

# Resumo de Aprendizagem de Máquina 2014-2

Eduardo M. B. de A. Tenório

embat@cin.ufpe.br

CIn-UFPE

## Resumo

*Este documento tem por finalidade ser um resumo dos assuntos abordados na disciplina Aprendizagem de Máquina do período 2014-2 do CIn-UFPE, ministrada pelos professores Francisco Carvalho e Teresa Ludermit. A maioria do documento referencia o livro “Pattern Classification”, de Duda, Hart e Stork. Os códigos utilizados como exercício de fixação encontram-se em [github.com/embatbr/resumo-aprendizagem](https://github.com/embatbr/resumo-aprendizagem).*

## 1 Teoria da Decisão Bayesiana

### 1.1 Introdução

Teoria da Decisão Bayesiana é uma abordagem estatística para a classificação de padrões, baseada em quantificar os tradeoffs associados a tomar uma determinada decisão (classificar) utilizando probabilidade e considerando os custos associados.

O **estado natural** é denotado por  $\omega$ , de modo que  $\omega = \omega_i$ , para  $i = 1, 2, \dots, c$ , significa que o exemplo foi classificado como pertencente à classe  $\omega_i$ . Cada uma dessas classes possui uma **probabilidade a priori**  $P(\omega_i)$ , com

$$\sum_{i=1}^c P(\omega_i) = 1, \quad (1)$$

refletindo o conhecimento prévio da chance de

um elemento da classe  $\omega_i$  aparecer. A **regra de decisão** fica:

$$\text{Decida } \omega_i \text{ para } \max_i P(\omega_i). \quad (2)$$

Neste caso a classe  $\omega_i$  sempre é escolhida e a probabilidade de erro é dada por:

$$P_{err}(\omega_i) = 1 - P(\omega_i). \quad (3)$$

Utilizando uma característica  $x$  que seja contínua e aleatória, sua **densidade de probabilidade estado-condicional** é dada por  $p(x|\omega)$ . Logo, a diferença entre  $p(x|\omega_i)$  e  $p(x|\omega_j)$  descreve a diferença da característica  $x$  entre as populações das classes  $\omega_i$  e  $\omega_j$ .

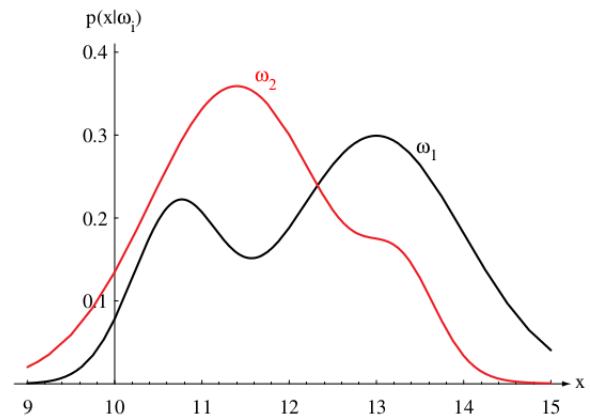


Figura 1: Para  $\omega_i = \omega_2$ , é mais frequente observar  $x$  entre 11 e 12 que  $x = 13$  (valor mais provável se  $\omega_i = \omega_1$ ).

Sabendo  $P(\omega_i)$  e  $p(x|\omega_i)$ , e medindo um valor  $x$ , a probabilidade conjunta de achar

um padrão na classe  $\omega_i$  e com  $x$  é dado por:  
 $p(\omega_i, x) = P(\omega_i|x)p(x) = p(x|\omega_i)P(\omega_i)$ , que  
 pela **fórmula de Bayes** fica:

