(\mathbf{x}_n)}. \tag{3}$$

Um critério comumente usado para tomada de decisão em classificação de padrões é o critério do *máximo a posteriori* (MAP). Ou seja, um determinado padrão $\mathbf{x}_n$ é atribuído à classe ω_j se a moda da densidade a posteriori $p(\omega_j|\mathbf{x}_n)$ for a maior dentre todas. Em outras palavras, tem-se a seguinte regra de decisão:

$$\text{Atribuir } \mathbf{x}_n \text{ à classe } \omega_j, \text{ se } p(\omega_j|\mathbf{x}_n) > p(\omega_i|\mathbf{x}_n), \quad \forall i \neq j. \tag{4}$$

O critério MAP também é comumente escrito como

$$\omega_j = \arg \max_{i=1,\ldots,K} \{p(\omega_i|\mathbf{x}_n)\}, \tag{5}$$

em que o operador "arg max" retorna o "argumento do máximo", ou seja, o conjunto de pontos para os quais a função de interesse atinge seu valor máximo.

Ao substituir a Eq. (3) na regra de decisão do critério MAP, obtém-se uma nova regra de decisão, dada por

$$\text{Atribuir } \mathbf{x}_n \text{ à classe } \omega_j, \text{ se } p(\omega_j)p(\mathbf{x}_n|\omega_j) > p(\omega_i)p(\mathbf{x}_n|\omega_i), \quad \forall i \neq j, \tag{6}$$

em que o termo $p(\mathbf{x}_n)$ é eliminado por estar presente em ambos os lados da inequação. Em outras palavras, o termo $p(\mathbf{x}_n)$ não influencia na tomada de decisão feita por meio do critério MAP.

Nota 1 - A regra de decisão do critério MAP, na forma como mostrado na Eq. (6), destaca a importância da informação *a priori* no processo de decisão. Assim, se as classes forem desbalanceadas, ou seja, com números muito díspares de exemplos, as classes com mais exemplos (i.e., com maior $p(\omega_i)$) tenderão a dominar o processo decisório, tornando o classificador tendecioso para tais classes.

Na verdade, o critério MAP pode ser generalizado para usar qualquer *função discriminante* $g_i(\mathbf{x}_n)$, passando a ser escrito como

$$\text{Atribuir } \mathbf{x}_n \text{ à classe } \omega_j, \text{ se } g_j(\mathbf{x}_n) > g_i(\mathbf{x}_n), \quad \forall i \neq j. \tag{7}$$

Nota 2 - Esta generalização do critério MAP é que permite entender a rede perceptron multicamadas (MLP, sigla em Inglês) como um aproximador do classificador bayesiano ótimo. Mais detalhes em [1].

[2]Do inglês, *likelihood function*.

Nota 3 - É importante ressaltar que, em um sentido amplo, uma função discriminante $g_i(\mathbf{x}_n)$ é qualquer função matemática que fornece um escore que permita quantificar a pertinência do padrão $\mathbf{x}_n$ à classe ω_i. Assim, as classes podem ser ranqueadas (i.e. ordenadas) em função dos valores de suas respectivas funções discriminantes.

No contexto dos classificadores bayesianos gaussianos, uma das funções discriminantes mais utilizadas envolve o logaritmo natural da densidade $p(\omega_i|\mathbf{x}_n)$. Assim, o critério MAP também é comumente escrito como

$$\begin{aligned} g_i(\mathbf{x}_n) &= \ln p(\omega_i|\mathbf{x}_n), \\ &= \ln p(\omega_i)p(\mathbf{x}_n|\omega_i), \\ &= \ln p(\omega_i) + \ln p(\mathbf{x}_n|\omega_i), \quad (8) \\ &= g_i^{(1)}(\mathbf{x}_n) + g_i^{(2)}(\mathbf{x}_n), \quad (9) \end{aligned}$$

em que $\ln(u)$ é a função logaritmo natural de u e a função $g_i^{(2)}(\mathbf{x}_n) = \ln p(\mathbf{x}_n|\omega_i)$ é chamada de *função log-verossimilhança* da classe ω_i.

