بازشناسی الگو

درس ۳

طبقه‌بندی مبتنی بر نظریه‌ی تصمیم بیز

Classification Based on Bayes Decision Theory

کاظم فولادی
دانشکده مهندسی برق و کامپیوتر
دانشگاه تهران

http://courses.fouladi.ir/pr

بازشناسی الگو

طبقه‌بندی مبتنی بر نظریه‌ی تصمیم بیز

۱

نظریه‌ی تصمیم بیز

نظریه‌ی تصمیم بیزی

BAYESIAN DECISION THEORY

نظریه‌ی تصمیم بیزی

Bayesian Decision Theory

یک روی‌کرد آماری پایه:

تصمیم‌گیری بر مبنای احتمالات و هزینه‌ها

Prepared by Kazim Fouladi | Spring 2017 | 2nd Edition

کاظم فولادی

نظریه‌ی تصمیم بیزی

BAYESIAN DECISION THEORY

نظریه‌ی تصمیم بیزی

Bayesian Decision Theory

یک روی‌کرد آماری پایه:
تصمیم‌گیری بر مبنای احتمالات و هزینه‌ها

هدف

طراحی یک طبقه‌بندی‌کننده
که یک الگوی مجهول را در **محتمل‌ترین طبقه**، طبقه‌بندی نماید.

فرض اولیه: تمام احتمالات و ساختار آماری مسئله معلوم است.

نظریه‌ی تصمیم بیزی

احتمال پیشین

A PRIORI PROBABILITY (PRIOR)

احتمال پیشین *A Priori Probability*	احتمال مشاهده‌ی هر یک از حالت‌ها (پیش از اینکه واقعاً مشاهده شوند)	$P(w_i)$
	بر اساس دانایی پیشین ما از حالت‌ها تعیین می‌شود.	

❋ مجموع احتمالات پیشین همه‌ی حالت‌ها برابر با یک است.

$$\sum_{i=1}^{c} P(w_i) = 1$$

❋ در بسیاری از کاربردها، احتمالات پیشین حالت‌های مختلف با هم برابر هستند (توزیع یکنواخت)

احتمالات پیشین را می‌توان با فراوانی نسبی حالت‌ها در الگوهای آموزشی تخمین زد: $P(w_i) \approx N_i/N$

نظریه‌ی تصمیم بیزی

تصمیم‌گیری بر اساس احتمالات پیشین

MAKING A DECISION

با فرض اینکه تنها دانایی موجود در مورد مسئله، فقط احتمالات پیشین باشد:

$$\text{Decide} \quad \begin{cases} w_1 & \text{if } P(w_1) > P(w_2) \\ w_2 & \text{otherwise} \end{cases}$$

$$P(w_1) \underset{w_2}{\overset{w_1}{\gtrless}} P(w_2)$$

w_1	$P(w_1) > P(w_2)$
w_2	$P(w_1) < P(w_2)$

$$P(error) = \min\{P(w_1), P(w_2)\}$$

نظریه‌ی تصمیم بیزی

تصمیم‌گیری بر اساس احتمالات پیشین: مشکلات

Making a Decision

با فرض اینکه تنها دانایی موجود در مورد مسئله، فقط احتمالات پیشین باشد:

$$\text{Decide} \quad \begin{cases} w_1 & \text{if } P(w_1) > P(w_2) \\ w_2 & \text{otherwise} \end{cases}$$

- همیشه یک طبقه انتخاب می‌شود (طبقه‌ی تصمیم همیشه معلوم و ثابت است)
- اگر توزیع یکنواخت باشد، تصمیم‌گیری ممکن نیست!

حل مشکل: استفاده از اطلاعات بیشتر

نظریه‌ی تصمیم بیزی

مثال: طبقه‌بندی دو نوع ماهی

BAYESIAN DECISION THEORY

طبقه‌بندی دو نوع ماهی

(w حالت طبیعت: یک متغیر تصادفی)

$w = w_1$ for sea bass

$w = w_2$ for salmon

$P(w_1)$ احتمال پیشین اینکه ماهی بعدی «ماهی خاردار sea bass» باشد.

$P(w_2)$ احتمال پیشین اینکه ماهی بعدی «ماهی قـــزل‌آلا salmon» باشد.

$$P(w_1) + P(w_2) = 1$$

(exclusivity and exhaustivity)

Prepared by Kazim Fouladi | Spring 2017 | 2nd Edition

کاظم فولادی

نظریه‌ی تصمیم بیزی

احتمال پسین

A POSTERIORI PROBABILITY (POSTERIOR)

احتمال پسین *A Posteriori Probability*	احتمال مشاهده‌ی هر یک از حالت‌ها (پس از مشاهده‌ی بردار ویژگی)	$P(w_i \mid \mathbf{x})$

احتمال اینکه الگوی مجهول متعلق به طبقه‌ی متناظر با حالت w_i باشد، در صورتی که بردار ویژگی، مقدار $\mathbf{x}$ را به خود بگیرد.

نظریه‌ی تصمیم بیزی

درست‌نمایی

LIKELIHOOD

درست‌نمایی *Likelihood*	احتمال مشاهده‌ی $\mathbf{x}$ در یک طبقه $p(\mathbf{x} \mid w_i)$
چگالی احتمال شرطی طبقه *Class-Conditional Probability Density*	مشاهده‌ی بردار ویژگی با مقدار $\mathbf{x}$ در یک طبقه w_i چه‌قدر محتمل است؟ (توصیف توزیع بردارهای ویژگی در هر یک از طبقه‌ها)

در حالت بردار ویژگی گسسته، تابع جرم احتمال $P(x|w_i)$ را خواهیم داشت.

نظریه‌ی تصمیم بیزی

احتمال مشاهده

EVIDENCE

احتمال مشاهده *Evidence*	احتمال مشاهده‌ی $\mathbf{x}$	$p(\mathbf{x})$
	برای نرمال‌سازی احتمال (مجموع یک) استفاده می‌شود.	
	در حالت بردار ویژگی گسسته، تابع جرم احتمال $P(\mathrm{x})$ را خواهیم داشت.	

قاعده‌ی بیز

مبنای اصلی بازشناسی آماری الگو

Bayes Rule

$$P(w_j|\mathbf{x}) = \frac{p(\mathbf{x}|w_j)P(w_j)}{p(\mathbf{x})}$$

where $p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x}|w_j)P(w_j)$.

$$posterior = \frac{likelihood \times prior}{evidence}$$

$$P(model \mid data) = \frac{P(data \mid model)P(model)}{P(data)}$$

Prepared by Kazim Fouladi | Spring 2017 | 2nd Edition

کاظم فولادی

MAKING A DECISION

تصمیم‌گیری پس از مشاهده‌ی ویژگی:

$$\text{Decide} \quad \begin{cases} w_1 & \text{if } P(w_1|x) > P(w_2|x) \\ w_2 & \text{otherwise} \end{cases}$$

$$P(w_1 \mid \mathbf{x}) \underset{w_2}{\overset{w_1}{\gtrless}} P(w_2 \mid \mathbf{x})$$

w_1	$P(w_1 \mid \mathbf{x}) > P(w_2 \mid \mathbf{x})$
w_2	$P(w_1 \mid \mathbf{x}) < P(w_2 \mid \mathbf{x})$

نظریه‌ی تصمیم بیزی

تصمیم‌گیری بر اساس احتمالات پسین

MAKING A DECISION

تصمیم‌گیری پس از مشاهده‌ی ویژگی:

$$\text{Decide} \quad \begin{cases} w_1 & \text{if } P(w_1|x) > P(w_2|x) \\ w_2 & \text{otherwise} \end{cases}$$

$$P(w_1 \mid \mathbf{x}) \underset{w_2}{\overset{w_1}{\gtrless}} P(w_2 \mid \mathbf{x})$$

با استفاده از قاعده‌ی بیز:

$$\text{Decide} \quad \begin{cases} w_1 & \text{if } \frac{p(x|w_1)}{p(x|w_2)} > \frac{P(w_2)}{P(w_1)} \\ w_2 & \text{otherwise} \end{cases}$$