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)}, \quad (4)$$

com a evidência para  $c$  classes

$$p(x) = \sum_{j=1}^c p(x|\omega_j)P(\omega_j). \quad (5)$$

A probabilidade a posteriori das classes  $\omega_1$  e  $\omega_2$  para um conjunto de valores de  $x$  é mostrada em Fig. (2). A regra de decisão fica:

$$\text{Decida } \omega_i \text{ se } \omega_i \text{ minimiza } P(\text{erro}|x), \quad (6)$$

onde

$$P(\text{erro}|x) = \sum_{j \neq i} P(\omega_j|x), \quad (7)$$

ou simplesmente  $P(\text{erro}|x) = 1 - P(\omega_i|x)$ .  
 Então a regra torna-se:

$$\text{Decida } \omega_i \text{ para } \max_i P(\omega_i|x), \quad (8)$$

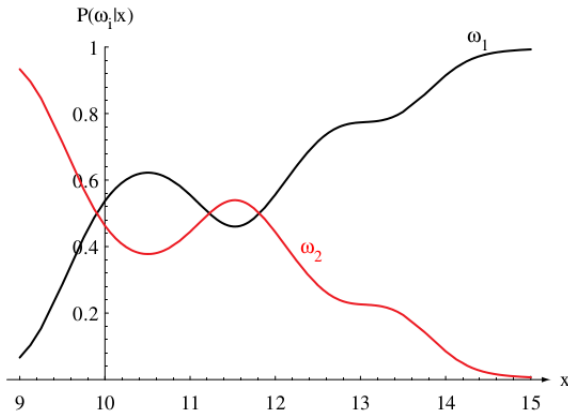


Figura 2: Probabilidades a posteriori para  $P(\omega_1) = \frac{2}{3}$  e  $P(\omega_2) = \frac{1}{3}$ , e para as densidades de probabilidade estado-condicional mostradas em Fig. (1).

Esta regra minimiza a probabilidade média de erro, dada por

$$P(\text{erro}) = \int_{-\infty}^{\infty} P(\text{erro}|x)p(x)dx. \quad (9)$$

## 1.2 Características Contínuas

É de fácil compreensão que a característica  $x$  pode ser trocada por um vetor de características  $\mathbf{x} = (x_1, x_2, \dots, x_d)$ , onde  $\mathbf{x}$  pertence ao espaço  $\mathbf{R}^d$  (espaço de características). A região que decide  $\omega_i$  é denotada por  $\mathcal{R}_i$ .

Outras ações além de apenas classificar um elemento podem ser tomadas, como por exemplo a **rejeição**: recusar-se a tomar uma decisão; uma opção válida quando o custo de ser indeciso é aceitável. Para isso **funções de custo** são inseridas, permitindo tratar de situações onde alguns erros de classificação são mais importantes que outros.

Seja  $\{\omega_1, \dots, \omega_c\}$  o conjunto finito de  $c$  classes e seja  $\{\alpha_1, \dots, \alpha_a\}$  o conjunto finito de possíveis ações. A função de custo  $\lambda(\alpha_i|\omega_j)$  descreve o custo de tomar a ação  $\alpha_i$  quando a classe é  $\omega_j$ . Logo, observado um  $\mathbf{x}$  em particular, tomar a ação  $\alpha_i$  quando a classe é  $\omega_j$  leva a um custo esperado (**risco**)

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^c \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}). \quad (10)$$

$R(\alpha_i|\mathbf{x})$  é chamado de **risco condicional**. Qualquer que seja o  $\mathbf{x}$  observado, o risco pode ser minimizado selecionando a ação que minimiza  $R(\alpha_i|\mathbf{x})$ .

A regra de decisão geral é uma função  $\alpha(\mathbf{x})$  que diz qual ação tomar para cada possível observação, ou seja, para cada  $\mathbf{x}$  a **função de decisão**  $\alpha(\mathbf{x})$  assume um dos  $a$  valores  $\alpha_1, \dots, \alpha_a$ . Logo, o risco global é dado por

$$R = \int R(\alpha(\mathbf{x}))p(\mathbf{x})d\mathbf{x}. \quad (11)$$

O risco global mínimo é chamado de **risco de Bayes**, denotado por  $R^*$ , sendo a melhor performance alcançável.

