شناسایی الگو
فصل دوم
طبقه‌بند بیز، مرز تصمیم، تخمین پارامترهای تابع چگالی احتمال

Textbook: Pattern Recognition, S. Theodoridis , K. Koutroumbas, Fourth Edition, 2009

کلاس‌بندی مبتنی بر نظریه بیز
CLASSIFIERS BASED ON BAYES DECISION THEORY

- با دانستن خصوصیات آماری ویژگی‌ها، به هر الگویی که با بردار ویژگی

$$\underline{x} = [x_1, x_2, x_3, \dots, x_l]^T$$

مشخص شده است **محتمل‌ترین** کلاس از مجموعه $\omega_1, \omega_2, \dots, \omega_M$ را تخصیص دهد. بدین معنی که

$$\underline{x} \rightarrow \omega_i : \underset{maximum}{P(\omega_i | \underline{x})}$$

کلاس‌بندی مبتنی بر نظریه بیز

- a-priori probabilities
  - احتمال پیشین کلاس ω_i

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

- the likelihood of $\underline{x}$ $w.r.\ to\ \omega_i$.
  - احتمال شرطی $\underline{x}$ نسبت به ω_i

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

قانون بیز (مساله دو کلاسه)

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

که در آن

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

مخرج می‌تواند حذف شود. زیرا برای همه کلاس‌ها یکسان است.

به تابع چگالی و تابع توزیع احتمال توجه شود.

قانون بیز (مساله دو کلاسه) – ادامه

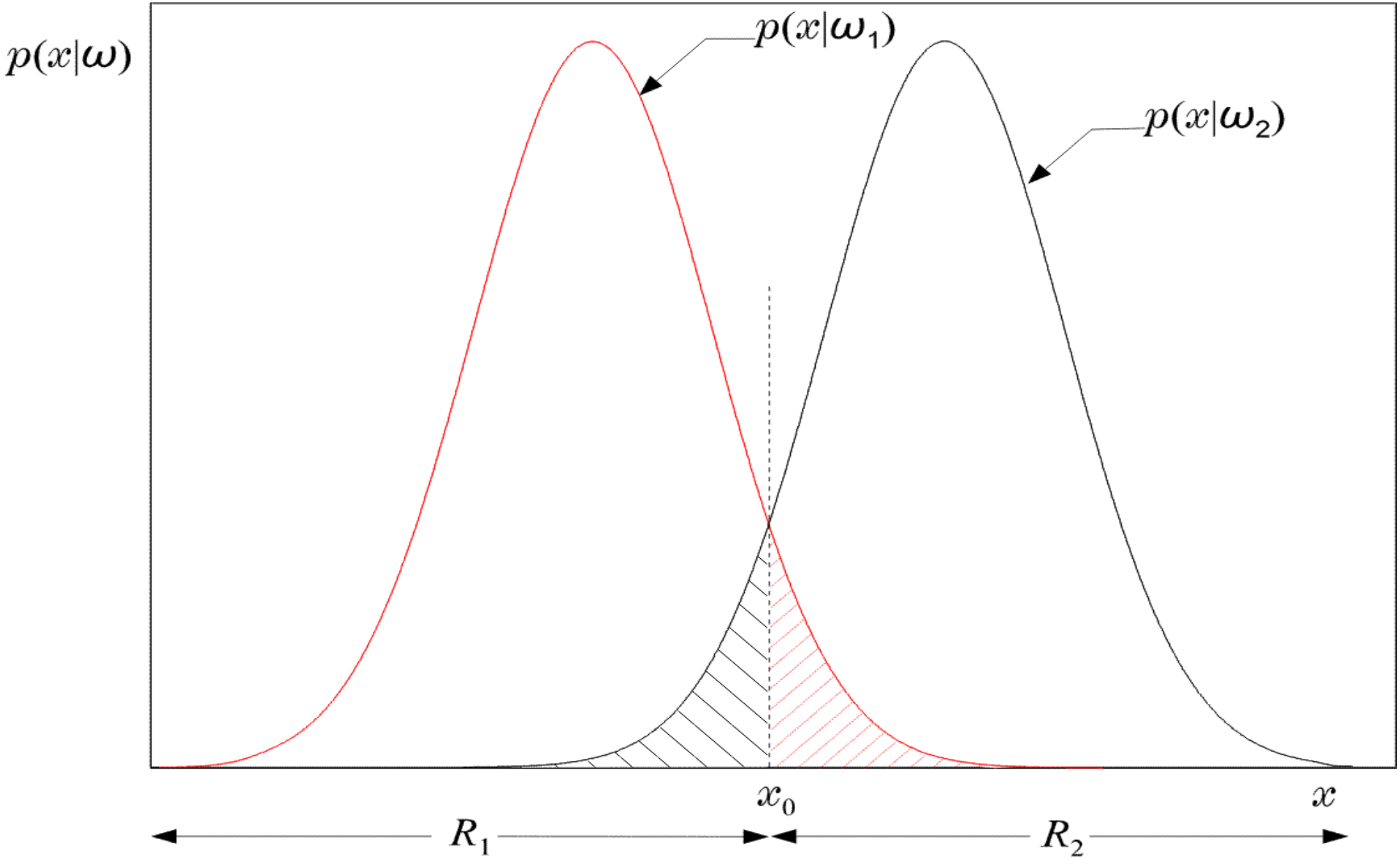
$$\text{If } P(\omega_1|\underline{x}) > P(\omega_2|\underline{x}) \Rightarrow \underline{x} \longrightarrow \omega_1$$

