Teoria delle Decisioni Bayesiana

Corso di Apprendimento Automatico

Laurea Magistrale in Informatica

Nicola Fanizzi

Dipartimento di Informatica
Università degli Studi di Bari

14 gennaio 2009

Sommario

- Introduzione
- Teoria delle decisioni Bayesiana - nel continuo
- Classificazione a Minimo Tasso d'Errore (Minimum-Error-Rate)
- Classificatori, funzioni discriminanti e superfici di decisione
- Teoria delle decisioni Bayesiana - nel discreto

Introduzione I

Esempio branzino/salmone

- Stato di natura, probabilità a *priori*
  - Lo stato di natura è una variabile aleatoria
  - La pesca di salmone o branzino è *equiprobabile*:

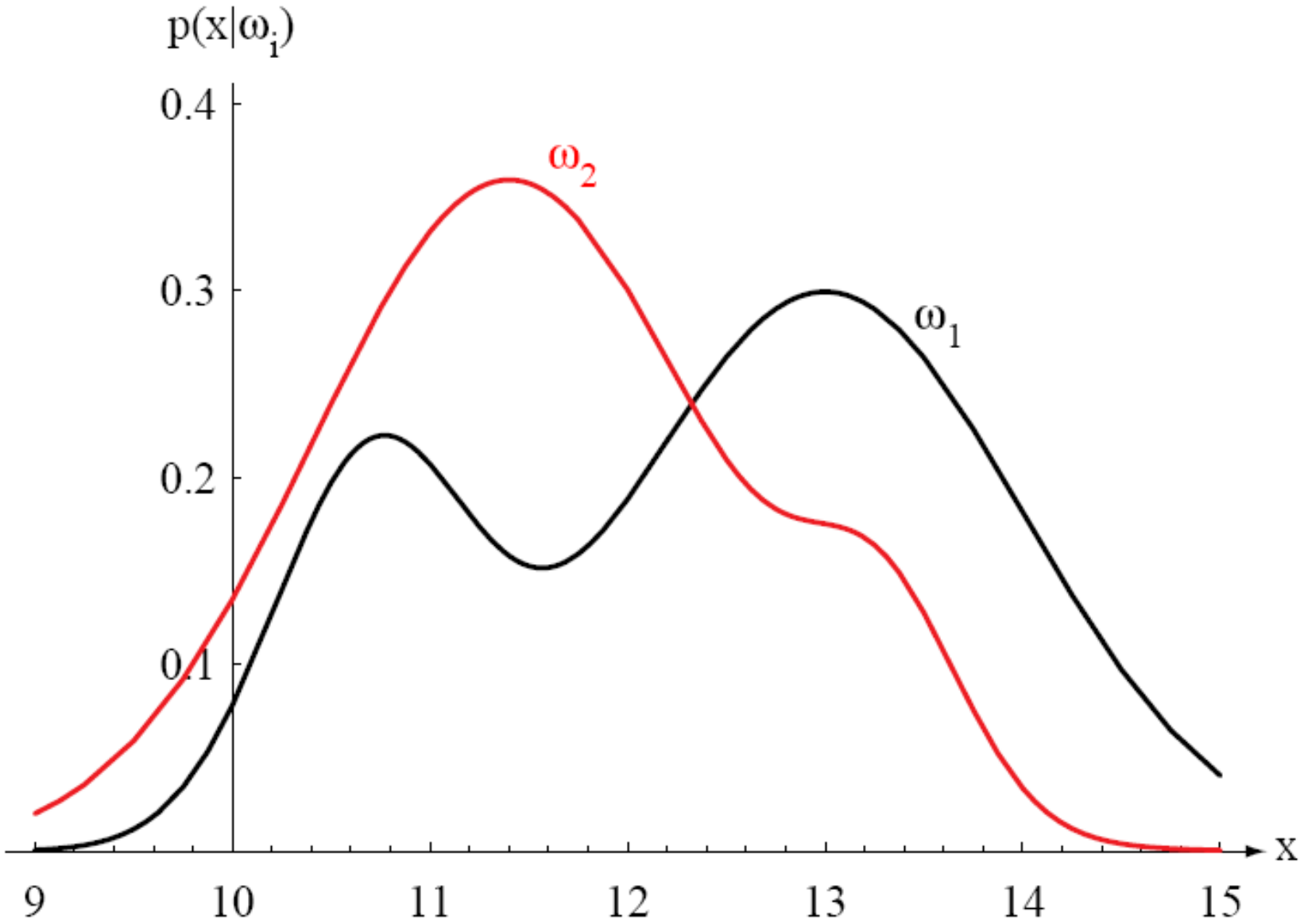
$P(\omega_1) = P(\omega_2)$ probabilità a priori *uniforme*

