Elementi di teoria bayesiana della decisione

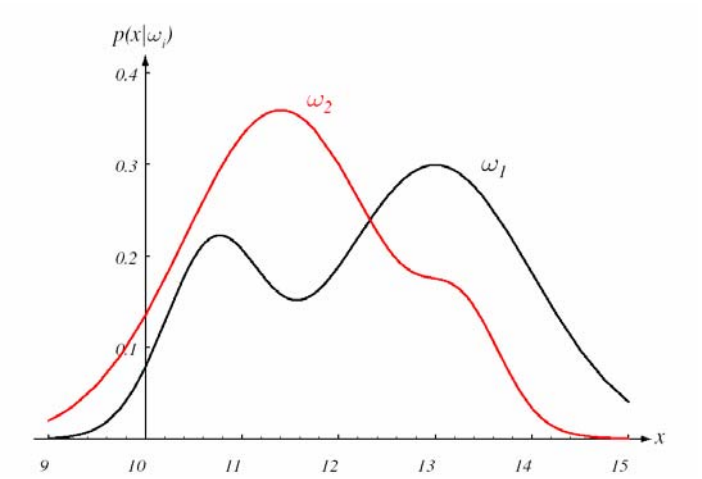
Teoria bayesiana della decisione: caratteristiche

- La teoria bayesiana della decisione è un approccio statistico fondamentale al problema del pattern recognition.
- Il suo obiettivo è quello di confrontare quantitativamente diverse decisioni di classificazione utilizzando le probabilità ed i costi cha accompagnano tali decisioni.
- Assunzioni fondamentali:
  - il problema della decisione è posto in termini probabilistici
  - sono noti i valori di tutte le probabilità rilevanti per il problema

Fondamenti

- Consideriamo un problema a C classi, con etichette ω_j con j=1,2,…,C.
- Etichettiamo con α_i i=1,2,…,*a* le decisioni che è possibile prendere.
- Supponiamo di conoscere la probabilità $P(\omega_j)$ che un campione appartenga ad una certa classe (*probabilità a priori*).
- Conosciamo inoltre la *funzione di costo* (*loss function*) $\lambda(\alpha_i|\ \omega_j)$ che descrive il costo indotto dall'aver preso la decisione α_i quando il campione appartiene alla classe ω_j.

- Se non avessimo altre informazioni, la regola di decisione sarebbe basata interamente sulle $P(\omega_j)$.
- Supponiamo, invece, di poter utilizzare un feature vector N-dimensionale x che, in questo ambito, è formalizzabile come una variabile aleatoria N-dimensionale.
- Conosciamo inoltre la funzione di densità di probabilità condizionata alla classe $p(x|\ \omega_j)$.



Un esempio di densità di probabilità condizionate alle classi con C=2.

Teorema di Bayes

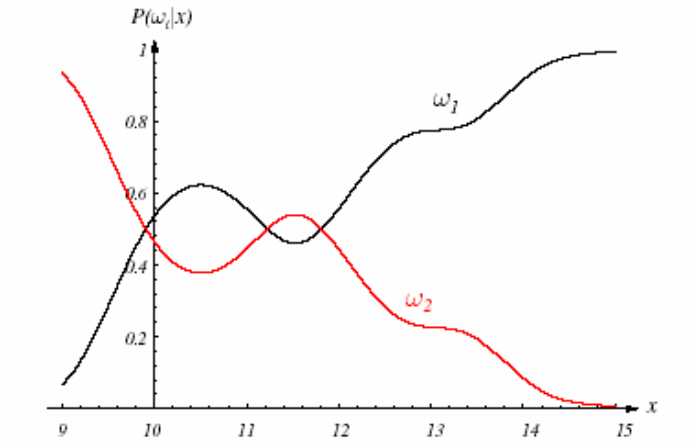
- A partire dalle conoscenze descritte, vorremmo stabilire quale sia la probabilità $P(\omega_j|x)$ (*probabilità a posteriori*)che il campione descritto da un feature vector x appartenga alla classe ω_j.
- E' possibile ottenere questa informazione grazie al teorema di Bayes per cui:

$$P(\omega_j|x) = \frac{p(x|\omega_j)\cdot P(\omega_j)}{p(x)} \quad \text{dove} \quad p(x) = \sum_{j=1}^{C} p(x|\omega_j)\cdot P(\omega_j)$$