معادل است با

$$\text{If } \mathrm{p}(\underline{x}|\omega_1)P(\omega_1) > \mathrm{p}(\underline{x}|\omega_2)P(\omega_2) \Rightarrow \underline{x} \longrightarrow \omega_1$$

اگر احتمال پیشین این دو کلاس باهم برابر باشد داریم:

$$\mathrm{p}(\underline{x}|\omega_1) > \mathrm{p}(\underline{x}|\omega_2) \Rightarrow \underline{x} \longrightarrow \omega_1$$



$$R_1(\rightarrow \omega_1) \ and \ R_2(\rightarrow \omega_2)$$

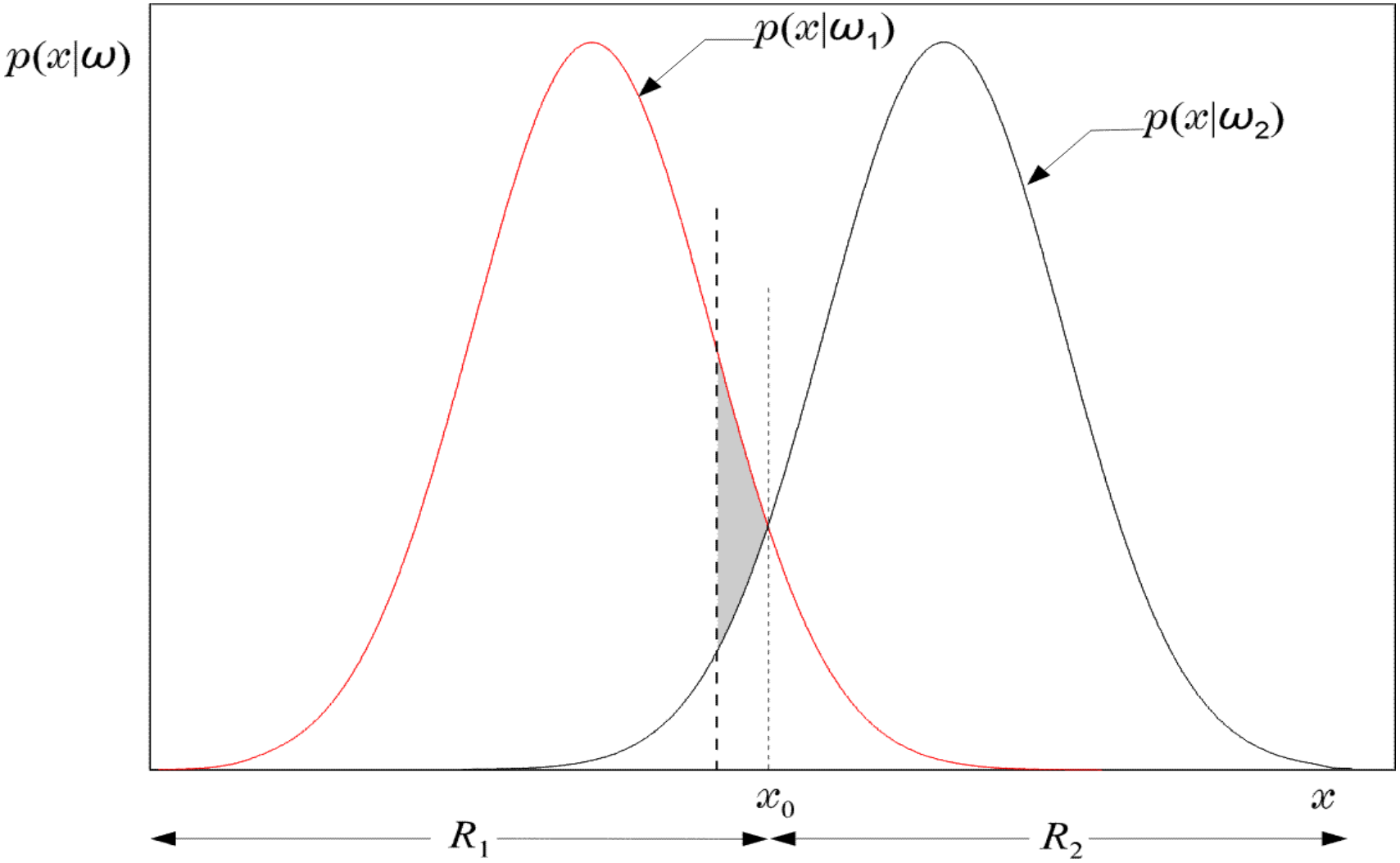
با توجه به مقادیر تابع چگالی احتمال می‌توان فضای مساله را به دو ناحیه تقسیم نمود.

خطای کلاس‌بندی

- مساحت ناحیه هاشور خورده در شکل صفحه قبل خطای سیستم را نشان می‌دهد

$$P_e = \frac{1}{2}\int_{-\infty}^{x_0} p(x|\omega_2)dx + \frac{1}{2}\int_{x_0}^{+\infty} p(x|\omega_1)dx$$

- نسبت به کمینه بودن خطای کلاس‌بندی طبقه‌بند بیز **بهینه** است.



- با جابه‌جایی مقدار آستانه مساحت ناحیه هاشور خورده افزایش یافته در نتیجه خطای کلاس‌بندی افزایش می‌یابد.

قانون بیز برای بیش از دو کلاس

- متغیر $\underline{x}$ متعلق به کلاس ω_i خواهد بود اگر

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

کمینه‌سازی میانگین ریسک
Minimizing the average risk

- به ازای هر تصمیم نادرست جریمه‌ای در نظر گرفته می‌شود چون برخی از تصمیم‌ها حساس‌ترند.
- به طور مثال هزینه فرد سالمی که مبتلا به بیماری واگیردار شناخته می‌شود؛ از بیماری که توسط طبقه‌بند سالم کلاس‌بندی می‌شود بسیار بیشتر است.
- این هزینه‌ها معمولا در قالب ماتریس اتلاف بیان می‌شوند.