Substituindo a função de verossimilhança mostrada na Eq. (2) em $g_i^{(2)}(\mathbf{x}_n)$, chega-se à seguinte expressão:

$$\begin{aligned} g_i^{(2)}(\mathbf{x}_n) &= \ln\left[\frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}_i|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}Q_i(\mathbf{x}_n)\right\}\right], \quad (10) \\ &= -\frac{1}{2}Q_i(\mathbf{x}_n) - \frac{p}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma}_i|, \quad (11) \end{aligned}$$

em que $Q_i(\mathbf{x}_n) = (\mathbf{x}_n - \mathbf{m}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}_n - \mathbf{m}_i)$.

Nota-se que o termo $-\frac{p}{2}\ln 2\pi$ é constante e aparece nas funções discriminates de todas as classes ($i = 1, \ldots, K$). Logo, este termo não influencia na tomada de decisão, podendo ser eliminado. Assim, a função discriminante geral do classificador bayesiano gaussiano é dada por

$$g_i(\mathbf{x}_n) = -\frac{1}{2}Q_i(\mathbf{x}_n) - \frac{1}{2}\ln|\mathbf{\Sigma}_i| + \ln p(\omega_i). \quad (12)$$

Uma suposição comumente feita na prática é a de que as densidades a priori das classes são iguais, ou seja

$$p(\omega_1) = P(\omega_2) = \cdots = P(\omega_K), \quad (13)$$

o que equivale a supor que as classes são equiprováveis[3]. Com isto, é possível simplificar ainda mais a função discriminante mostrada na Eq. (12):

$$g_i(\mathbf{x}_n) = -\frac{1}{2}Q_i(\mathbf{x}_n) - \frac{1}{2}\ln|\mathbf{\Sigma}_i|, \quad (14)$$

uma vez que o termo $\ln p(\omega_i)$ é igual para todas as K funções discriminantes. Vale ressaltar que usar esta função discriminante equivale a reescrever o critério MAP como

$$\text{Atribuir } \mathbf{x}_n \text{ à classe } \omega_j, \text{ se } \ln p(\mathbf{x}_n|\omega_j) > \ln p(\mathbf{x}_n|\omega_i), \quad \forall i \neq j \quad (15)$$

de tal forma que a regra de decisão passa a depender somente das funções de log-verossimilhança das classes. Neste caso, o critério MAP passa a ser chamado de critério da máxima verossimilhança (*maximum likelihood criterion*, ML).

[3]Esta suposição pode ser encontrada na prática em situações nas quais o número de exemplos (padrões) por classe é aproximadamente igual.

2 Casos Particulares

Nesta seção vamos considerar duas suposições simplificadoras para a matriz de covariância $\mathbf{\Sigma}$ a fim de derivar dois classificadores gaussianos muito utilizados na prática. Mesmo que os dados não possuam a estrutura de covariância descrita nos casos particulares, é possível aplicar uma transformação linear aos dados originais de modo que a satisfazer às suposições simplificadoras.

- **Caso 1**: As estruturas de covariâncias das K classes são iguais, ou seja, suas matrizes de covariância são iguais. Em outras palavras,

$$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_K = \mathbf{\Sigma}. \tag{16}$$

  Neste caso, a função discriminante da classe ω_i passa a ser escrita simplesmente como

$$g_i(\mathbf{x}_n) = -\frac{1}{2}Q_i(\mathbf{x}_n) = -\frac{1}{2}(\mathbf{x}_n - \mathbf{m}_i)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_n - \mathbf{m}_i), \tag{17}$$

  em que o termo $-\frac{1}{2}\ln|\mathbf{\Sigma}_i|$ foi eliminado por não influenciar mais na tomada de decisão. Note que a função discriminante $g_i(\mathbf{x}_n)$ é proporcional a $Q_i(\mathbf{x}_n)$, que é a distância de Mahalanobis quadrática. Assim, para todos os efeitos, pode-se fazer $g_i(\mathbf{x}_n) = Q_i(\mathbf{x}_n)$, de tal forma que o critério de decisão passa a ser escrito como

$$\text{Atribuir } \mathbf{x}_n \text{ à classe } \omega_j, \text{ se } Q_j(\mathbf{x}_n) < Q_i(\mathbf{x}_n), \quad \forall i \neq j, \tag{18}$$

  o que, em palavras, significa classificar $\mathbf{x}_n$ como sendo da classe ω_j se a distância (de Mahalanobis) de $\mathbf{x}_n$ ao centróide da classe ω_j (i.e. $\mathbf{m}_j$) for *menor* que as distâncias de $\mathbf{x}_n$ aos centróides restantes.