نظریه‌ی تصمیم بیزی

تصمیم‌گیری بر اساس احتمالات پسین: حالات خاص

MAKING A DECISION

تصمیم‌گیری پس از مشاهده‌ی ویژگی:

$$\text{Decide} \quad \begin{cases} w_1 & \text{if } P(w_1|x) > P(w_2|x) \\ w_2 & \text{otherwise} \end{cases}$$

$$P(w_1 \mid \mathbf{x}) \underset{w_2}{\overset{w_1}{\gtrless}} P(w_2 \mid \mathbf{x})$$

اگر درست‌نمایی‌ها مساوی بودند،
مبنای تصمیم احتمالات پیشین است:

$$p(\mathbf{x} \mid w_1) = p(\mathbf{x} \mid w_2) \quad \Rightarrow \quad P(w_1) \underset{w_2}{\overset{w_1}{\gtrless}} P(w_2)$$

اگر احتمالات پیشین مساوی بودند،
مبنای تصمیم احتمالات پسین است:

$$P(w_1) = P(w_2) \quad \Rightarrow \quad P(\mathbf{x} \mid w_1) \underset{w_2}{\overset{w_1}{\gtrless}} P(\mathbf{x} \mid w_2)$$

نظریه‌ی تصمیم بیزی

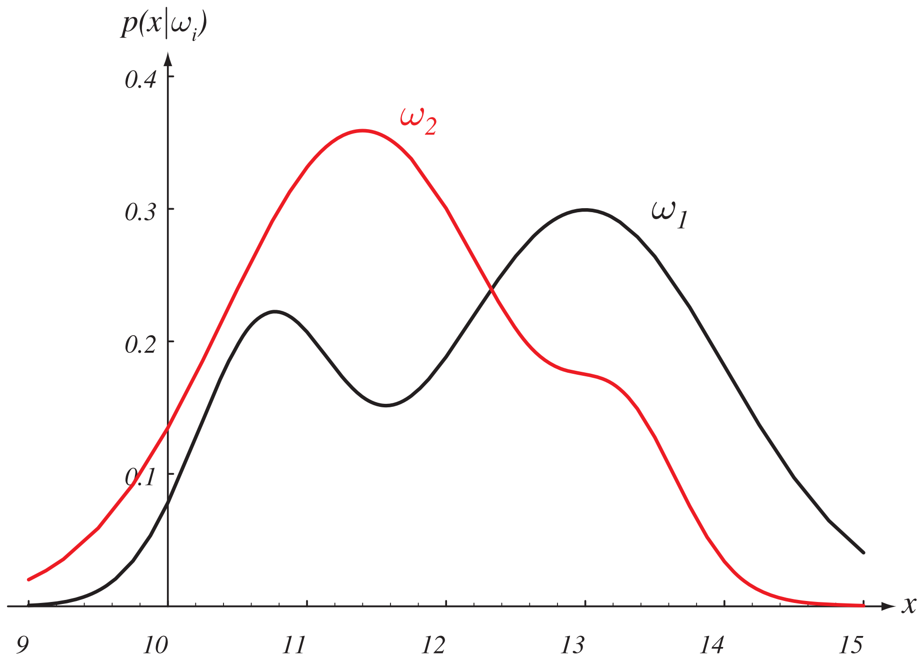
مثال: طبقه‌بندی دو نوع ماهی: توابع چگالی احتمال طبقه‌ها

Bayesian Decision Theory

توابع چگالی احتمال طبقه‌ها (فرضی):
چگالی احتمال اندازه‌گیری یک مقدار خاص از ویژگی x با دانستن اینکه الگو در کلاس ω_i است.

اگر x میزان سبکی (وزن) یک ماهی باشد، این دو منحنی می‌توانند تفاوت وزن دو جمعیت از دو گونه ماهی را توصیف کنند.

توابع چگالی نرمال‌شده هستند و بنابراین مساحت زیر هر منحنی برابر با ۱ است.



From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification, John Wiley & Sons, Inc., 2001.

نظریه‌ی تصمیم بیزی

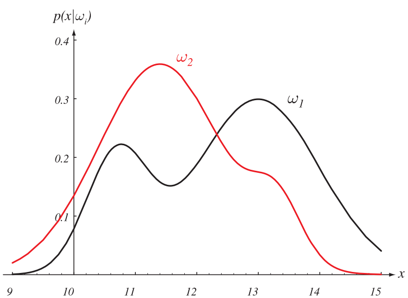
مثال: طبقه‌بندی دو نوع ماهی: توابع احتمال پسین طبقه‌ها

Bayesian Decision Theory

توابع احتمال پسین طبقه‌ها برای احتمالات پیشین

$$P(\omega_1) = \frac{2}{3}, \qquad P(\omega_2) = \frac{1}{3}$$

و توابع چگالی احتمال طبقه‌های زیر:

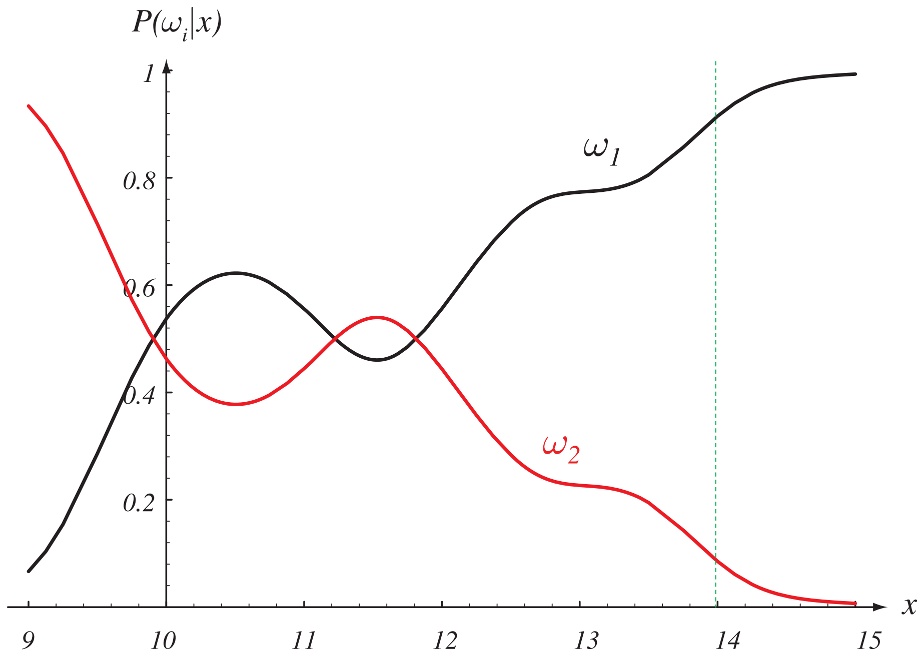


برای این مثال:

اگر مقدار ویژگی اندازه‌گیری شده برای یک الگو $x = 14$ باشد، داریم

$$P(\omega_1|x) = 0.92$$
$$P(\omega_2|x) = 0.08$$

در هر مقدار x مجموع احتمالات پسین ۱ است.



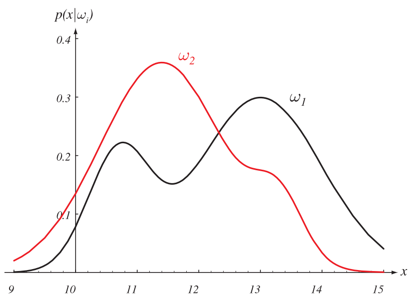
From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification, John Wiley & Sons, Inc., 2001.