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

کمینه‌سازی میانگین ریسک

- ریسک نسبت به کلاس ω_1

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

- ریسک نسبت به کلاس ω_2

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

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

کمینه‌سازی میانگین ریسک

- باید نواحی R_1 و R_2 به نحوی انتخاب شود که مقدار r کمینه شود.
- داده x متعلق به کلاس ω_1 است اگر:

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

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

کمینه‌سازی میانگین ریسک (حالت خاص)

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

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

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

$$\text{if } \lambda_{21} = \lambda_{12} \Rightarrow \text{Minimum classification error probability}$$

مثال

- $p(x|\omega_1) = \frac{1}{\sqrt{\pi}} \exp(-x^2)$
- $p(x|\omega_2) = \frac{1}{\sqrt{\pi}} \exp(-(x-1)^2)$
- $P(\omega_1) = P(\omega_2) = \frac{1}{2}$
- $L = \begin{pmatrix} 0 & 0.5 \\ 1.0 & 0 \end{pmatrix}$

مثال

- Then the threshold value is:

$$x_0 \text{ for minimum } P_e :$$

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

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

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

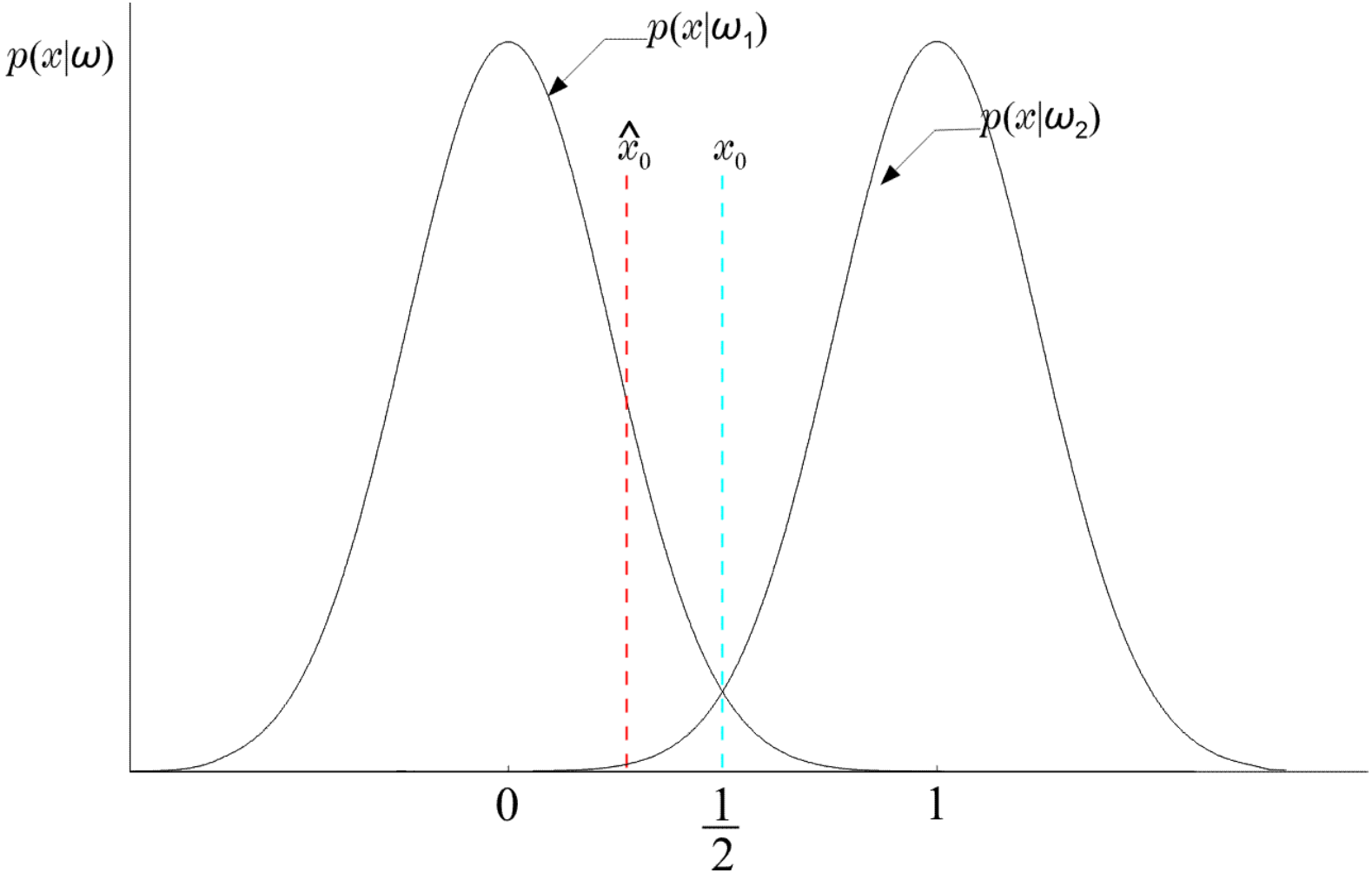
$$\hat{x}_0 : \exp(-x^2) = 2\exp(-(x-1)^2) \Rightarrow$$