$P(\omega_1) + P(\omega_2) = 1$ *esclusività* ed *esaustività*

Introduzione II

- Regola di decisione con la sola informazione delle probabilità a priori:
  "Se $P(\omega_1) > P(\omega_2)$ *allora decidi per* ω_1
  altrimenti decidi per ω_2*"*
- Usare l'informazione condizionale sulle classe
  - Sia X una variabile aleatoria che misura il peso
  - $P(x|\omega_1)$ e $P(x|\omega_2)$ descrivono la differente leggerezza tra le due popolazioni di pesci

Introduzione III



Probabilità a posteriori, verosimilianza, evidenza I

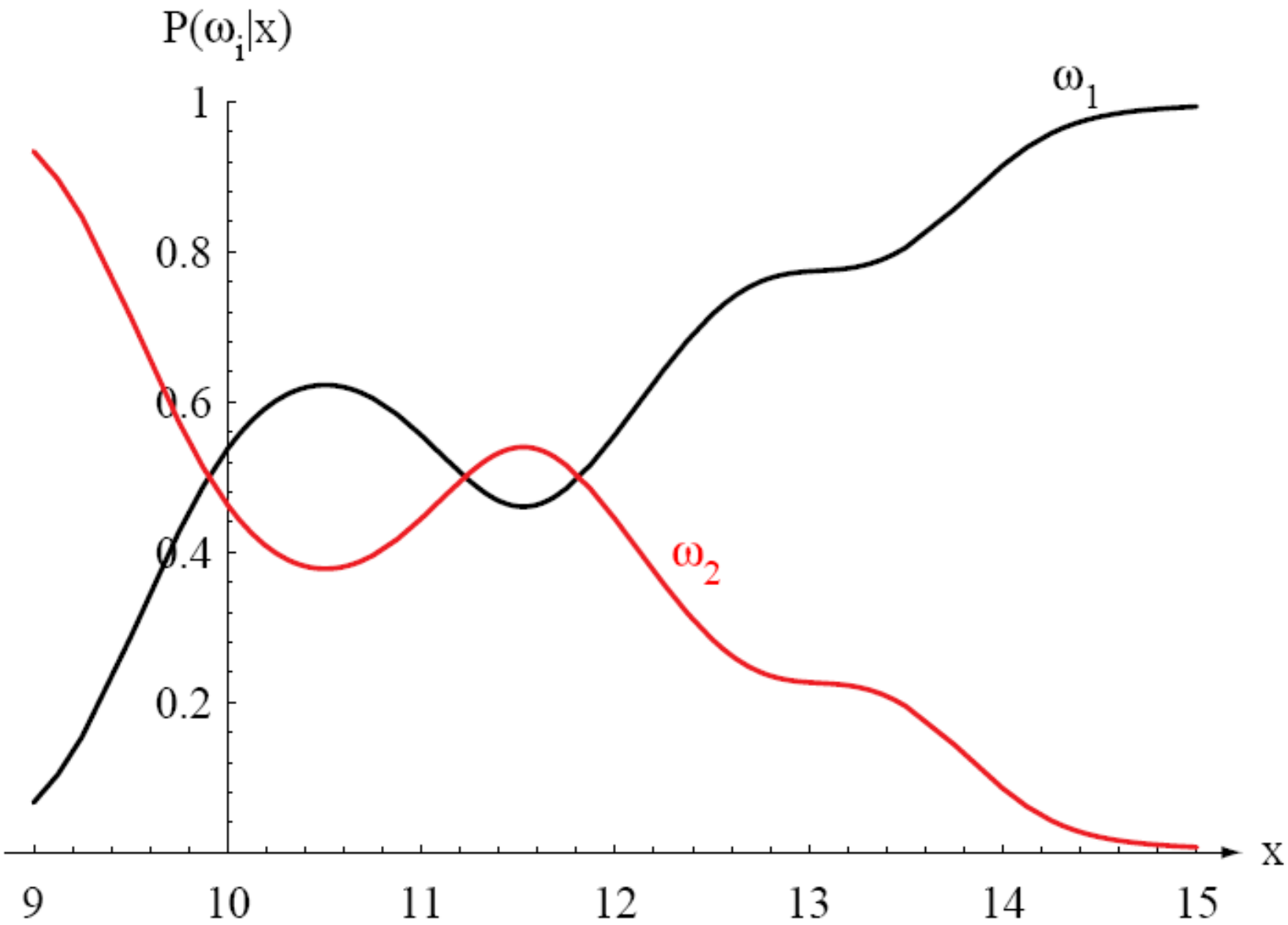
$$\underbrace{P(\omega_j|x)}_{\text{p. a posteriori}} = \frac{\overbrace{P(x|\omega_j)}^{\text{verosimilianza}} \overbrace{P(\omega_j)}^{\text{p. a priori}}}{\underbrace{P(x)}_{\text{evidenza}}}$$