  Na prática, opta-se por projetar o classificador baseado em distância de Mahalanobis conforme mostrado na Eq. (18) quando a matriz de covariância de alguma das classes existentes é singular, ou seja, não é invertível. A seguir, apresentaremos diferentes maneiras de gerar uma única matriz de covariância a partir das matrizes de covariância das classes, fato este que diminui a chance de a matriz comum não ser invertível.

  (i) **Matriz de Covariância Agregada - $\mathbf{\Sigma}_{pool}$**: Uma forma muito comum de se implementar o classificador gaussiano cuja função descriminante é mostrada na Eq. (17) envolve o uso da matriz de covariância agregada, definida como

$$\begin{aligned}\mathbf{\Sigma}_{pool} &= \left(\frac{n_1}{N}\right)\mathbf{\Sigma}_1 + \left(\frac{n_2}{N}\right)\mathbf{\Sigma}_2 + \cdots + \left(\frac{n_K}{N}\right)\mathbf{\Sigma}_K, \\ &= p(\omega_1)\mathbf{\Sigma}_1 + p(\omega_2)\mathbf{\Sigma}_2 + \cdots + p(\omega_K)\mathbf{\Sigma}_K, \\ &= \sum_{i=1}^{K} p(\omega_i)\mathbf{\Sigma}_i,\end{aligned} \tag{19}$$

  em que $p(\omega_i)$ é a probabilidade a priori da classe i. Percebe-se assim que a matriz $\mathbf{\Sigma}_{pool}$ é a média ponderada das matrizes de covariância das K classes, com os coeficientes de ponderação sendo dados pelas respectivas probabilidades a priori.

  A matriz $\mathbf{\Sigma}_{pool}$ costuma ser mais bem condicionada que as matrizes de covariância individuais e, por isso, sua inversa tende a causar menos problemas de instabilidade numérica.

(*ii*) **Método de Regularização de Friedman** [4] - Este método consiste na combinação linear da matriz de covariância agregada com as matrizes de covariância das classes. A matriz resultante é dada por

$$\boldsymbol{\Sigma}_i^{\lambda} = \frac{(1-\lambda)\mathbf{S}_i + \lambda\mathbf{S}_{pool}}{(1-\lambda)n_i + \lambda N}, \tag{20}$$

tal que $\mathbf{S}_i = n_i\boldsymbol{\Sigma}_i$, $\mathbf{S}_{pool} = N\boldsymbol{\Sigma}_{pool}$, e $0 \leq \lambda \leq 1$. Para os valores de λ nos extremos do intervalo, chegamos aos seguintes casos:

$$\boldsymbol{\Sigma}_i^{\lambda} = \begin{cases} \boldsymbol{\Sigma}_i, & \text{se } \lambda = 0 \\ \boldsymbol{\Sigma}_{pool}, & \text{se } \lambda = 1 \end{cases} \tag{21}$$

No primeiro caso ($\lambda = 0$), devemos usar a função discriminante da Eq. (14). Já para o segundo caso ($\lambda = 1$), devemos usar a função discriminante da Eq. (17). Caso contrário, o valor do hiperparâmentro λ pode ser encontrado, por exemplo, via busca em grade (*grid search*) dentro do intervalo $0 < \lambda < 1$.

- **Caso 2**: Para este caso, assume-se que os atributos de $\mathbf{x}_n$ são descorrelacionados entre si e possuem variâncias diferentes. Assim, tem-se que a matriz de covariância comum a todas as classes tem a seguinte estrutura:

$$\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2), \tag{22}$$

em que σ_k^2 é a variância do k-ésimo atributo, $k = 1, \ldots, p$. A principal vantagem deste caso particular está no fato de sempre existir a inversa de $\boldsymbol{\Sigma}$, que é dada por

$$\boldsymbol{\Sigma}^{-1} = \text{diag}\left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \ldots, \frac{1}{\sigma_p^2}\right). \tag{23}$$

Sabemos que o fato de a matriz de covariância ser diagonal corresponde a assumir que os atributos são descorrelacionados. Se, além disso, os atributos são gaussianos, então podemos afirmar que os atributos são estatisticamente independentes. Esta é a mesma suposição feita pelo popular classificador *Gaussian naive Bayes* (Bayes ingênuo gaussiano)[5]. Assim, uma forma alternativa de se implementar o classificador naive Bayes gaussiano, se dá através do uso da distância de Mahalanobis com a matriz inversa da Eq. (23).