$$\hat{x}_0 = \frac{(1-\ell n2)}{2} < \frac{1}{2}$$

مثال (ادامه)

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

(WHY?)



تابع توزیع نرمال یک بعدی

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

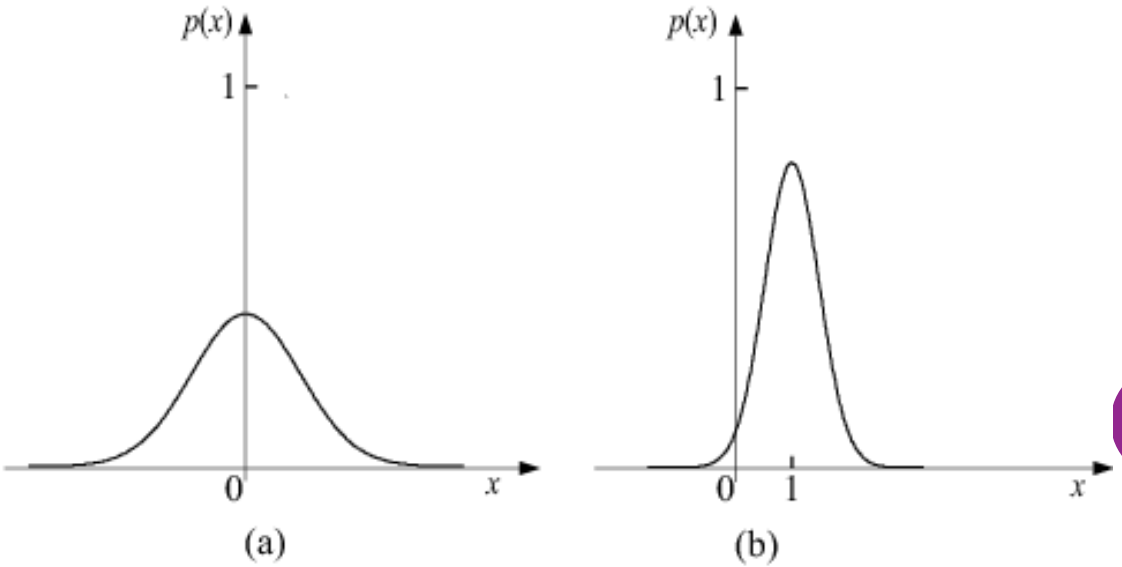
where

μ is the mean value, i.e.:

$$\mu = E[x] = \int_{-\infty}^{+\infty} x p(x) dx$$

σ^2 is the variance,

$$\sigma^2 = E\left[(x - E[x])^2\right] = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x) dx$$



تابع چگالی نرمال چند بعدی

$$p(\underline{x}) = \frac{1}{(2\pi)^{\frac{\ell}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \right)$$

$\mu: Expected\ value, l \times 1, E[\underline{x}]$

$\Sigma: covariance\ matric, l \times l, E\left[(\underline{x} - \mu)(\underline{x} - \mu)^T\right]$

$|.|: determinant$

$l: number\ of\ features$

$-\frac{1}{2}(\underline{x} - \mu)^T \Sigma^{-1} (\underline{x} - \mu): Scaler\ Number$

معکوس ماتریس دو در دو

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$determinant\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