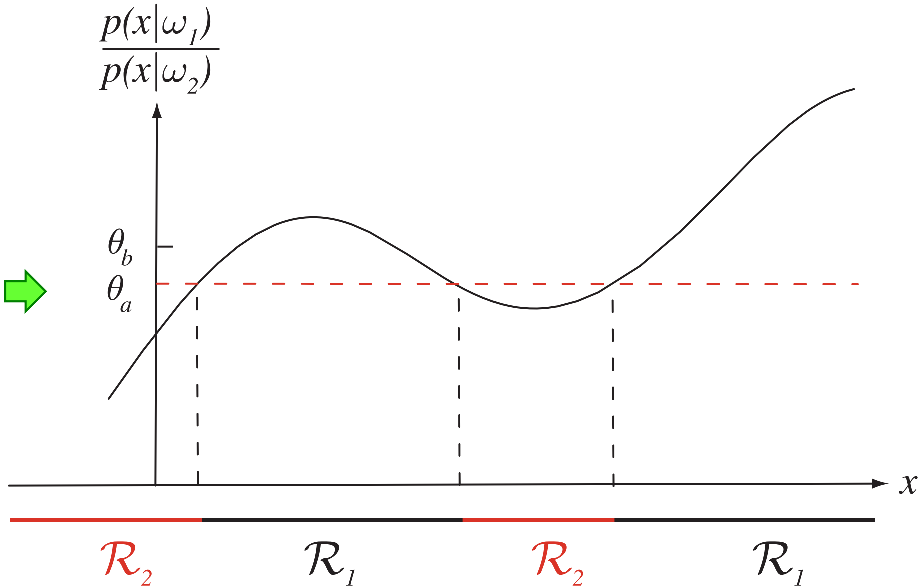
نظریه‌ی تصمیم بیزی

مثال: طبقه‌بندی دو نوع ماهی: نسبت درست‌نمایی

THE LIKELIHOOD RATIO

استفاده از نسبت درست‌نمایی
برای تصمیم‌گیری





From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification, John Wiley & Sons, Inc., 2001.

نظریه‌ی تصمیم بیزی

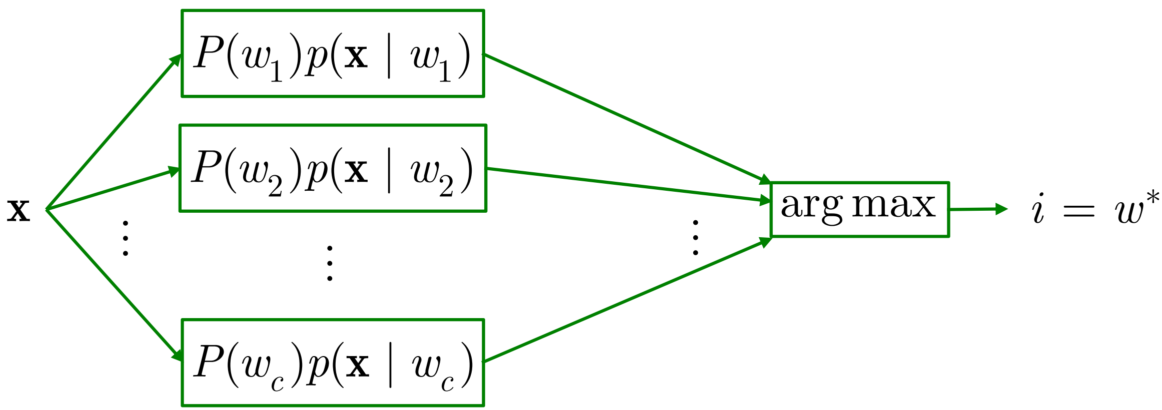
تصمیم‌گیری بر اساس احتمالات پسین (در حالت چند طبقه‌ای)

Making a Decision

برای c طبقه: $w_1,\ w_2,\ \ldots,\ w_c$:

$$\boxed{w^* = \arg\max_i P(w_i \mid \mathbf{x})}$$

$$= \arg\max_i P(w_i)p(\mathbf{x} \mid w_i)$$



PROBABILITY OF ERROR

$$\text{Decide} \quad \begin{cases} w_1 & \text{if } P(w_1|x) > P(w_2|x) \\ w_2 & \text{otherwise} \end{cases}$$

احتمال خطای این تصمیم چیست؟

$$P(error|x) = \begin{cases} P(w_1|x) & \text{if we decide } w_2 \\ P(w_2|x) & \text{if we decide } w_1 \end{cases}$$

احتمال متوسط خطا عبارت است از:

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

قاعده‌ی تصمیم بیز، این خطا را می‌نیمم می‌کند، زیرا:

$$P(error|x) = \min\{P(w_1|x), P(w_2|x)\}$$

قاعده‌ی تصمیم بیزی

احتمالات و انتگرال‌های خطا: احتمال خطا (حالت دو دسته‌ای)

For the two-category case

$$\begin{aligned} P(error) &= P(\mathbf{x} \in \mathcal{R}_2, w_1) + P(\mathbf{x} \in \mathcal{R}_1, w_2) \\ &= P(\mathbf{x} \in \mathcal{R}_2 | w_1) P(w_1) + P(\mathbf{x} \in \mathcal{R}_1 | w_2) P(w_2) \\ &= \int_{\mathcal{R}_2} p(\mathbf{x}|w_1)\, P(w_1)\, d\mathbf{x} + \int_{\mathcal{R}_1} p(\mathbf{x}|w_2)\, P(w_2)\, d\mathbf{x} \end{aligned}$$

From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification, John Wiley & Sons, Inc., 2001.

قاعده‌ی تصمیم بیزی

احتمالات و انتگرال‌های خطا: احتمال خطا (حالت چند دسته‌ای)

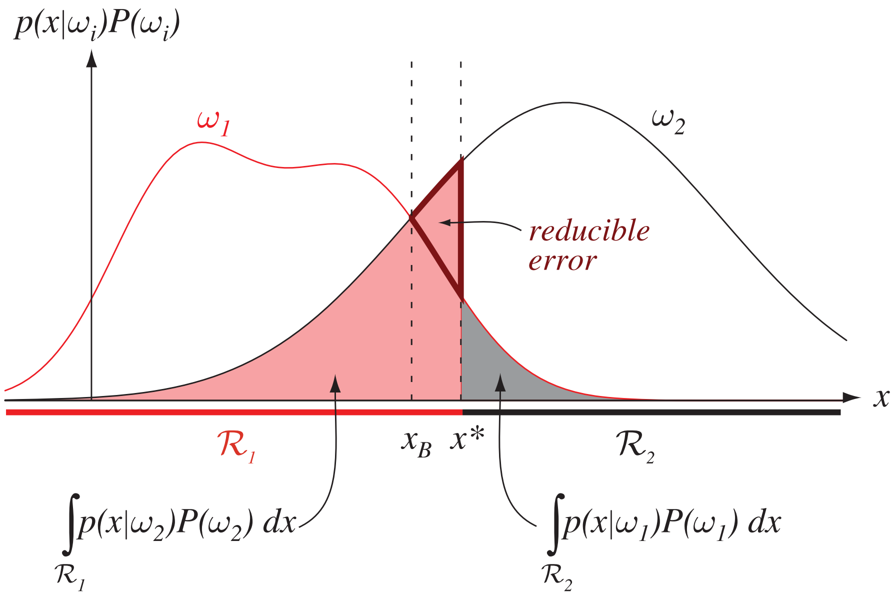
- For the multicategory case

$$\begin{aligned} P(error) &= 1 - P(correct) \\ &= 1 - \sum_{i=1}^{c} P(\mathbf{x} \in \mathcal{R}_i, w_i) \\ &= 1 - \sum_{i=1}^{c} P(\mathbf{x} \in \mathcal{R}_i | w_i) P(w_i) \\ &= 1 - \sum_{i=1}^{c} \int_{\mathcal{R}_i} p(\mathbf{x}|w_i)\, P(w_i)\, d\mathbf{x} \end{aligned}$$

From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification, John Wiley & Sons, Inc., 2001.

قاعده‌ی تصمیم بیزی

تابع چگالی گاوسی چندمتغیره: توابع تفکیک



مؤلفه‌های احتمال خطا
برای احتمالات پیشین مساوی و
نقطه‌ی تصمیم (غیربهینه) x^*:

ناحیه‌ی صورتی رنگ، متناظر با
احتمال خطاها برای انتخاب ω_1 است
وقتی که حالت طبیعت واقعاً ω_2 است.

ناحیه‌ی خاکستری رنگ، متناظر با
احتمال خطاها برای انتخاب ω_2 است
وقتی که حالت طبیعت واقعاً ω_1 است.

اگر مرز تصمیم در نقطه‌ی تساوی دو
احتمال پسین x_B قرار گیرد،
آن‌گاه خطای قابل کاهش حذف می‌شود
و مساحت کل سایه‌دار به می‌نیمم ممکن
می‌رسد:
این تصمیم بیز است
و نرخ خطای می‌نیمم بیز را می‌دهد.