Teorema di Bayes

Grazie al teorema di Bayes, è possibile risalire alla probabilità che il feature vector osservato x sia stato prodotto da un campione appartenente alla classe ω_j (prob. a posteriori) a partire dalla probabilità a priori $P(\omega_j)$ e dalle *verosimiglianze* $p(x|\omega_j)$.

Rev. Thomas Bayes
b. 1702, London
d. 1761, Tunbridge Wells, Kent



Le probabilità a posteriori relative alle due classi viste prima, assumendo $P(\omega_1)=2/3$ e $P(\omega_2)=1/3$.

Il costo atteso

- A fronte dell'osservazione di un f.v. x, qual è il costo che dobbiamo aspettarci nel caso prendiamo una decisione α_i ?
- Tale costo (che va sotto il nome di *rischio condizionale*, *conditional risk*) viene valutato come:

$$R(\alpha_i|x) = \sum_{j=1}^{C} \lambda(\alpha_i|\omega_j) \cdot P(\omega_j|x)$$

La regola di decisione

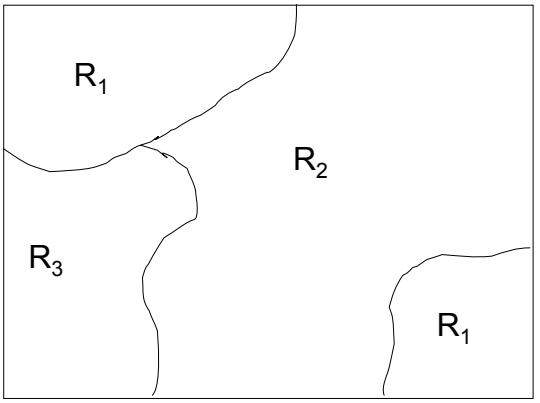
- Una *regola di decisione* è una funzione $\alpha(x)$ che indica quale azione intraprendere per ogni possibile valore di x osservato.
- In questo contesto, la *regola di decisione ottima* è quella per cui si ha il minimo rischio condizionale:

$$\alpha(x) = \underset{1 \le j \le C}{\operatorname{argmin}} R(\alpha_i|x)$$

Regioni di decisione

La regola di decisione induce nello spazio delle fatures un insieme di regioni di decisione.

$x \in R_i \Leftrightarrow \alpha(x) = \alpha_i$



Problemi a due classi

- Nel caso particolare di problemi a due classi, indichiamo con α_i la decisione per la classe ω_i con i=1,2.
- Definiamo $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$. I rischi condizionali sono:

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

Problemi a due classi

- Ovviamente la regola di decisione farà scegliere ω_1 se $R(\alpha_1|x)<R(\alpha_2|x)$.
- La stessa condizione si può porre in modo equivalente in termini di probabilità a posteriori:

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

oppure:

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

Problemi a due classi

- Ricordando il teorema di Bayes, la condizione si può scrivere:

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

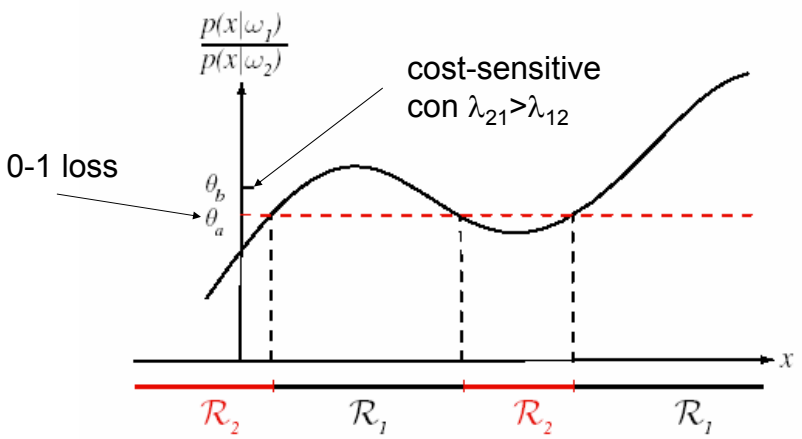
dove il membro di sinistra si definisce *rapporto di verosimiglianza* (*likelihood ratio*)

Problemi a due classi non cost-sensitive

- Nel caso di classificazione non cost-sensitive la decisione deve minimizzare il numero di errori.
- Questo caso ricade nel precedente a patto di porre $\lambda_{21}=\lambda_{12}=1$ e $\lambda_{11}=\lambda_{22}=0$ (*zero-one loss*).
- La condizione diventa quindi:

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)}$$

Problemi a due classi



Probabilità minima di errore

- E' importante valutare quale sia la minima probabilità di errore, che fornisce il miglior risultato raggiungibile.
- Consideriamo il problema a due classi. Indichiamo con X un generico campione e con x il f.v. corrispondente.
- Siano inoltre R_1 e R_2 le due regioni di decisione e $T=R_1 \cup R_2$ il dominio di x.

Probabilità minima di errore Problemi a due classi

- Per una regola di decisione che genera le due regioni di decisione R_1 e R_2 la probabilità di errore è:

$$
\begin{aligned}
P_e &= p(x \in R_2, X \in \omega_1) + p(x \in R_1, X \in \omega_2) = \\
&= p(x \in R_2 \mid \omega_1) P(\omega_1) + p(x \in R_1 \mid \omega_2) P(\omega_2) = \\
&= \int_{R_2} p(x \mid \omega_1) dx \, P(\omega_1) + \int_{R_1} p(x \mid \omega_2) dx \, P(\omega_2) = \\
&= \int_{R_2} p(x \mid \omega_1) P(\omega_1) dx + \int_{R_1} p(x \mid \omega_2) P(\omega_2) dx
\end{aligned}
$$

Probabilità minima di errore Problemi a due classi

- La probabilità di errore è limitata inferiormente:

$$P_e = \int_{R_2} p(x|\omega_1)P(\omega_1)dx + \int_{R_1} p(x|\omega_2)P(\omega_2)dx \geq$$

$$\geq \int_T \min\{p(x|\omega_1)P(\omega_1), p(x|\omega_2)P(\omega_2)\}dx$$

- La probabilità minima di errore è quindi:

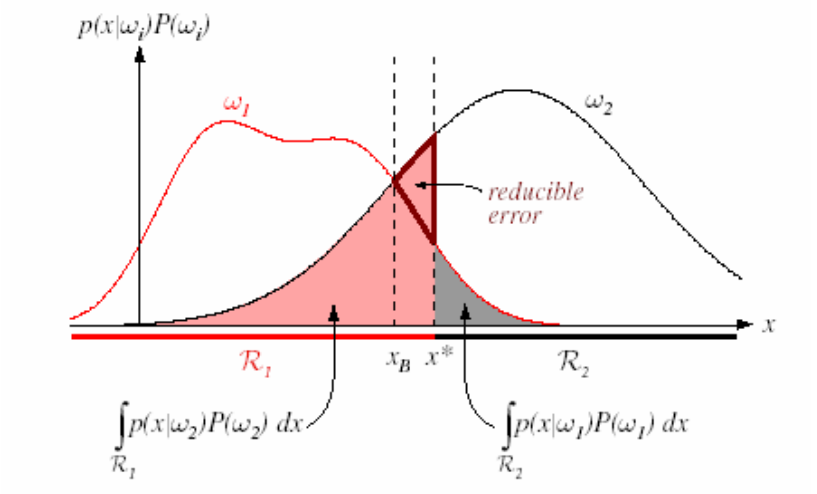
$$P_e^* = \int_T \min\{p(x|\omega_1)P(\omega_1), p(x|\omega_2)P(\omega_2)\}dx$$

Quando viene ottenuta ?