- $P(x)$ meno importante di $P(\omega_j|x)$ e $P(\omega_j)$

In caso di c categorie

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

Probabilità a posteriori, verosimilianza, evidenza II



Errore I

La decisione è conseguenza dalle probabilità a posteriori
X è un'osservazione per la quale:
se $P(\omega_1|x) > P(\omega_2|x) \rightarrow$ stato di natura reale = ω_1
se $P(\omega_1|x) < P(\omega_2|x) \rightarrow$ stato di natura reale = ω_2

Pertanto:
quando si osserva una particolare x,
la probabilità d'errore è:

- $P(error|x) = P(\omega_1|x)$ decidendo per ω_2
- $P(error|x) = P(\omega_2|x)$ decidendo per ω_1

Errore II

Minimizzare la probabilità d'errore

- Se $P(\omega_1|x) > P(\omega_2|x)$ allora decidi per ω_1 altrimenti per ω_2

Vale anche in media:

$$P(errore) = \int_{-\infty}^{-\infty} P(errore, x)dx = \int_{-\infty}^{-\infty} P(errore|x)P(x)dx$$

Pertanto:

$$P(errore|x) = \min\{P(\omega_1|x), P(\omega_2|x)\}$$

(regola di decisione Bayesiana)

Nel caso del continuo I

Generalizzazione delle idee precedenti:

- Usare più d'una *feature*
- Usare più di due *stati* di natura
- Permettere *azioni* non decidere solo per lo stato di natura
  - Permettere altre azioni oltre alla classificazione permette anche la possibilità di rigetto
  - Rifiutare di prendere una decisione in casi difficili o cattivi!
- Introdurre una *loss function* più generale della probabilità d'errore
  - La loss function stabilisce il costo di ogni azione intrapresa

Nel caso del continuo II

- Sia $\{\omega_1, \omega_2, \ldots, \omega_c\}$ l'insieme di c stati di natura ("categorie")
- Sia $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ l'insieme delle azioni possibili
- Sia $\lambda(\alpha_i|\omega_j)$ il costo dell'azione α_i quando lo stato di natura è ω_j

Nel caso del continuo III

Rischio globale

R si ottiene sommando $\underbrace{R(\alpha_i|x)}_{\textit{rischio condizionato}}$ per $i = 1, \ldots, a$

$$R = \int R(\alpha(x)|x)p(x)dx$$

Minimizzare $R \Leftrightarrow$ Minimizzare $R(\alpha_i|x)$ per $i = 1, \ldots, a$

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|x) \qquad i = 1, \ldots, a$$

Selezionare l'azione α_i per la quale $R(\alpha_i|x)$ sia minima
$\rightarrow$ R minimale (*rischio di Bayes*, miglior performance ottenibile)

Classificazione binaria I

- α_1: decidere per ω_1
- α_2: decidere per ω_2

$$\lambda_{ij} = \lambda(\alpha_i|\omega_j)$$

costo della decisione per ω_i quando il vero stato di natura è ω_j

Rischio condizionato:

- $R(\alpha_1|x) = \lambda_{11}P(\omega_1|x) + \lambda_{12}P(\omega_2|x)$
- $R(\alpha_2|x) = \lambda_{21}P(\omega_1|x) + \lambda_{22}P(\omega_2|x)$

Classificazione binaria II

La nostra regola è la seguente:
Se $R(\alpha_1|x) < R(\alpha_2|x)$ allora
si compie l'azione α_1 ossia "decidi per ω_1"