No caso de os atributos serem correlacionados, é possível aplicar uma transformação linear ao conjunto de dados, de modo a descorrelacioná-los. Isto equivale a diagonalizar a matriz de covariância original, de tal modo a fazer com que a suposição do classificador naive Bayes seja de fato observada nos dados. Esta operação pode ser realizada através da aplicação de PCA (*principal component analysis*) ao conjunto original de dados. O novo conjunto de dados gerado terá uma matriz de covariância diagonal.

- **Caso 3**: Para este caso, os atributos de $\mathbf{x}_n$ são descorrelacionados entre si e possuem variâncias iguais. Neste caso, tem-se que a matriz de covariância comum a todas as classes é dada por

$$\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_p, \tag{24}$$

[4] Jerome H. Friedman (1989). "Regularized Discriminant Analysis", *Journal of the American Statistical Association*, 84(405):165–175.

[5] Domingos, Pedro; Pazzani, Michael (1997). "On the optimality of the simple Bayesian classifier under zero-one loss". *Machine Learning*. 29(2–3):103–137.

em que $\mathbf{I}_p$ é a matriz identidade de ordem p. Logo, tem-se que $\mathbf{\Sigma}^{-1} = \frac{1}{\sigma^2}\mathbf{I}_p$. Neste caso, a função discriminante da classe ω_i passa a ser escrita como

$$\begin{aligned} g_i(\mathbf{x}_n) &= -\frac{1}{2\sigma^2}(\mathbf{x}_n - \mathbf{m}_i)^T\mathbf{I}_p(\mathbf{x}_n - \mathbf{m}_i), & (25)\\ &= -\frac{1}{2\sigma^2}(\mathbf{x}_n - \mathbf{m}_i)^T(\mathbf{x}_n - \mathbf{m}_i), & (26)\\ &= -\frac{1}{2\sigma^2}\|\mathbf{x}_n - \mathbf{m}_i\|^2, & (27)\end{aligned}$$

em que $\|\mathbf{u}\|^2$ denota a norma euclidiana quadrática de $\mathbf{u}$. Assim, usando esta função discriminante e o critério de decisão MAP da Eq. (7) obtemos

$$\text{Atribuir } \mathbf{x}_n \text{ à classe } \omega_j, \text{ se } -\frac{1}{2\sigma^2}\|\mathbf{x}_n - \mathbf{m}_j\|^2 > -\frac{1}{2\sigma^2}\|\mathbf{x}_n - \mathbf{m}_i\|^2, \quad \forall i \neq j, \tag{28}$$

que é equivalente à seguinte regra de decisão baseada na distância ao centróide da classe:

$$\text{Atribuir } \mathbf{x}_n \text{ à classe } \omega_j, \text{ se } \|\mathbf{x}_n - \mathbf{m}_j\|^2 < \|\mathbf{x}_n - \mathbf{m}_i\|^2, \quad \forall i \neq j. \tag{29}$$

Em resumo, quando as matrizes de covariância das classes são diagonais e as variâncias dos atributos são iguais, o classificador gaussiano baseado em distância de Mahalanobis reduz-se ao classificador de distância euclidiana mínima ao centróide.

Observação Importante 1 - Conforme já mencionado, conjuntos de dados cujas matrizes de covariância das classes não são diagonais, ou seja, cujos os atributos são correlacionados, podem ser processados de forma a diagonalizar as matrizes de covariância. Para este propósito, pode-se utilizar a técnica PCA. Em processamento de sinais e imagens, este procedimento é comumente chamado de *embranquecimento* dos dados (*data whitening*).