Ottimalità del classificatore bayesiano

- La probabilità di errore minima viene raggiunta con la regola di decisione bayesiana $\alpha(x)$ = argmax{$P(\omega_1|x)$, $P(\omega_2|x)$ }.
- Di conseguenza, nei problemi a due classi, il classificatore costruito con questa regola (classificatore bayesiano) è il classificatore ottimo.

Probabilità minima di errore
Problemi a due classi



Probabilità minima di errore
Problemi multiclasse

- In maniera analoga si può calcolare la minima probabilità di errore per problemi a C classi:

$$P_e = 1 - P(\text{correct}) = 1 - \sum_{i=1}^{C} \int_{R_i} p(x \mid \omega_i) P(\omega_i) dx$$

- Siccome :

$$\sum_{i=1}^{C} \int_{R_i} p(x \mid \omega_i) P(\omega_i) dx \le \int_{T} \max_{1 \le i \le C} \{p(x \mid \omega_i) P(\omega_i)\} dx$$

$$P_e^* = 1 - \int_{T} \max_{1 \le i \le C} \{p(x \mid \omega_i) P(\omega_i)\} dx$$

Ottimalità del classificatore bayesiano

- La probabilità di errore minima viene raggiunta con la regola di decisione bayesiana $\alpha(x) = \text{argmax } \{P(\omega_i|x)\}$.
- Di conseguenza, anche nei problemi multiclasse, il classificatore classificatore bayesiano risulta il classificatore ottimo.

Classificazione con rigetto

- Nella classificazione cost-sensitive, ci possono essere casi in cui il costo di un errore è così elevato che è conveniente astenersi dal fornire una risposta piuttosto che rischiare un errore.
- In questi casi, alle decisioni possibili si aggiunge la “decisione di non decidere”, detta anche *rigetto*.

Regola di decisione con rigetto

- La regola di decisione viene ampliata per considerare il nuovo tipo di decisione (*regola di decisione con rigetto*).
- Le condizioni per le quali viene sospesa la decisione vanno sotto il nome di *regola di rigetto* (*reject rule*).

Classificazione con rigetto

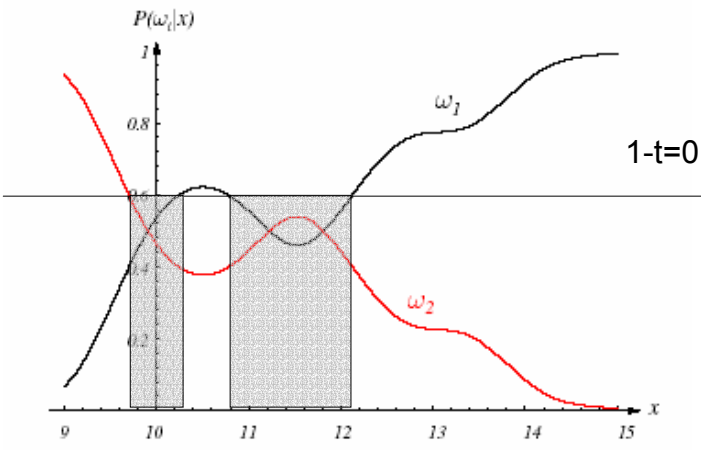
- Per il classificatore bayesiano, la probabilità di errore su un campione x è $P_e(x) = 1-\max\{P(\omega_i|x)\}$.
- Supponiamo di non voler procedere alla classificazione se la P_e supera una soglia t (P_e massima tollerabile).