Questo porta alla regola equivalente:
decidi per ω_1 se

$$(\lambda_{21} - \lambda_{11})P(x|\omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22})P(x|\omega_2)P(\omega_2)$$

altrimenti decidi per ω_2

Tasso di verosimiglianza

La regola precedente equivale alla seguente:
Se

$$\frac{P(x|\omega_1)}{P(x|\omega_2)} > \frac{(\lambda_{12} - \lambda_{22})P(\omega_2)}{(\lambda_{21} - \lambda_{11})P(\omega_1)}$$

allora compi l'azione α_1 (decidere per ω_1)
altrimenti compi l'azione α_2 (decidere per ω_2)

$$\frac{P(x|\omega_1)}{P(x|\omega_2)}$$ **likelihood ratio**

Proprietà della decisione ottimale
"Se il grado di verosimiglianza eccede una soglia indipendente dall'esempio di input x, si possono intraprendere azioni ottimali"

Classificazione per minimo tasso d'errore I

- Le azioni sono decisioni sulle classi
  Se α_i viene intrapresa ed il vero stato di natura è ω_j allora:
  la decisione è *corretta* se $i = j$ ed *erronea* se $i \neq j$
- Si cerca una regola di decisione che
  minimizza la probabilità d'errore che è il *tasso d'errore*

Introduzione della *loss function zero-uno*:

$$\lambda(\alpha_i, \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

Classificazione per minimo tasso d'errore II

Perciò, il rischio condizionato è:

$$\begin{aligned} R(\alpha_i|x) &= \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P(\omega_j|x) \\ &= \sum_{j \neq i} P(\omega_j|x) = 1 - P(\omega_i|x) \end{aligned}$$

Il rischio corrispondente a questa loss function è la probabilità d'errore media

- Minimizzare il rischio richiede di massimizzare $P(\omega_i|x)$ (dato che $R(\alpha_i|x) = 1 - P(\omega_i|x)$)
- Per il minimo tasso d'errore:
  Decidere ω_i if $P(\omega_i|x) > P(\omega_j|x) \quad \forall j \neq i$

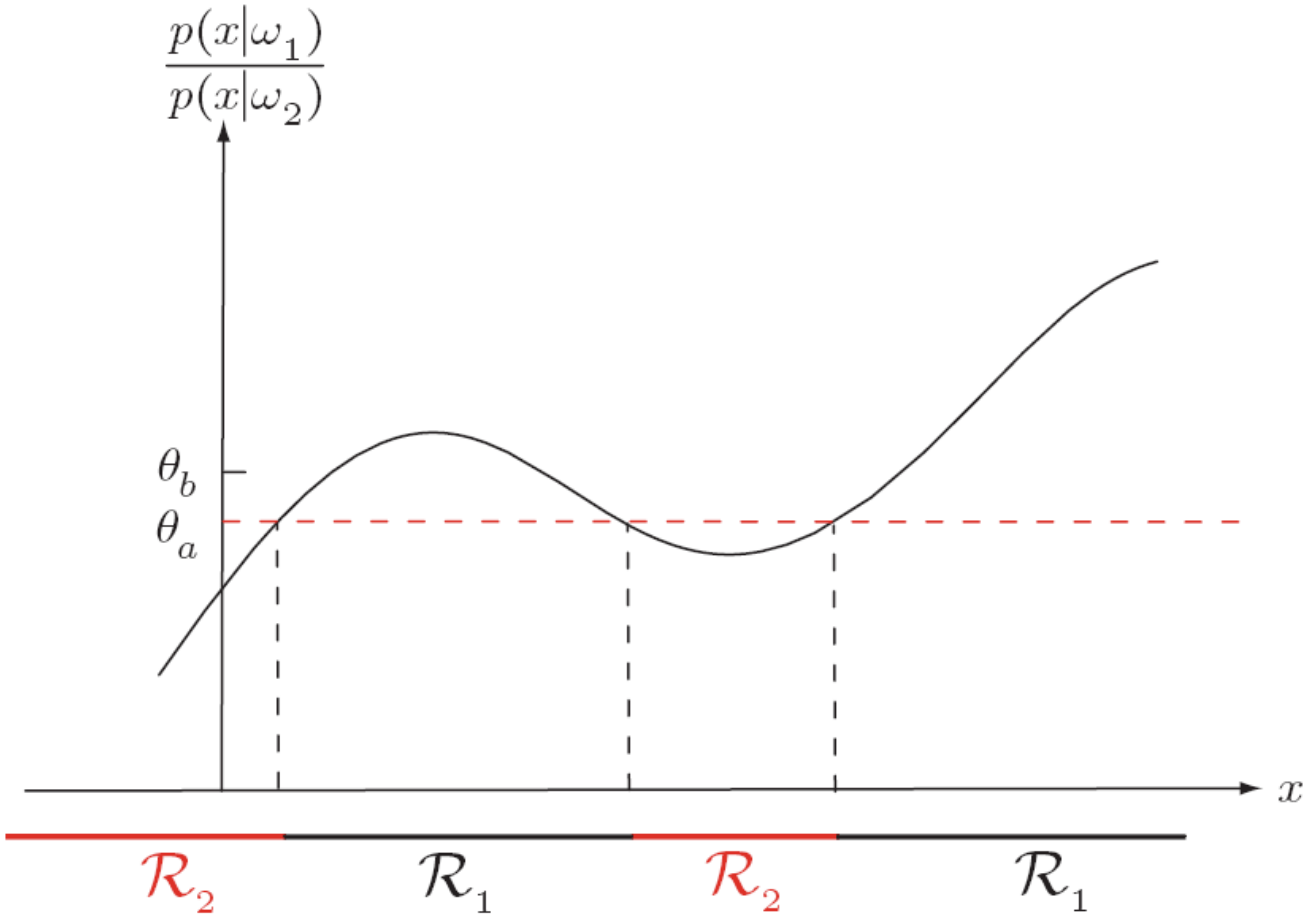
Classificazione per minimo tasso d'errore III