تابع توزیع نرمال دو بعدی

$$p(\underline{x}) = p(x_1, x_2) = \frac{1}{(2\pi|\Sigma|)^{\frac{1}{2}}} xp\left(-\frac{1}{2}[x_1-\mu_1, x_2-\mu_2]\Sigma^{-1}\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}\right)$$

$$\Sigma = E\left[\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}[x_1-\mu_1, x_2-\mu_2]\right] = \begin{bmatrix} \sigma_1^2 & \sigma \\ \sigma & \sigma_2^2 \end{bmatrix}$$

$$\sigma = E[(x_1-\mu_1)(x_2-\mu_2)]$$

$$\underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} E[x_1] \\ E[x_2] \end{bmatrix}$$

شناسایی الگو
فصل دوم
مرز تصمیم

طبقه‌بند بیز برای توزیع احتمال نرمال

- تابع چگالی نرمال چند بعدی (متغیره)

$$p(\underline{x}) = \frac{1}{(2\pi)^{\frac{\ell}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu})^T \Sigma^{-1} (\underline{x}-\underline{\mu})\right)$$

$\mu: Expected\ value, l \times 1, E[\underline{x}]$

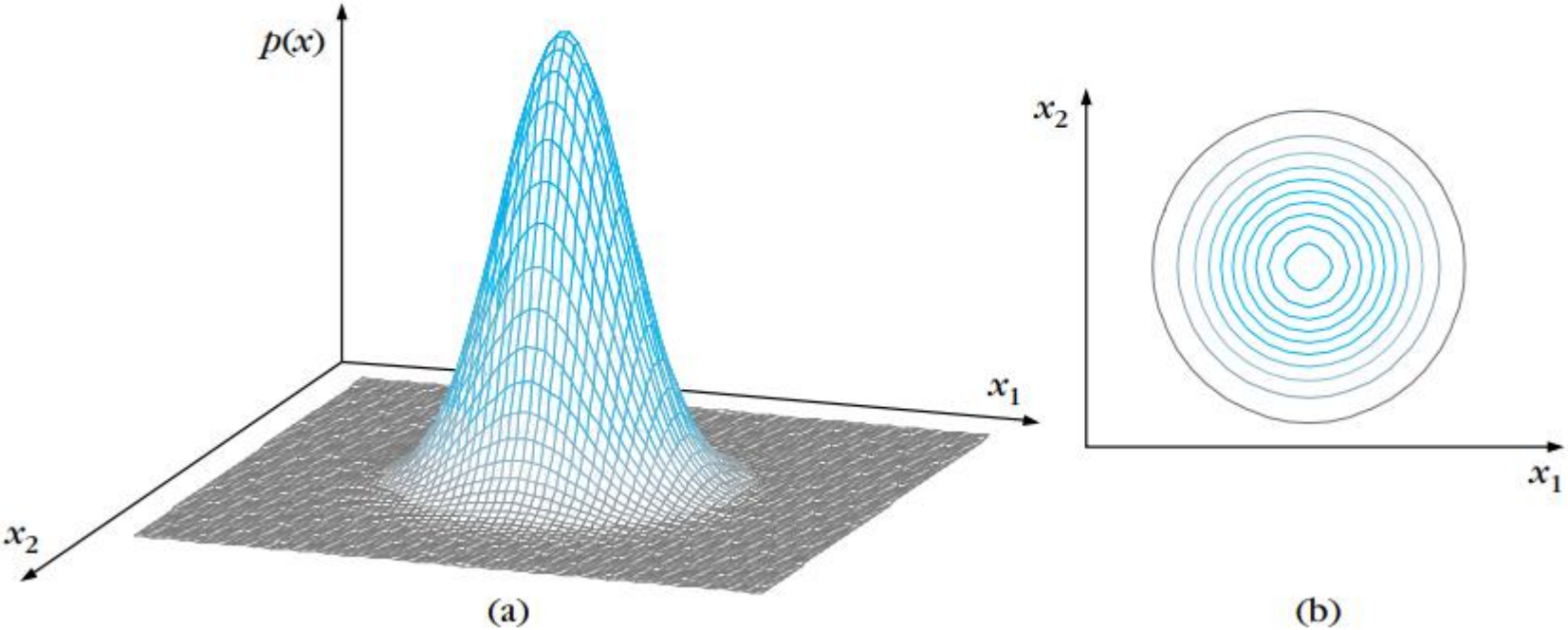
$\Sigma: covariance\ matric, l \times l, E\left[(\underline{x}-\mu)(\underline{x}-\mu)^T\right]$