From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification, John Wiley & Sons, Inc., 2001.

CLASSIFIERS BASED ON BAYES DECISION THEORY

- Statistical nature of feature vectors

$$\underline{x} = \left[x_1, x_2, \ldots, x_l\right]^T$$

- Assign the pattern represented by feature vector $\underline{x}$ to the most probable of the available classes

$$\omega_1, \omega_2, \ldots, \omega_M$$

That is $\quad \underline{x} \rightarrow \omega_i : P(\omega_i | \underline{x})$

maximum

- Computation of a-posteriori probabilities
  - Assume known
    - a-priori probabilities

      $$P(\omega_1), P(\omega_2)..., P(\omega_M)$$

    - $p(\underline{x}|\omega_i), i = 1,2..., M$

This is also known as the likelihood of

$$\underline{x} \text{ w.r. to } \omega_i.$$

- The Bayes rule ($M = 2$)

$$p(\underline{x})P(\omega_i|\underline{x}) = p(\underline{x}|\omega_i)P(\omega_i) \Rightarrow$$

$$P(\omega_i \mid \underline{x}) = \frac{p(\underline{x}|\omega_i)P(\omega_i)}{p(\underline{x})}$$

where

$$p(\underline{x}) = \sum_{i=1}^{2} p(\underline{x}|\omega_i)P(\omega_i)$$

❖ The Bayes classification rule (for two classes $M = 2$)

- Given $\underline{x}$ classify it according to the rule

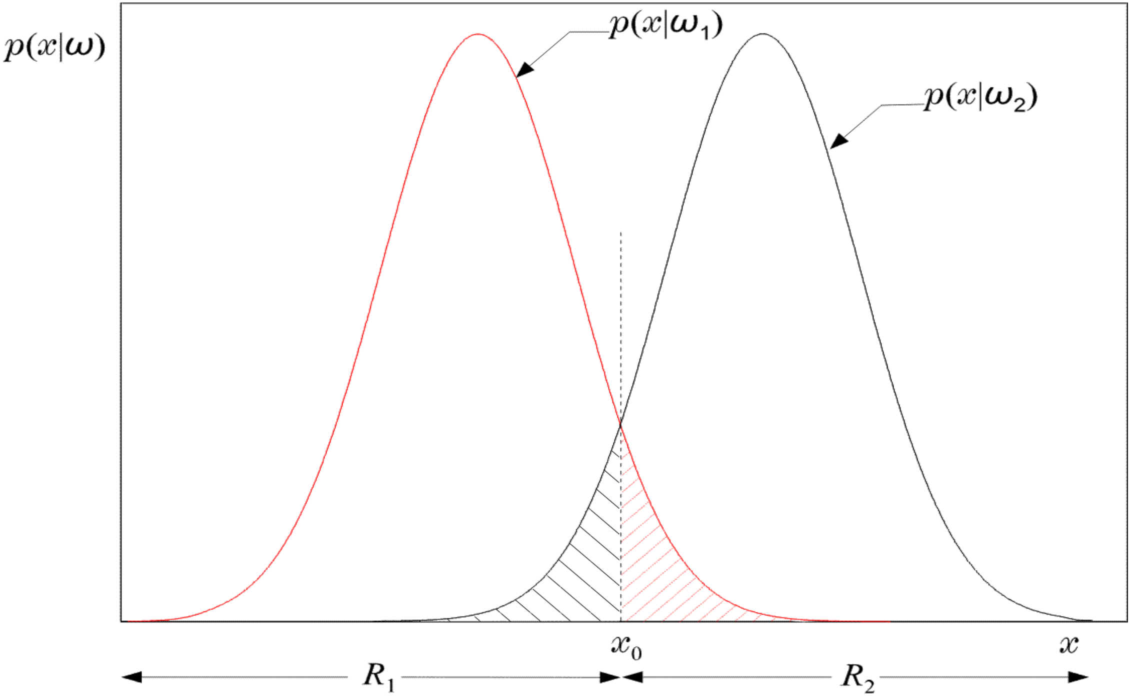
$$\boxed{\begin{array}{l} \text{If } P(\omega_1|\underline{x}) > P(\omega_2|\underline{x}) \;\; \underline{x} \to \omega_1 \\ \text{If } P(\omega_2|\underline{x}) > P(\omega_1|\underline{x}) \;\; \underline{x} \to \omega_2 \end{array}}$$

- Equivalently: classify $\underline{x}$ according to the rule

$$\boxed{p(\underline{x}|\omega_1)P(\omega_1) \gtrless p(\underline{x}|\omega_2)P(\omega_2)}$$

- For equiprobable classes the test becomes

$$\boxed{p(\underline{x}|\omega_1) \gtrless P(\underline{x}|\omega_2)}$$



$$R_1(\to \omega_1) \text{ and } R_2(\to \omega_2)$$

❖ Equivalently in words: Divide space in two regions

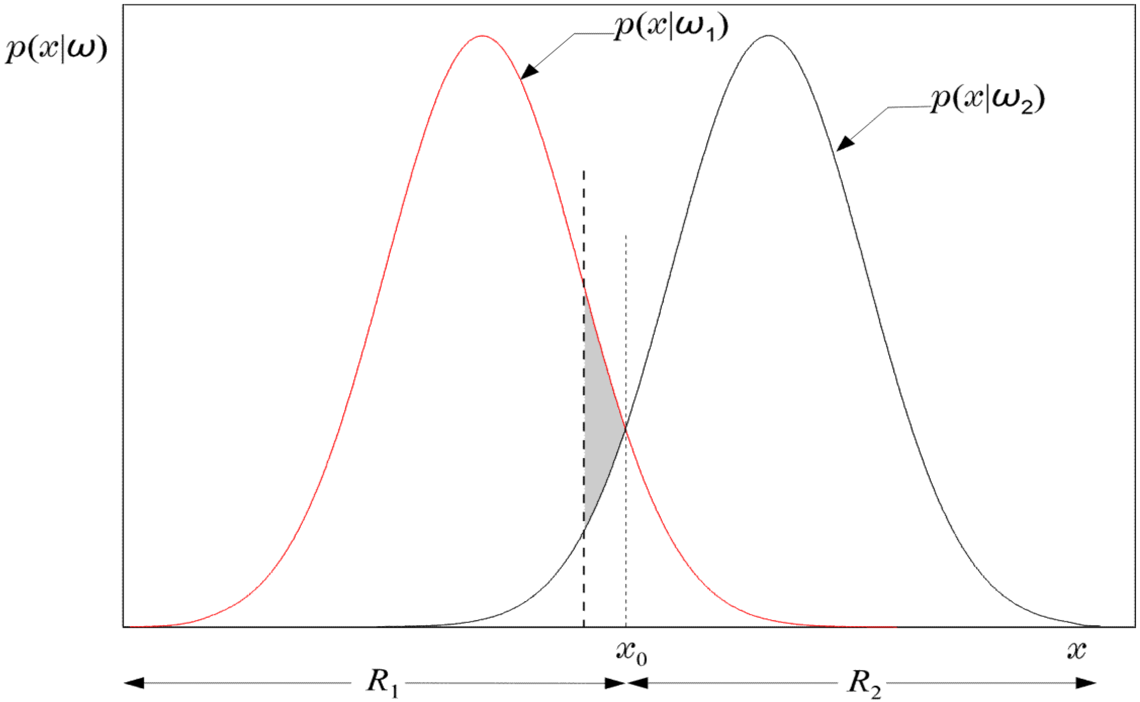
$$\boxed{\begin{array}{l} \text{If } \underline{x} \in R_1 \Rightarrow \underline{x} \text{ in } \omega_1 \\ \text{If } \underline{x} \in R_2 \Rightarrow \underline{x} \text{ in } \omega_2 \end{array}}$$

❖ **Probability of error**

- Total shaded area

$$P_e = \int_{-\infty}^{x_0} p(x|\omega_2)dx + \int_{x_0}^{+\infty} p(x|\omega_1)dx$$

❖ Bayesian classifier is OPTIMAL with respect to minimizing the classification error probability!!!!



➢ **Indeed**: Moving the threshold the total shaded area increases by the extra "grey" area.

