Análise Discriminante (AD)

Discriminar significa separar, classificar. Análise discriminante é, portanto, uma técnica que pode ser utilizada para classificação de uma amostra população. Para isso é necessário ter um conhecimento a priori dos vários grupos de uma população (treinamento). Este conhecimento permite a elaboração de uma função matemática chamada regra de classificação ou discriminação, que é utilizada para classificar novos elementos amostrais nos grupos já existentes. Exemplos: Na ecologia, algumas espécies de insetos são muito similares. A AD pode ajudar na classificação de insetos, bastando que, para cada espécie candidata, se tenham disponíveis, informações sobre o seu perfil geral em relação a algumas características morfológicas, como por exemplo: comprimento das patas, das asas, da mandíbula, das antenas, do tórax, etc. A tradução destas medidas em uma função matemática fará com que a classificação do inseto possa ser mais precisa do que a mera inspeção visual.

Construção de uma função de classificação

A idéia básica por trás da AD é determinar se os grupos são diferentes com relação à média de uma variável, e então usar essa variável para prever a que grupo um novo caso pertence.

Vamos considerar o caso de que dispomos de apenas duas populações e um conjunto de observações importantes de cada população. Se a distribuição de elementos for conhecida podemos usar o princípio da máxima versossimilhança. Suponha que uma escola adote um critério de classificação baseados em duas fases, na fase 1 todos os candidatos fazem as provas e somente os candidatos aprovados seguem para a fase 2. Seja X a nota de matemática e vamos considerar duas populações de alunos:

- população 1: aqueles que passaram na primeira fase, mas não conseguiram ser aprovados na segunda fase;
- população 2: aqueles que conseguiram ser aprovados em ambas as fases.

O objetivo é a partir destes dados, construir uma regra de classificação que permita identificar dentre aqueles que passaram na primeira fase, quais provavelmente serão classificados na segunda fase. Seja μ_1 a média da população 1 e μ_2 a média da população 2 e σ^2 a variância das duas populações. Calculemos a razão de verossimilhança para as duas populações:

$$\lambda(x) = \frac{\text{função densidade da população 1}}{\text{função densidade da população 2}} = \frac{f_1(x)}{f_2(x}$$

que no caso de uma distribuição normal torna-se:

$$\lambda(x) = \frac{\frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2\right\}}{\frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma}\right)^2\right\}} = exp\left\{\frac{1}{2}\left[\left(\frac{x-\mu_2}{\sigma}\right)^2 - \left(\frac{x-\mu_1}{\sigma}\right)^2\right]\right\} \quad (1)$$

Para uma nota x fixa, quando $\lambda(x) > 1$, o valor da função de densidade da população 1 ("não aprovados") calculada pelo respectivo valor x é maior do que aquele obtido usando-se a densidade da população 2, assim, se a razão $\lambda(x) > 1$, seria razoável classificar o candidato como um provável não aprovado na fase 2. Se $\lambda(x) < 1$, seria razoável classificá-lo como pertencente a população 2.

Se $\lambda(x) = 1$, temos um impasse, podendo classificá-lo como pertencente a população 1 ou a população 2. Este é um caso, em que necessitamos adicionar mais informações ao problema.

Vamos um exemplo, suponha $\mu_1 = 16$, $\mu_2 = 18$, $\sigma^2 = 4$. Veja que $\mu_2 > \mu_1$, pois num critério classificatório, a maioria fica abaixo da média de classificação. Logo, para um candidato que tivesse conseguido $x = 16,5$ na prova de matemática o valor de $\lambda(x)$ seria:

$$\hat{\lambda}(16,5) = exp\left\{\frac{1}{2}\left[\left(\frac{16,5-18}{2}\right)^2 - \left(\frac{16,5-16}{2}\right)^2\right]\right\} = 1,284$$

o que classificaria o candidato na população 1. Ou seja, $16,5$ está bem mais distante de 18, que é média da população 2, do que de 16, que é média da população 1. Veja a **Figura 1**, onde as duas curvas representam a a densidade da população 1 e 2 respectivamente. A linha tracejada seria o ponto de separação ou de corte entre as duas populações.

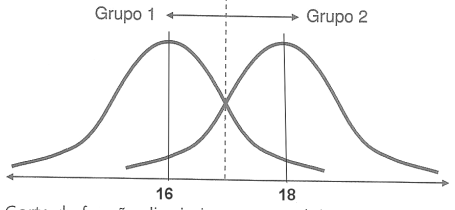


Figura 1 Corte da função discriminante para dois grupos com variâncias iguais.

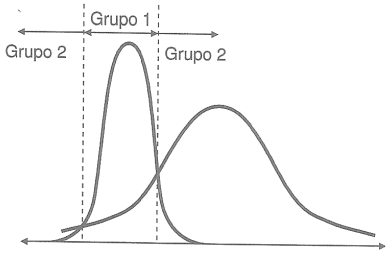


Figura 2 Corte da função discriminante para dois grupos com variâncias desiguais.