Regioni di decisione e loss function zero-uno

- Pertanto si ha la regola:
  Sia $\frac{(\lambda_{12}-\lambda_{22})P(\omega_2)}{(\lambda_{21}-\lambda_{11})P(\omega_1)} = \theta_\lambda$
  allora decidere per ω_1 se $\frac{P(x|\omega_1)}{P(x|\omega_2)} > \theta_\lambda$
- Se λ è la loss function zero-uno che significa:
  Se $\lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ allora $\theta_\lambda = \frac{P(\omega_2)}{P(\omega_1)} = \theta_a$

  Se $\lambda = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}$ allora $\theta_\lambda = \frac{2P(\omega_2)}{P(\omega_1)} = \theta_b$

Classificazione per minimo tasso d'errore IV



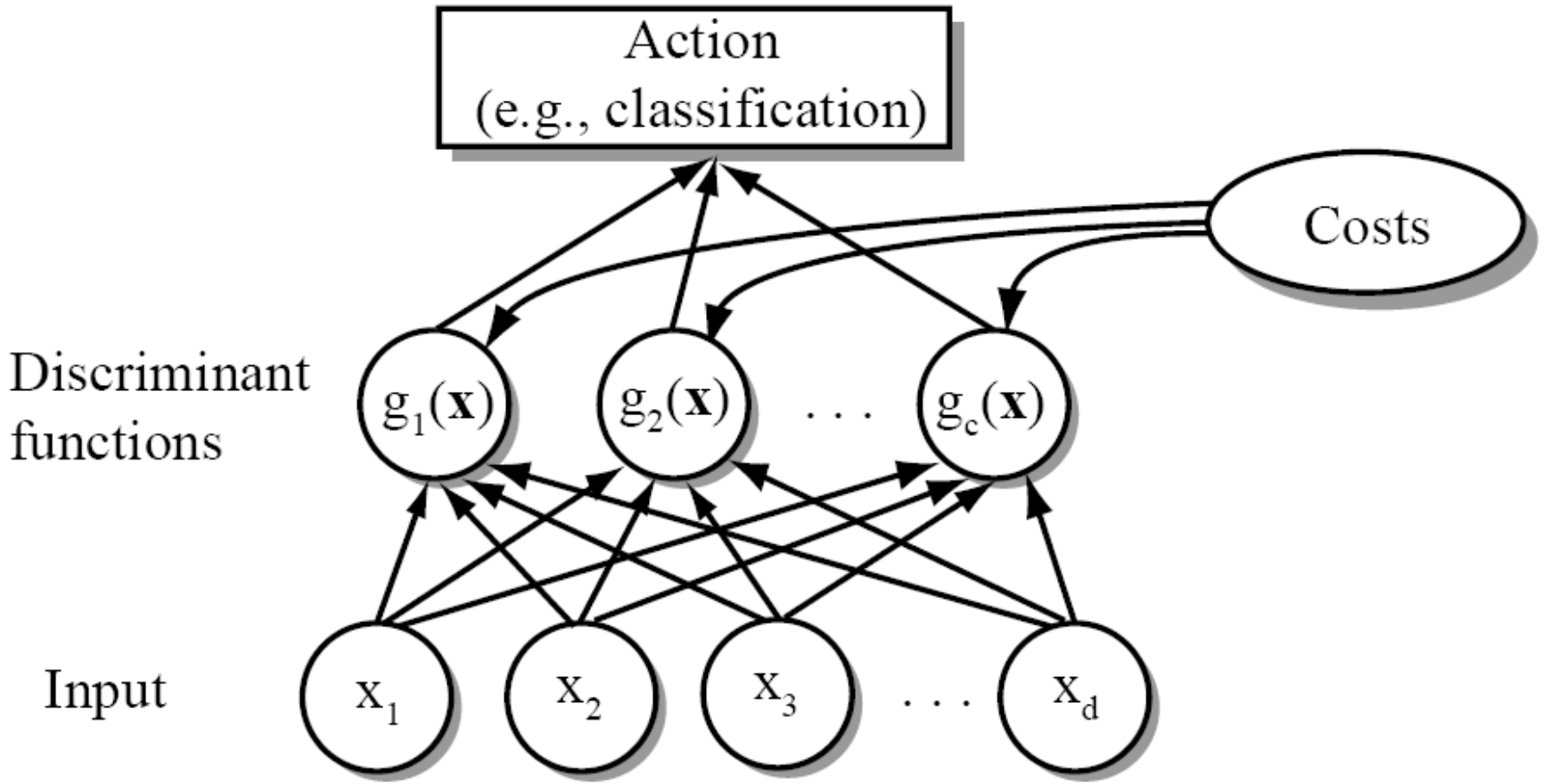
Con una loss function 0/1 o basata sulla classificazione, i limiti di decisione sono determinati da θ_a. Se la loss function penalizza la miscategorizzazione di ω_2, si passa a soglie più ampie θ_b, e $\mathcal{R}_1$ diventa più piccola

Il caso multi-categorico

- Insieme di funzioni discriminanti $g_i(x),\ i = 1, \ldots, c$
- Il classificatore assegna un vettore x alla classe ω_i se:

$$g_i(x) > g_j(x) \qquad \forall j \neq i$$

Struttura funzionale di un classificatore



Un passo successivo determina quale dei valori discriminanti sia il massimo, e assegna la classe di conseguenza

Minimizzazione del rischio

- Sia $g_i(x) = -R(\alpha_i|x)$
  La discriminazione massima corrisponde al minimo rischio!
- Per il minimum error rate, considerare

$$g_i(x) = P(\omega_i|x)$$

La discriminazione massima corrisponde alla massima prob. a posteriori!