❖ The Bayes classification rule for many ($M > 2$) classes:

- Given $\underline{x}$ classify it to ω_i if:

$$\boxed{P(\omega_i | \underline{x}) > P(\omega_j | \underline{x}) \quad \forall j \neq i}$$

- Such a choice also minimizes the classification error probability

تابع زیان

Loss Function

از تابع زیان برای بیان هزینه‌ی هر تصمیم/کنش استفاده می‌کنیم:

c طبقه داریم: $\{w_1, \ldots, w_c\}$

a کنش ممکن: $\{\alpha_1, \ldots, \alpha_a\}$

تابع زیان
Loss Function

بیانگر میزان زیان وارده در اثر انتخاب کنش α_i به‌ازای طبقه‌ی ω_j

$$\lambda(\alpha_i|w_j)$$

ماتریس زیان
Loss Matrix

$$\mathbf{\Lambda} = [\lambda(\alpha_i \mid w_j)]_{a\times c}$$

نظریه‌ی تصمیم بیزی

تابع زیان: مثال (تابع زیان دودویی / صفر و یک)

Loss Function

انتخاب درست طبقه: هزینه‌ی صفر
انتخاب نادرست طبقه: هزینه‌ی یک

$$\lambda(\alpha_i|w_j) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases} \qquad i, j = 1, \ldots, c$$

(همه‌ی خطاها هزینه‌ی یکسان و واحد دارند.)

نظریه‌ی تصمیم بیزی

ریسک شرطی

CONDITIONAL RISK

ریسک شرطی، امید ریاضی زیان است
(زیان مورد انتظار):

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|w_j)P(w_j|\mathbf{x})$$

مثال: ریسک شرطی برای تابع زیان دودویی را محاسبه کنید.

نظریه‌ی تصمیم بیزی

ریسک کل

Overall Risk

ریسک کل، امید ریاضی ریسک شرطی است:

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$$

$\alpha(\mathbf{x})$ یک قاعده برای انتساب مقدار مشاهده $\mathbf{x}$ به یکی از تصمیم‌های α است.

برای می‌نیمم کردن ریسک کل، نیاز به قاعده‌ای داریم که $R(\alpha(\mathbf{x})|\mathbf{x})$ را می‌نیمم کند.

نظریه‌ی تصمیم بیزی

قاعده‌ی تصمیم‌گیری بیز برای می‌نیمم‌سازی ریسک کل

BAYES DECISION RULE FOR MINIMIZING THE OVERALL RISK

تصمیم بهینه با معیار «حداقل ریسک کل»:

$$\alpha^* = \arg\min_{\alpha_i} R(\alpha_i \mid \mathbf{x}) = \arg\min_{\alpha_i} \sum_{j=1}^{c} \lambda(\alpha_i \mid w_j) P(w_j \mid \mathbf{x})$$

نظریه‌ی تصمیم بیزی

قاعده‌ی تصمیم‌گیری بیز برای می‌نیمم‌سازی ریسک کل: مثال (طبقه‌بندی در دو دسته) (۱ از ۲)

Bayes Decision Rule for Minimizing The Overall Risk

$$\alpha^* = \arg\min_{\alpha_i} R(\alpha_i | \mathbf{x}) = \arg\min_{\alpha_i} \sum_{j=1}^{c} \lambda(\alpha_i | w_j) P(w_j | \mathbf{x})$$

- Define

  α_1: deciding w_1

  α_2: deciding w_2

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

- Conditional risks can be written as

$$R(\alpha_1|\mathbf{x}) = \lambda_{11} P(w_1|\mathbf{x}) + \lambda_{12} P(w_2|\mathbf{x})$$
$$R(\alpha_2|\mathbf{x}) = \lambda_{21} P(w_1|\mathbf{x}) + \lambda_{22} P(w_2|\mathbf{x})$$

قاعده‌ی تصمیم بیز برای می‌نیمم ریسک کل؟

Prepared by Kazim Fouladi | Spring 2017 | 2nd Edition
کاظم فولادی

نظریه‌ی تصمیم بیزی

قاعده‌ی تصمیم‌گیری بیز برای می‌نیمم‌سازی ریسک کل: مثال (طبقه‌بندی در دو دسته) (۱ از ۲)

Bayes Decision Rule for Minimizing The Overall Risk

قاعده‌ی تصمیم بیز برای می‌نیمم ریسک کل می‌شود:

$$\text{Decide} \begin{cases} w_1 & \text{if } (\lambda_{21}-\lambda_{11})P(w_1|\mathbf{x}) > (\lambda_{12}-\lambda_{22})P(w_2|\mathbf{x}) \\ w_2 & \text{otherwise} \end{cases}$$

که متناظر با تصمیم w_1 است اگر:

$$\frac{p(\mathbf{x}|w_1)}{p(\mathbf{x}|w_2)} > \underbrace{\frac{(\lambda_{12}-\lambda_{22})}{(\lambda_{21}-\lambda_{11})}\,\frac{P(w_2)}{P(w_1)}}_{\theta_\lambda}$$

یعنی: نسبت درست‌نمایی باید با یک مقدار آستانه‌ی ثابت (مستقل از $\mathbf{x}$) مقایسه شود.

Prepared by Kazim Fouladi | Spring 2017 | 2nd Edition

کاظم فولادی

نظریه‌ی تصمیم بیزی

قاعده‌ی تصمیم‌گیری بیز برای می‌نیمم‌سازی ریسک کل: (تابع زیان دودویی $\Leftarrow$ طبقه‌بندی با می‌نیمم نرخ خطا)

Minimum-Error-Rate Classification

اگر تابع زیان دودویی استفاده شود، طبقه‌بندی می‌نیمم ریسک کل به طبقه‌بندی می‌نیمم خطا تبدیل می‌شود:

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

$$
\lambda(\alpha_i|w_j) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases} \qquad i, j = 1, \dots, c
$$

برای می‌نیمم کردن ریسک باید احتمال ماکزیمم شود، پس:

$$
\text{Decide } w_i \text{ if } P(w_i|\mathbf{x}) > P(w_j|\mathbf{x}) \quad \forall j \neq i
$$

که همان قاعده‌ی بیز برای می‌نیمم احتمال خطا است.

نظریه‌ی تصمیم بیزی

معیارهای طبقه‌بندی با می‌نیمم نرخ خطا

MINIMUM (BAYES) ERROR

ریسک کل *Overall Risk*	با هدف می‌نیمم‌سازی ریسک کل
معیار می‌نیماکس *Minimax Criterion*	با هدف می‌نیمم‌سازی ماکزیمم ریسک کل ممکن (با احتمالات پیشین مختلف)
معیار نِی‌من–پیرسون *Neyman-Pearson Criterion*	با هدف می‌نیمم‌سازی ریسک کل در معرض یک قید

نظریه‌ی تصمیم بیزی

خطای می‌نیمم بیز: معیار می‌نیماکس

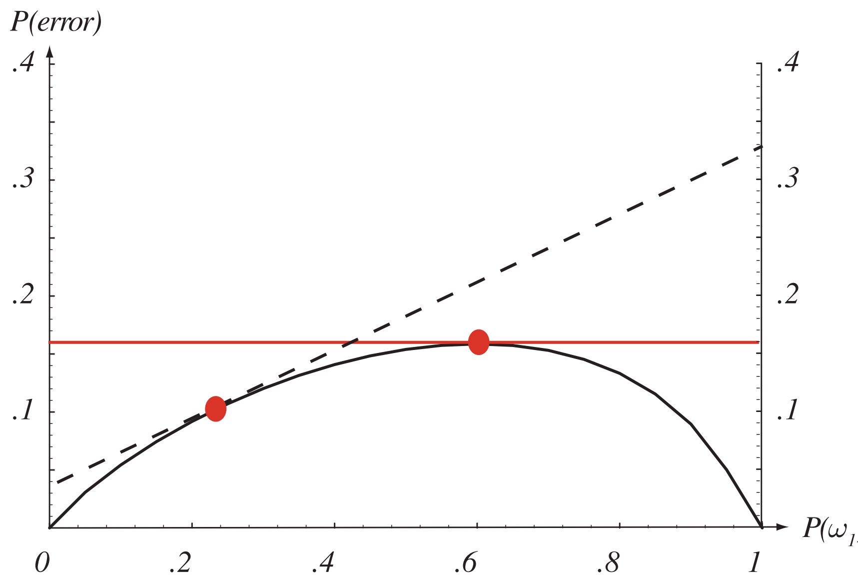
MINIMUM (BAYES) ERROR

برای هر مقدار احتمالات پیشین
مثلاً $P(\omega_1) = 0.25$
یک مرز تصمیم بهینه‌ی متناظر
و یک نرخ خطای بیز وابسته وجود دارد.