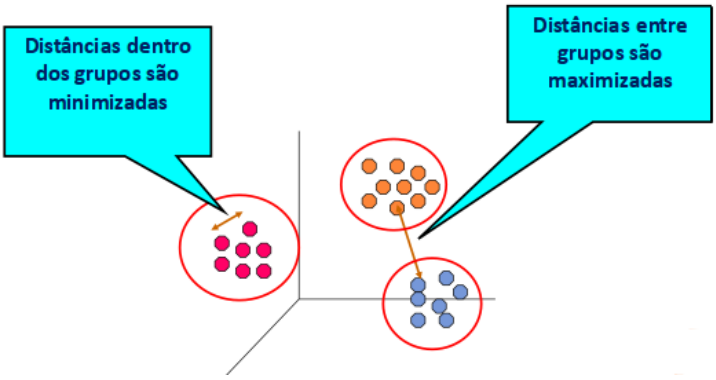
Para um candidato que obtivesse $x = 17,5$, o valor de $\lambda(x)$ seria $0,7788$, indicando que este seria um provável aprovado na segunda fase. Aplicando $-2ln$ em ambos os lados da equação (1), sendo ln o logaritmo natural, teremos:

$$-2ln(\lambda(x)) = \left(\frac{x-\mu_1}{\sigma}\right)^2 - \left(\frac{x-\mu_2}{\sigma}\right)^2 = \frac{1}{\sigma^2}\left[(x-\mu_1)^2 - (x-\mu_2)^2\right] \quad (2)$$

Lembrando que se $p \geq 1$, $log(p) \geq 0$ e se $p < 1$, $log(p) < 0$, ou seja, o logaritmo (expoente) de base positiva onde o número é menor que 1 (0 vírgula alguma coisa) é sempre negativo.

Assim, quando $\lambda(x) > 1$, $-ln(\lambda(x)) < 0$ e neste caso está mais próximo de μ_1 do que μ_2, ou seja, $(x-\mu_1)^2 < (x-\mu_2)^2$. Quando $\lambda(x) < 1$, $-ln(\lambda(x)) > 0$ e neste caso está mais próximo de μ_2 do que μ_1.

A regra de classificação pode ser escrita em função $\lambda(x)$ ou em função de $-ln(\lambda(x))$. Estas funções são denominadas funções discriminantes.



Populações com variâncias diferentes

Nem sempre as populações tem a mesma variância, veja o gráfico da **Figura 2**. Temos que analisar em que as variabilidades diferem. Sejam σ_1^2 e σ_2^2 as variâncias das populações 1 e 2. Logo a razão entre as duas funções será:

$$\lambda(x) = \frac{\frac{1}{\sqrt{2\pi}\sigma_1} exp\left\{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right\}}{\frac{1}{\sqrt{2\pi}\sigma_2} exp\left\{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right\}} = \quad (3)$$

$$= \frac{\sigma_2}{\sigma_1} exp\left\{\frac{1}{2}\left[\left(\frac{x-\mu_2}{\sigma_2}\right)^2 - \left(\frac{x-\mu_1}{\sigma_1}\right)^2\right]\right\}$$

Se for considerado $-2ln(\lambda(x))$, temos:

$$-2ln(\lambda(x)) = -2ln\left(\frac{\sigma_2}{\sigma 1}\right) + \left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_2}{\sigma_2}\right)^2\right] \quad (4)$$

Assim novamente, se $-2ln(\lambda(x)) < 0$, x é da população 1, e se $-2ln(\lambda(x)) = 0$ pode pertencer a qualquer das duas populações.

E no caso de populações com mais de uma variável?

No caso de p **variáveis** (ou p campos) podemos estender o conceito. Suponha que para a população 1, o **vetor** $\vec{X}$ seja normal com **vetor de médias** $\vec{\mu_1}$ e **matriz de covariâncias** Σ_1, e que para a população 1 sejam o vetor de médias $\vec{\mu_2}$ e a matriz de covariâncias Σ_2, onde:

$$\vec{X} = \begin{bmatrix} -\vec{X}_1- \\ -\vec{X}_2- \\ \vdots \\ -\vec{X}_p- \end{bmatrix}$$

.

Perceba que o vetor $\vec{X}$ é um vetor de vetores, ou melhor, uma matriz com p linhas e n colunas. A linha i corresponde a i-sima variável, e as instâncias estão nas colunas. Se fossemos transformar $\vec{X}$ em tabela de banco de dados, temos transpor $\vec{X}$. Repetindo, em um banco de dados, cada coluna corresponde a uma variável, e cada linha é uma instância das variáveis, em operações matemáticas, pensa-se de modo transposto.

Um vetor de observações $\vec{x}$ fixo, ou uma instância, seria:

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

No exemplo anterior, supondo que o candidato tivesse as notas de português, matemática e física, ou seja $p = 3$, $\vec{x}$ teria 3 elementos. A razão entre as funções de densidade será:

$$-2ln(\lambda(x)) = -2ln\left\{\frac{(2\pi)^{\frac{p}{2}}(|\Sigma_1|^{\frac{1}{2}})^{-1}}{(2\pi)^{\frac{p}{2}}(|\Sigma_1|^{\frac{1}{2}})^{-1}}\left[\frac{exp\left\{-\frac{1}{2}(x-\mu_1)^{'}\Sigma_1^{-1}(x-\mu_1)\right\}}{exp\left\{-\frac{1}{2}(x-\mu_2)^{'}\Sigma_2^{-1}(x-\mu_2)\right\}}\right]\right\} = \tag{5}$$

$$= \left[(x-\mu_1)^{'}\Sigma_1^{-1}(x-\mu_1)\right] - \left[(x-\mu_2)^{'}\Sigma_2^{-1}(x-\mu_2)\right] + [ln|\Sigma_1| - ln|\Sigma_2|]$$