$$g_i(x) \equiv P(x|\omega_i)P(\omega_i)$$

ossia

$$g_i(x) = \ln P(x|\omega_i) + \ln P(\omega_i)$$

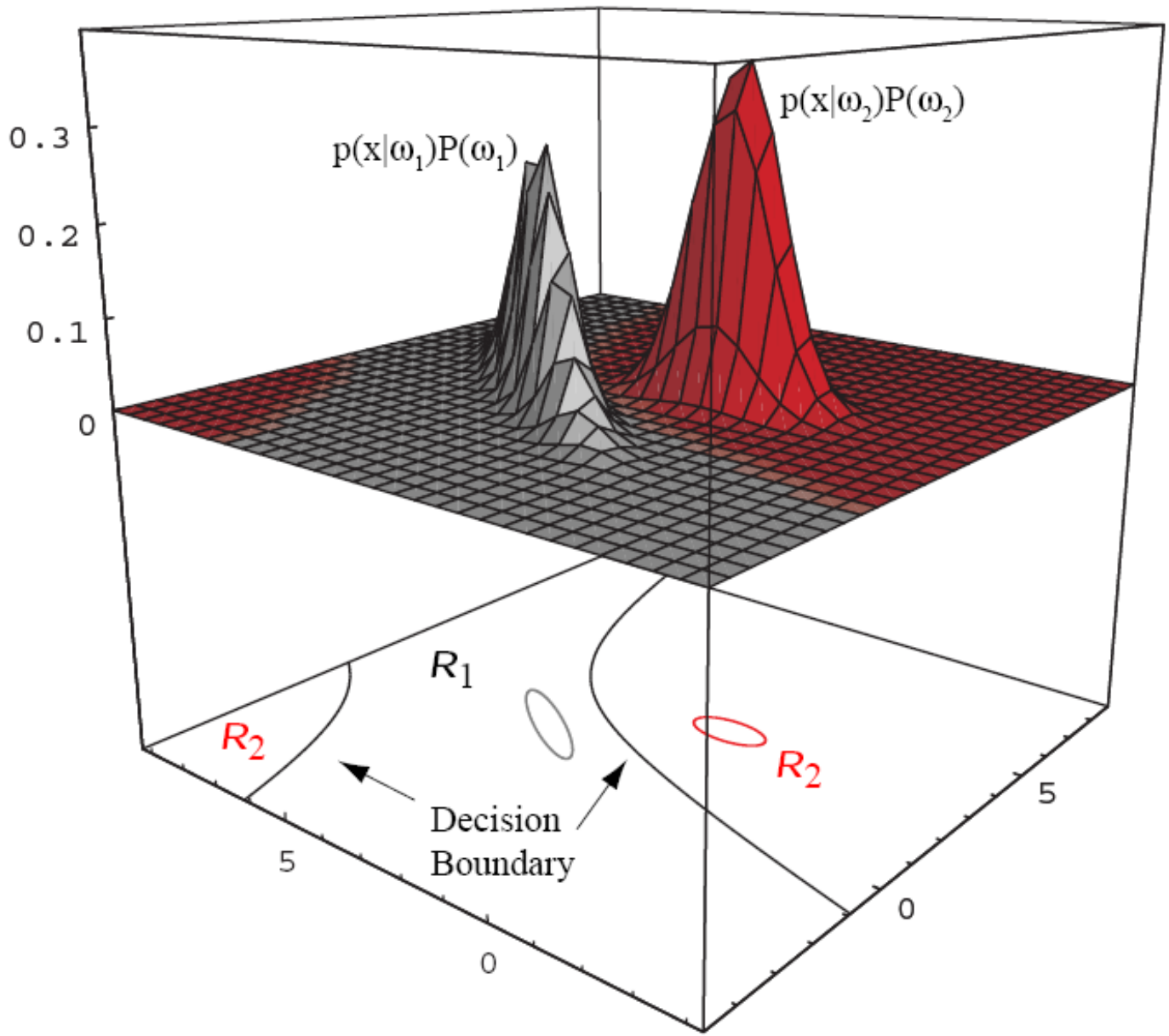
Regioni di decisione I

- Lo spazio delle feature viene diviso in c regioni di decisione
  Se $g_i(x) > g_j(x)\ \forall j \neq i$ allora x è in $\mathcal{R}_i$
  ($\mathcal{R}_i$ significa assignare x a ω_i)
- Caso binario
  - Un classificatore detto *dicotomizzatore* con due funzioni discriminanti g_1 e g_2
  - Sia $g(x) = g_1(x) - g_2(x)$

    Decidere per ω_1 se $g(x) > 0$; altrimenti decidere per ω_2
  - Calcolo di $g(x)$

$$g(x) = P(\omega_1|x) - P(\omega_2|x) = \ln \frac{P(x|\omega_1)}{P(x|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

Regioni di decisione II



Caso discreto I

- Le componenti di x sono a valori binari o interi, x prende solo uno degli m valori discreti

$$v_1, v_2, \ldots, v_m$$

- Caso di features binarie indipendenti nel problema binario
  Sia $x = [x_1, x_2, \ldots, x_d]^t$ dove ogni x_i è 0 o 1, con le probabilità:

$$p_i = P(x_i = 1|\omega_1) \text{ e } q_i = P(x_i = 1|\omega_2)$$

Caso discreto II

La funzione discriminante in tal caso sarà:

$$g(x) = \sum_{i=1}^{d} w_i x_i + w_0$$

dove

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad i = 1, \ldots, d$$

e

$$w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

Decidere ω_1 se $g(x) > 0$ e ω_2 se $g(x) \leq 0$

Credits

- R. Duda, P. Hart, D. Stork: *Pattern Classification*, Wiley