برای هر یک از چنین مرزهای (ثابت)
اگر احتمالات پیشین تغییر کند،
احتمال خطا به‌صورت تابعی خطی از
$P(\omega_1)$ تغییر خواهد کرد (خط چین)
و ماکزیمم مقدار این خطا در مقدار
اکستریم احتمال پیشین $P(\omega_1) = 1$
رخ می‌دهد.

برای **می‌نیمم‌سازی ماکزیمم** این خطا،
باید مرز تصمیم را برای خطای ماکزیمم
بیز (مثلاً $P(\omega_1) = 0.6$ در اینجا)،
طراحی کنیم
و بنابراین خطا به‌صورت تابعی از احتمال
پیشین تغییر نمی‌کند (خط ممتد قرمز).

خطای می‌نیمم (بیز)
به‌صورت تابعی از احتمال پیشین $P(\omega_1)$
در یک مسئله‌ی طبقه‌بندی دو دسته‌ای با توزیع‌های ثابت



From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification, John Wiley & Sons, Inc., 2001.

❖ Minimizing the average risk

- For each wrong decision, a penalty term is assigned since some decisions are more sensitive than others

- For $M = 2$
  - Define the loss matrix

$$L = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix}$$

  - λ_{11} penalty term for deciding class ω_1, although the pattern belongs to ω_1, etc.

- Risk with respect to ω_1

$$r_1 = \lambda_{11} \int_{R_1} p(\underline{x}|\omega_1)\, d\underline{x} + \boxed{\lambda_{12} \int_{R_2} p(\underline{x}|\omega_1)\, d\underline{x}}$$

- Risk with respect to ω_2

$$r_2 = \boxed{\lambda_{21} \int_{R_1} p(\underline{x}|\omega_2)\,d\underline{x}} + \lambda_{22} \int_{R_2} p(\underline{x}|\omega_2)\,d\underline{x}$$

$\Rightarrow$ Probabilities of wrong decisions, weighted by the penalty terms

- Average risk

$$r = r_1 P(\omega_1) + r_2 P(\omega_2)$$

❖ Choose R_1 and R_2 so that r is minimized

❖ Then assign $\underline{x}$ to ω_i if

$$\ell_1 \equiv \lambda_{11} p(\underline{x}|\omega_1)P(\omega_1) + \lambda_{21} p(\underline{x}|\omega_2)P(\omega_2) \quad < $$

$$\ell_2 \equiv \lambda_{12} p(\underline{x}|\omega_1)P(\omega_1) + \lambda_{22} p(\underline{x}|\omega_2)P(\omega_2)$$

❖ Equivalently:

assign $\underline{x}$ in ω_1 (ω_2) if

$$\boxed{\ell_{12} \equiv \frac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)} > (<) \frac{P(\omega_2)}{P(\omega_1)} \frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}}}$$

ℓ_{12} : likelihood ratio

❖ If $P(\omega_1) = P(\omega_2) = \dfrac{1}{2}$ and $\lambda_{11} = \lambda_{22} = 0$

$$\boxed{\begin{aligned} &\underline{x} \rightarrow \omega_1 \text{ if } P(\underline{x}|\omega_1) > P(\underline{x}|\omega_2)\frac{\lambda_{21}}{\lambda_{12}} \\ &\underline{x} \rightarrow \omega_2 \text{ if } P(\underline{x}|\omega_2) > P(\underline{x}|\omega_1)\frac{\lambda_{12}}{\lambda_{21}} \end{aligned}}$$

if $\lambda_{21} = \lambda_{12} \Rightarrow$

Minimum classification error probability

❖ An example:

– $p(x|\omega_1) = \dfrac{1}{\sqrt{\pi}} \exp(-x^2)$

– $p(x|\omega_2) = \dfrac{1}{\sqrt{\pi}} \exp(-(x-1)^2)$

– $P(\omega_1) = P(\omega_2) = \dfrac{1}{2}$

– $L = \begin{pmatrix} 0 & 0.5 \\ 1.0 & 0 \end{pmatrix}$

- Then the threshold value is:

x_0 for minimum P_e :

$$x_0 : \exp(-x^2) = \exp(-(x-1)^2) \Rightarrow$$

$$x_0 = \frac{1}{2}$$

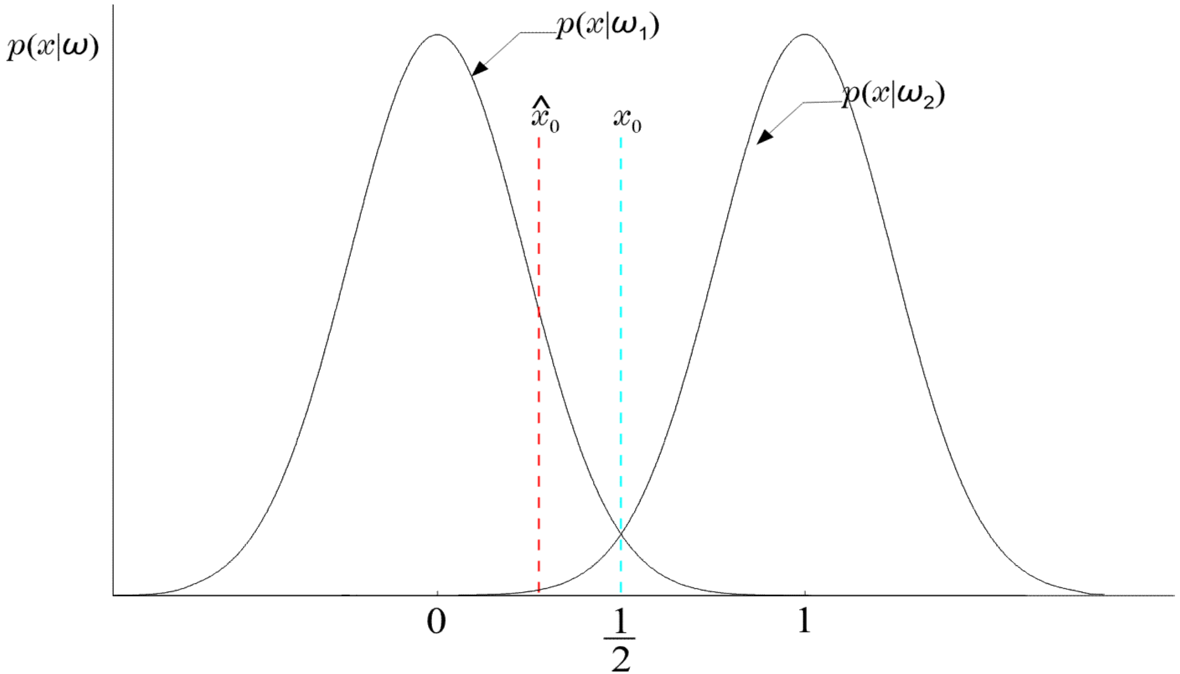
- Threshold $\hat{x}_0$ for minimum r

$$\hat{x}_0 : \exp(-x^2) = 2\exp(-(x-1)^2) \Rightarrow$$

$$\hat{x}_0 = \frac{(1-\ln 2)}{2} < \frac{1}{2}$$

Thus $\hat{x}_0$ moves to the left of $\dfrac{1}{2} = x_0$

(WHY?)