Observação Importante 2 - Além de *embranquecer* (i.e. descorrelacionar) as variáveis de entrada, pode-se forçar que todas elas tenham variância unitária aplicando a seguinte transformação a cada uma das variáveis:

$$x'_j = \frac{x_j - \mu_j}{\sigma_j}, \qquad j = 1, \ldots, p, \tag{30}$$

em que μ_j e σ_j são, respectivamente, o valor médio e o desvio-padrão da j-ésima variável. Pode-se facilmente mostrar que x'_j tem média nula e a variância igual a 1. Tente!

3 Complexidade dos Classificadores Gaussianos

Nesta seção a complexidade das funções discriminantes dos classificadores gaussianos será analisada. Por complexidade da função discriminante, entende-se a forma matemática resultante para a fronteira de decisão entre as classes, que pode ser linear ou não.

Vamos considerar inicialmente os casos particulares. Primeiro, analisaremos o Caso 2 ($\mathbf{\Sigma} = \mathbf{I}_p$) da seção anterior. Em seguida, discutiremos o Caso 1 ($\mathbf{\Sigma}_i = \mathbf{\Sigma}$). Finalmente, trataremos do caso mais geral.

- $\mathbf{\Sigma} = \mathbf{I}_p$: Para este caso particular, vamos iniciar nossa análise usando a função discriminante mostrada na Eq. (26), ou seja

$$g_i(\mathbf{x}_n) = -\frac{1}{2}(\mathbf{x}_n - \mathbf{m}_i)^T(\mathbf{x}_n - \mathbf{m}_i). \tag{31}$$

Distribuindo os produtos no lado direito da equação anterior, chegamos ao seguinte resultado:

$$g_i(\mathbf{x}_n) = -\frac{1}{2}\left[\mathbf{x}_n^T\mathbf{x}_n - \mathbf{x}_n^T\mathbf{m}_i - \mathbf{m}_i^T\mathbf{x}_n + \mathbf{m}_i^T\mathbf{m}_i\right], \quad (32)$$

$$= -\frac{1}{2}\left[\mathbf{x}_n^T\mathbf{x}_n - 2\mathbf{m}_i^T\mathbf{x}_n + \mathbf{m}_i^T\mathbf{m}_i\right], \quad (33)$$

em que usamos o fato de que o produto escalar entre 2 vetores é uma operação comutável (daí, $\mathbf{x}_n^T\mathbf{m}_i = \mathbf{m}_i^T\mathbf{x}_n$).

Percebe-se que o termo $\mathbf{x}_n^T\mathbf{x}_n$ não influencia na tomada de decisão, uma vez que é independente da classe (i.e. não depende de i). Em outras palavras, este termo aparece com mesmo valor nas funções discriminantes de todas as classes. Logo, o termo $\mathbf{x}_n^T\mathbf{x}_n$ pode ser eliminado das funções discriminantes sem prejuízo ao resultado da classificação. Neste caso, a função discriminante da i-ésima classe passa a ser escrita como

$$g_i(\mathbf{x}_n) = \mathbf{m}_i^T\mathbf{x}_n - \frac{1}{2}\mathbf{m}_i^T\mathbf{m}_i. \quad (34)$$

Note que se fizermos $\boldsymbol{\beta}_i = \mathbf{m}_i$ e $b_i = -\frac{1}{2}\mathbf{m}_i^T\mathbf{m}_i$, a função discriminante da Eq. (34) pode ser escrita como $g_i(\mathbf{x}_n) = \boldsymbol{\beta}_i^T\mathbf{x}_n + b_i$, que nada mais é do que a equação de um hiperplano no espaço $p+1$. Conclui-se, portanto, que este classificador é linear.

- $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$: Para este outro caso particular, a análise é feita usando-se a função discriminante mostrada na Eq. (17). Neste caso, tem-se que

$$g_i(\mathbf{x}_n) = -\frac{1}{2}(\mathbf{x}_n - \mathbf{m}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \mathbf{m}_i). \quad (35)$$

Assim como foi feito na análise anterior, distribuindo os produtos no lado direito da equação acima, chegamos ao seguinte resultado:

$$g_i(\mathbf{x}_n) = -\frac{1}{2}(\mathbf{x}_n - \mathbf{m}_i)^T(\boldsymbol{\Sigma}^{-1}\mathbf{x}_n - \boldsymbol{\Sigma}^{-1}\mathbf{m}_i), \quad (36)$$

$$= -\frac{1}{2}\left[\mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n - \mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_i - \mathbf{m}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n + \mathbf{m}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_i\right], \quad (37)$$

$$= -\frac{1}{2}\left[\mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n - 2\mathbf{m}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n + \mathbf{m}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_i\right], \quad (38)$$

em que usamos o fato de que $\mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_i = \mathbf{m}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n$.