Regola di decisione con rigetto per il classificatore bayesiano

- La regola di decisione diventa quindi:

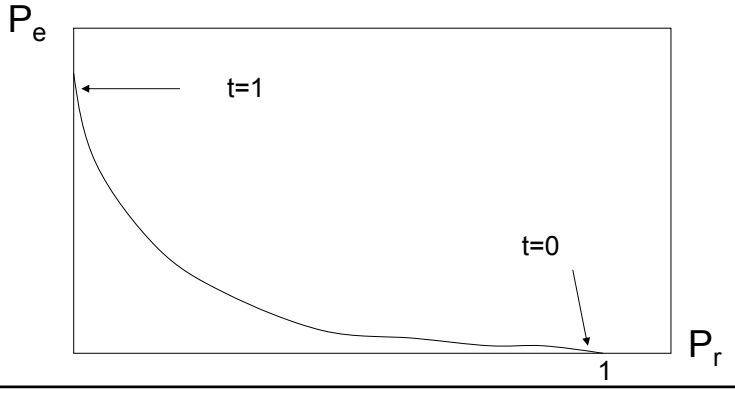
$$\alpha(x) = \begin{cases} \omega_i & \text{se } P(\omega_i|x) > P(\omega_j|x)\ \forall i \neq j \text{ and } \\ & P(\omega_i|x) > 1-t \\ \text{'rigetto'} & \text{altrimenti} \end{cases}$$

Regioni di rigetto



Curva error/reject

- Al variare di t variano la probabilità di errore e la probabilità di rigetto secondo una curva che si definisce curva error/reject



Classificazione con rigetto

- Anche il rigetto avrà un suo costo (inferiore a quello di un errore).
- Assumiamo una funzione di costo del tipo:

$$\lambda_{ij} = \begin{cases} c \text{ se } i=j \\ e \text{ se } i \neq j \\ r \text{ se } i=\text{'rigetto'} \end{cases}$$

Classificazione con rigetto

- Il rischio condizionale diventa:

$$R(\alpha|x) = \begin{cases} r \text{ se } \alpha = \text{'rigetto'} \\ c\,P(\omega_i|x) + e\,(1 - P(\omega_i|x)) \text{ se } \alpha = \omega_i \end{cases}$$

- Quindi la regola di decisione diventa:

$$\alpha(x) = \begin{cases} \omega_i & \text{se } P(\omega_i|x) > P(\omega_j|x)\ \forall i \neq j \text{ and} \\ & P(\omega_i|x) > (e-r)/(e-c) \\ \text{'rigetto'} & \text{altrimenti} \end{cases}$$

Regola di Chow

Funzioni discriminanti

- Un classificatore può essere rappresentato tramite un insieme di *funzioni discriminanti* $g_i(x)$ i=1,…,C.
- Un campione con f.v. viene assegnato alla classe ω_i se $g_i(x) > g_j(x)$ $j \neq i$.
- Un classificatore bayesiano può essere rappresentato in termini di funzioni discriminanti in diversi modi.

Funzioni discriminanti

- Caso generale: $g_i(x) = -R(\alpha_i|x)$
- Non cost-sensitive: $g_i(x) = P(\omega_i|x)$
- In generale, ogni funzione monotona di $P(\omega_i|x)$ può essere usata, es.: $g_i(x) = \ln P(\omega_i|x)$
- Per il classificatore due classi (*dichotomizer*) si definisce un'unica funzione discriminante:

$$g(x) \equiv g_1(x) - g_2(x) \underset{<}{\overset{>}{\underset{\omega_2}{\overset{\omega_1}{}}}} 0$$

$$g(x) = \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} + \ln \frac{\lambda_{21} - \lambda_{11}}{\lambda_{12} - \lambda_{22}} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

Funzioni discriminanti