بازشناسی الگو

طبقه‌بندی مبتنی بر نظریه‌ی تصمیم بیز

۲

توابع تفکیک و مرزهای تصمیم

تابع تفکیک

DISCRIMINANT FUNCTION

تابع تفکیک
Discriminant Function

تابعی برای تصمیم‌گیری در مورد کلاس‌ها

تابع تفکیک برای کلاس i

$$g_i(\mathbf{x}), i = 1, \ldots, c$$

classifier assigns a feature vector $\mathbf{x}$ to class w_i if

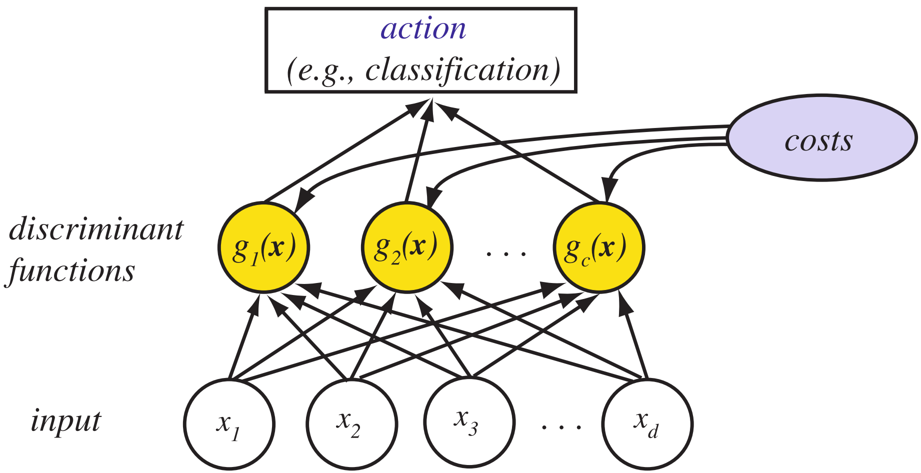
$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i$$

$$\boxed{i^* = \arg\max_i g_i(\mathbf{x})}$$

تابع تفکیک

Discriminant Function

ساختار تابعی یک طبقه‌بندی کننده‌ی الگوی آماری عمومی با توابع تفکیک



From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification, John Wiley & Sons, Inc., 2001.

تابع تفکیک

مثال

DISCRIMINANT FUNCTION

- For the classifier that minimizes conditional risk

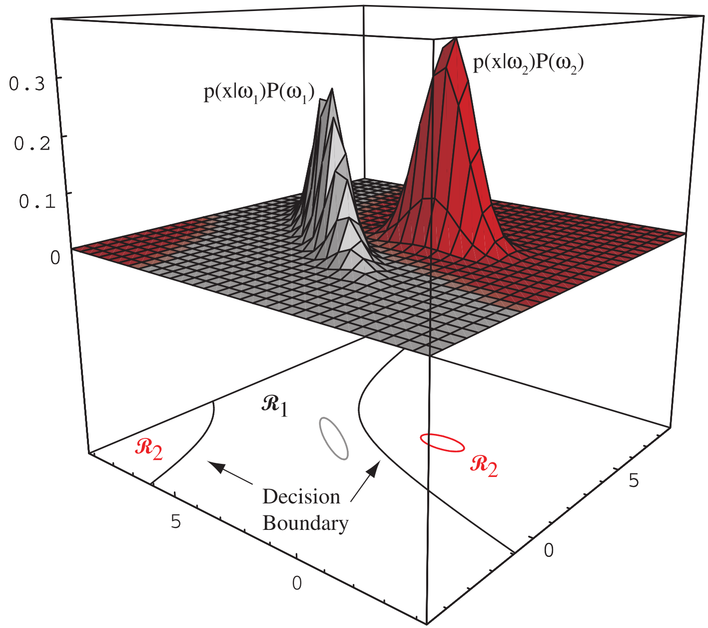
$$g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$$

- For the classifier that minimizes error

$$g_i(\mathbf{x}) = P(w_i|\mathbf{x})$$

تابع تفکیک

DISCRIMINANT FUNCTION



These functions divide the feature space into c *decision regions* $\mathcal{R}_1, \ldots, \mathcal{R}_c$ separated by *decision boundaries*

From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification, John Wiley & Sons, Inc., 2001.

تابع تفکیک

خصوصیات

DISCRIMINANT FUNCTION

هر تابع صعودی از تابع تفکیک، خود یک تابع تفکیک است.

مثل: ضرب در عدد مثبت، جمع با یک عدد، لگاریتم گرفتن، ... ($\Leftarrow$ درک بهتر / تسریع محاسبات)

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j)}$$

$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$$

$$g_i(\mathbf{x}) = \ln\, p(\mathbf{x}|\omega_i) + \ln\, P(\omega_i)$$

تابع تفکیک

دایکوتومایزر (تفکیک‌گر به دو بخش)

DICHOTOMIZER

تابع دایکوتومایزر
Dichotomizer Function

تابع تفکیک برای جداسازی دو کلاس

تابع دایکوتومایزر برای کلاس 1 و 2:

$$g(\mathbf{x}) \equiv g_1(\mathbf{x}) - g_2(\mathbf{x})$$

Decide ω_1 if $g(\mathbf{x}) > 0$; otherwise decide ω_2.

تابع تفکیک

دایکوتومایزر (تفکیک‌گر به دو بخش): مثال

DICHOTOMIZER

$$g(\mathbf{x}) \equiv g_1(\mathbf{x}) - g_2(\mathbf{x})$$

$$g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x})$$

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

DISCRIMINANT FUNCTIONS & DECISION SURFACES

- If R_i, R_j are contiguous: $g(\underline{x}) \equiv P(\omega_i|\underline{x}) - P(\omega_j|\underline{x}) = 0$

$$R_i : P(\omega_i|\underline{x}) > P(\omega_j|\underline{x})$$

$$\begin{array}{l} + \\ \hline - \end{array} \quad g(\underline{x}) = 0$$

$$R_j : P(\omega_j|\underline{x}) > P(\omega_i|\underline{x})$$

is the surface separating the regions.
On one side is positive (+), on the other is negative (-).
It is known as **Decision Surface**

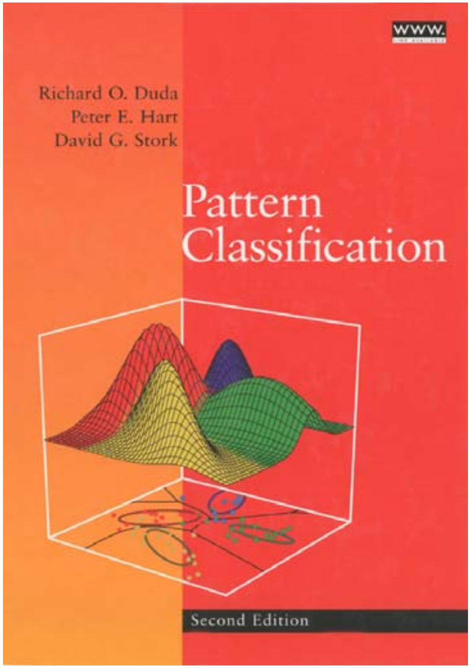
- If $f(.)$ monotonic, the rule remains the same if we use:

$$\underline{x} \rightarrow \omega_i \text{ if : } f(P(\omega_i|\underline{x})) > f(P(\omega_j|\underline{x})) \quad \forall i \neq j$$

- $g_i(\underline{x}) \equiv f(P(\omega_i|\underline{x}))$ is a **discriminant function**

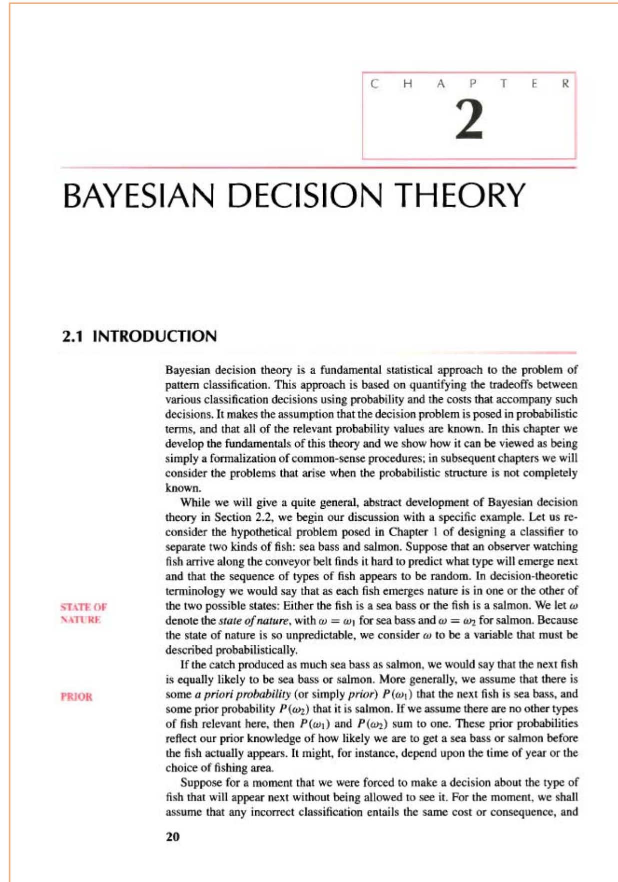
- In general, discriminant functions can be defined independent of the Bayesian rule.
  They **lead to suboptimal solutions**, yet if chosen appropriately, can be computationally more tractable.