De modo muito semelhante ao caso anterior, o termo $\mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n$ não influencia na tomada de decisão, logo pode ser eliminado das funções discriminantes sem prejuízo ao resultado da classificação. Assim, a função discriminante da i-ésima classe passa a ser escrita como

$$g_i(\mathbf{x}_n) = \mathbf{m}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n - \frac{1}{2}\mathbf{m}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_i. \quad (39)$$

Note que se fizermos $\boldsymbol{\beta}_i = \boldsymbol{\Sigma}^{-1}\mathbf{m}_i$ e $b_i = -\frac{1}{2}\mathbf{m}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_i$, a função discriminante da Eq. (39) também pode ser escrita como $g_i(\mathbf{x}_n) = \boldsymbol{\beta}_i^T\mathbf{x}_n + b_i$, o que nos leva a concluir que este classificador também é linear.

- **Caso geral**: Este cenário utiliza a função discriminante mostrada na Eq. (14). Para este caso, tem-se que

$$g_i(\mathbf{x}_n) = -\frac{1}{2}(\mathbf{x}_n - \mathbf{m}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_n - \mathbf{m}_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i|. \tag{40}$$

Distribuindo os produtos no lado direito da equação acima, chegamos ao seguinte resultado:

$$\begin{aligned} g_i(\mathbf{x}_n) &= -\frac{1}{2}(\mathbf{x}_n - \mathbf{m}_i)^T(\boldsymbol{\Sigma}_i^{-1}\mathbf{x}_n - \boldsymbol{\Sigma}_i^{-1}\mathbf{m}_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i|, & (41)\\ &= -\frac{1}{2}\left[\mathbf{x}_n^T\boldsymbol{\Sigma}_i^{-1}\mathbf{x}_n - \mathbf{x}_n^T\boldsymbol{\Sigma}_i^{-1}\mathbf{m}_i - \mathbf{m}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{x}_n + \mathbf{m}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{m}_i\right] - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i|, & (42)\\ &= -\frac{1}{2}\left[\mathbf{x}_n^T\boldsymbol{\Sigma}_i^{-1}\mathbf{x}_n - 2\mathbf{m}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{x}_n + \mathbf{m}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{m}_i\right] - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i|, & (43) \end{aligned}$$

em que usamos o fato de que $\mathbf{x}_n^T\boldsymbol{\Sigma}_i^{-1}\mathbf{m}_i = \mathbf{m}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{x}_n$.

Ao contrário dos 2 casos particulares anteriores, não podemos desprezar o termo $\mathbf{x}_n^T\boldsymbol{\Sigma}_i^{-1}\mathbf{x}_n$, pois este assume valores distintos para classes distintas. Assim, a função discriminante da i-ésima classe passa a ser escrita como

$$g_i(\mathbf{x}_n) = -\frac{1}{2}\mathbf{x}_n^T\boldsymbol{\Sigma}_i^{-1}\mathbf{x}_n + \mathbf{m}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{x}_n - \frac{1}{2}\mathbf{m}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{m}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i|, \tag{44}$$

Note que se fizermos $\mathbf{B}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}$, $\boldsymbol{\beta}_i = \boldsymbol{\Sigma}_i^{-1}\mathbf{m}_i$ e $b_i = -\frac{1}{2}\mathbf{m}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{m}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i|$, a função discriminante da Eq. (44) pode ser escrita como $g_i(\mathbf{x}_n) = \mathbf{x}_n^T\mathbf{B}_i\mathbf{x}_n + \boldsymbol{\beta}_i^T\mathbf{x}_n + b_i$, que é a expressão geral de um hiperparabolóide. Isto nos leva a concluir que este classificador é não-linear, sendo comumente chamado de *classificador gaussiano quadrático*.

Referências

[1] D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley, and B. W. Suter. The multilayer perceptron as an approximation to a bayes optimal discriminant function. *IEEE Transactions on Neural Networks*, 1(4):296–298, 1990.