### 1.3 Classificação por Taxa de Erro Mínima

Para evitar erros, a regra de decisão procurada é aquela que minimiza a probabilidade de erro, i.e. minimiza a **taxa de erro**. A função de custo de interesse para este caso é chamada de **simétrica** ou **zero-um**,

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & \text{se } i = j \\ 1 & \text{se } i \neq j \end{cases} \quad i, j = 1, \dots, c. \quad (12)$$

Como todos os erros tem custo igual, o risco condicional é dado por

$$R(\alpha_i|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x}) \quad (13)$$

com  $P(\omega_i|\mathbf{x})$  sendo a probabilidade condicional da ação  $\alpha_i$  estar correta. A regra de decisão neste caso continua:

$$\text{Decida } \omega_i \text{ para } \max_i P(\omega_i|x). \quad (14)$$

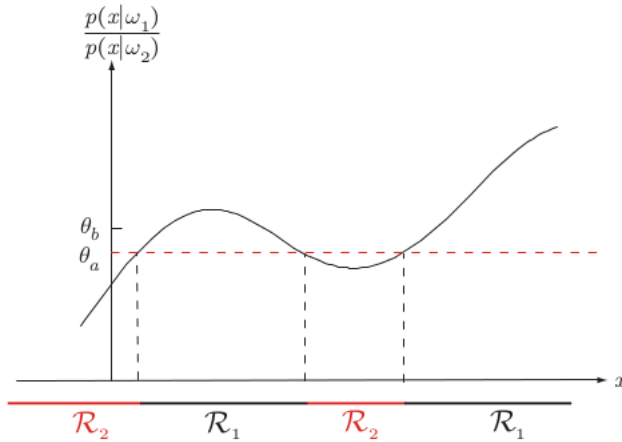


Figura 3: Se a penalização de classificar  $\omega_1$  como  $\omega_2$  for maior que o oposto, então a razão tende ao threshold  $\theta_b$ .

### 1.4 Classificadores, Funções Discriminantes e Superfícies de Decisão

A maneira mais usual de representar classificadores de padrões é através de um con-

junto de **funções discriminantes**  $g_i(\mathbf{x})$ ,  $i = 1, \dots, c$ . O classificador atribui um vetor de características  $\mathbf{x}$  à classe  $\omega_i$  se

$$g_i(\mathbf{x}) = \max_j g_j(\mathbf{x}) \quad (15)$$

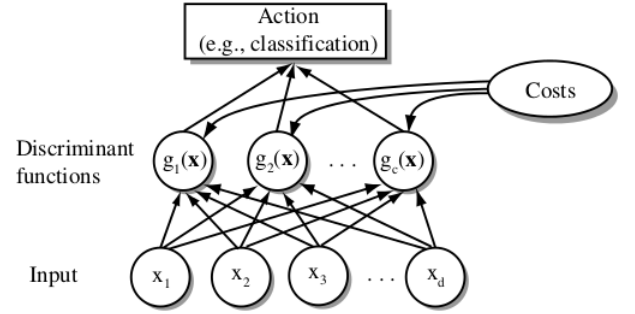


Figura 4: Classificador com  $c$  funções discriminantes e entradas  $d$ -dimensional. A ação geralmente é “escolher o maior  $g_i(\mathbf{x})$ ”.

Para o caso geral com riscos, pode-se fazer  $g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$ , enquanto para o caso “taxa de erro mínima”,  $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x})$ . A função discriminante  $g_i(\mathbf{x})$  pode ser substituída por  $f(g_i(\mathbf{x}))$ , com  $f(\cdot)$  sendo uma função monotonicamente crescente, com o resultado da classificação ficando inalterado.

Para o caso em que  $c = 2$ , o classificador é chamado **dicotomizador**, e apenas uma função discriminante  $g(\mathbf{x}) \equiv g_1(\mathbf{x}) - g_2(\mathbf{x})$  é necessária. Logo a regra de decisão torna-se:

$$\text{Decida } \omega_1 \text{ se } g(\mathbf{x}) > 0; \text{ senão, } \omega_2. \quad (16)$$