منبع اصلی



R.O. Duda, P.E. Hart, and D.G. Stork,
Pattern Classification,
Second Edition, John Wiley & Sons, Inc., 2001.

Chapter 2



C H A P T E R 2

BAYESIAN DECISION THEORY

2.1 INTRODUCTION

Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. This approach is based on quantifying the tradeoffs between various classification decisions using probability and the costs that accompany such decisions. It makes the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known. In this chapter we develop the fundamentals of this theory and we show how it can be viewed as being simply a formalization of common-sense procedures; in subsequent chapters we will consider the problems that arise when the probabilistic structure is not completely known.

While we will give a quite general, abstract development of Bayesian decision theory in Section 2.2, we begin our discussion with a specific example. Let us reconsider the hypothetical problem posed in Chapter 1 of designing a classifier to separate two kinds of fish: sea bass and salmon. Suppose that an observer watching fish arrive along the conveyor belt finds it hard to predict what type will emerge next and that the sequence of types of fish appears to be random. In decision-theoretic terminology we would say that as each fish emerges nature is in one or the other of the two possible states: Either the fish is a sea bass or the fish is a salmon. We let ω denote the *state of nature*, with $\omega = \omega_1$ for sea bass and $\omega = \omega_2$ for salmon. Because the state of nature is so unpredictable, we consider ω to be a variable that must be described probabilistically.

STATE OF NATURE

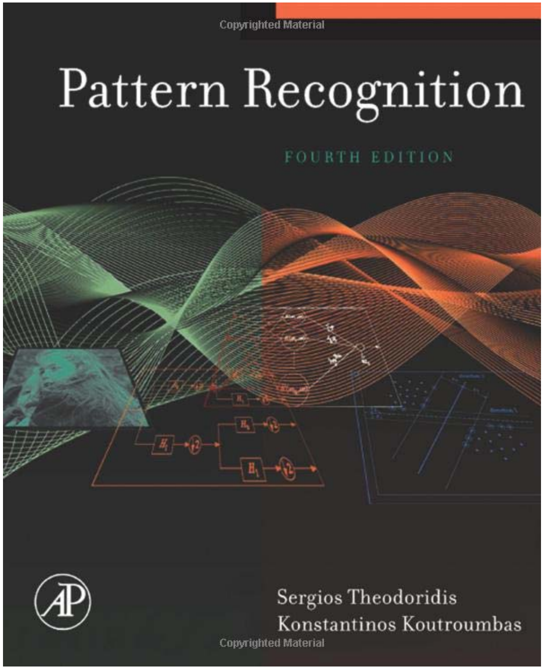
If the catch produced as much sea bass as salmon, we would say that the next fish is equally likely to be sea bass or salmon. More generally, we assume that there is some *a priori probability* (or simply *prior*) $P(\omega_1)$ that the next fish is sea bass, and some prior probability $P(\omega_2)$ that it is salmon. If we assume there are no other types of fish relevant here, then $P(\omega_1)$ and $P(\omega_2)$ sum to one. These prior probabilities reflect our prior knowledge of how likely we are to get a sea bass or salmon before the fish actually appears. It might, for instance, depend upon the time of year or the choice of fishing area.

PRIOR

Suppose for a moment that we were forced to make a decision about the type of fish that will appear next without being allowed to see it. For the moment, we shall assume that any incorrect classification entails the same cost or consequence, and

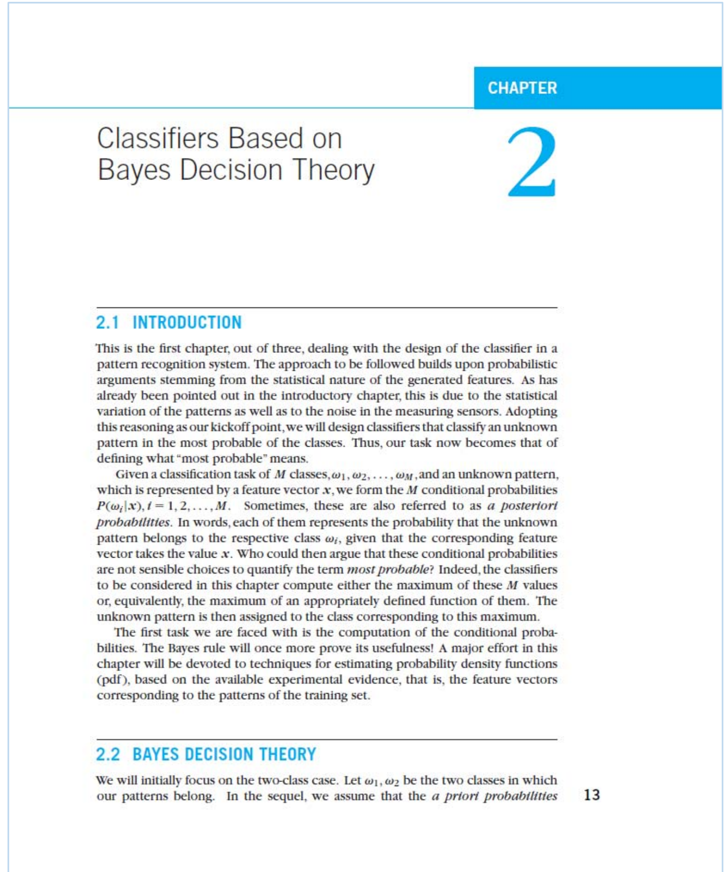
20

منبع اصلی



S. Theodoridis, K. Koutroumbas,
Pattern Recognition,
Fourth Edition, Academic Press, 2009.

Chapter 2



CHAPTER

Classifiers Based on Bayes Decision Theory 2

2.1 INTRODUCTION

This is the first chapter, out of three, dealing with the design of the classifier in a pattern recognition system. The approach to be followed builds upon probabilistic arguments stemming from the statistical nature of the generated features. As has already been pointed out in the introductory chapter, this is due to the statistical variation of the patterns as well as to the noise in the measuring sensors. Adopting this reasoning as our kickoff point, we will design classifiers that classify an unknown pattern in the most probable of the classes. Thus, our task now becomes that of defining what "most probable" means.

Given a classification task of M classes, $\omega_1, \omega_2, \ldots, \omega_M$, and an unknown pattern, which is represented by a feature vector x, we form the M conditional probabilities $P(\omega_i|x), i = 1, 2, \ldots, M$. Sometimes, these are also referred to as *a posteriori probabilities*. In words, each of them represents the probability that the unknown pattern belongs to the respective class ω_i, given that the corresponding feature vector takes the value x. Who could then argue that these conditional probabilities are not sensible choices to quantify the term *most probable*? Indeed, the classifiers to be considered in this chapter compute either the maximum of these M values or, equivalently, the maximum of an appropriately defined function of them. The unknown pattern is then assigned to the class corresponding to this maximum.

The first task we are faced with is the computation of the conditional probabilities. The Bayes rule will once more prove its usefulness! A major effort in this chapter will be devoted to techniques for estimating probability density functions (pdf), based on the available experimental evidence, that is, the feature vectors corresponding to the patterns of the training set.

2.2 BAYES DECISION THEORY

We will initially focus on the two-class case. Let ω_1, ω_2 be the two classes in which our patterns belong. In the sequel, we assume that the *a priori probabilities*

13