$|.|: determinant$

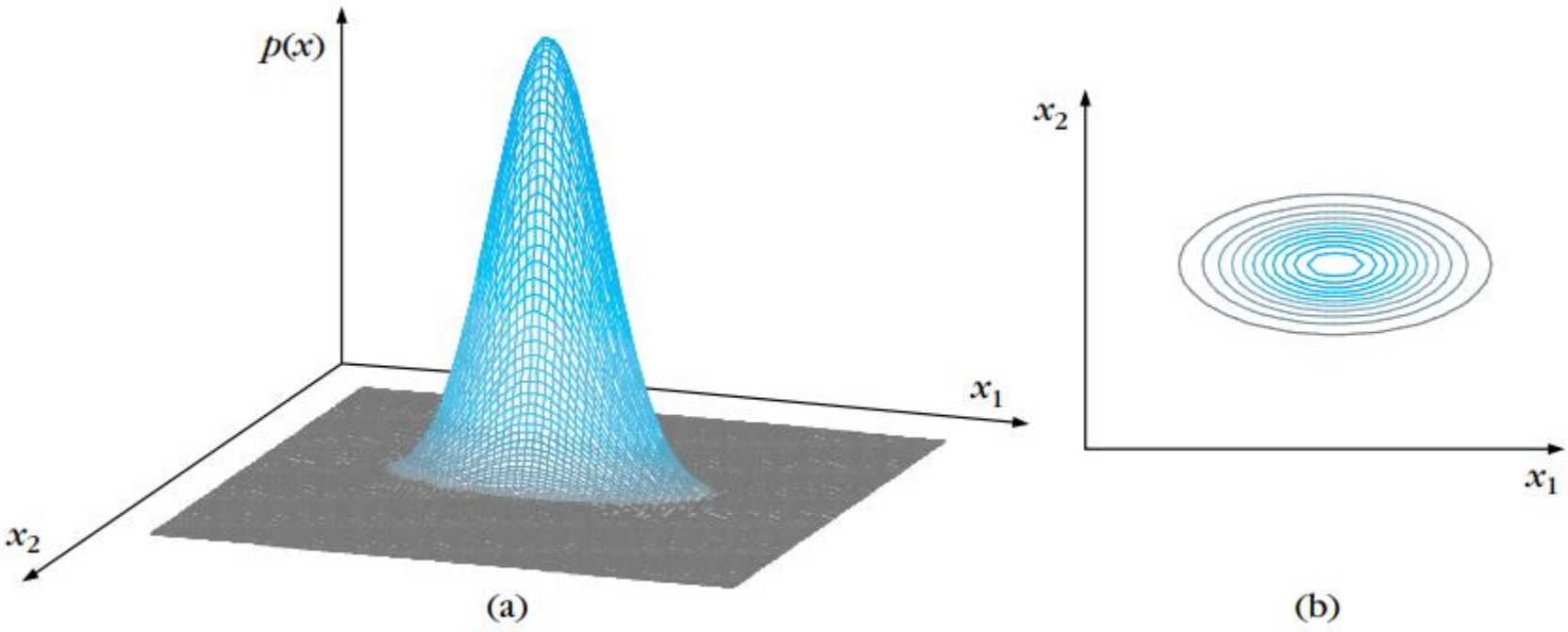
$l: number\ of\ features$

$-\frac{1}{2}(\underline{x}-\mu)^T \Sigma^{-1} (\underline{x}-\mu): Scaler\ Number$

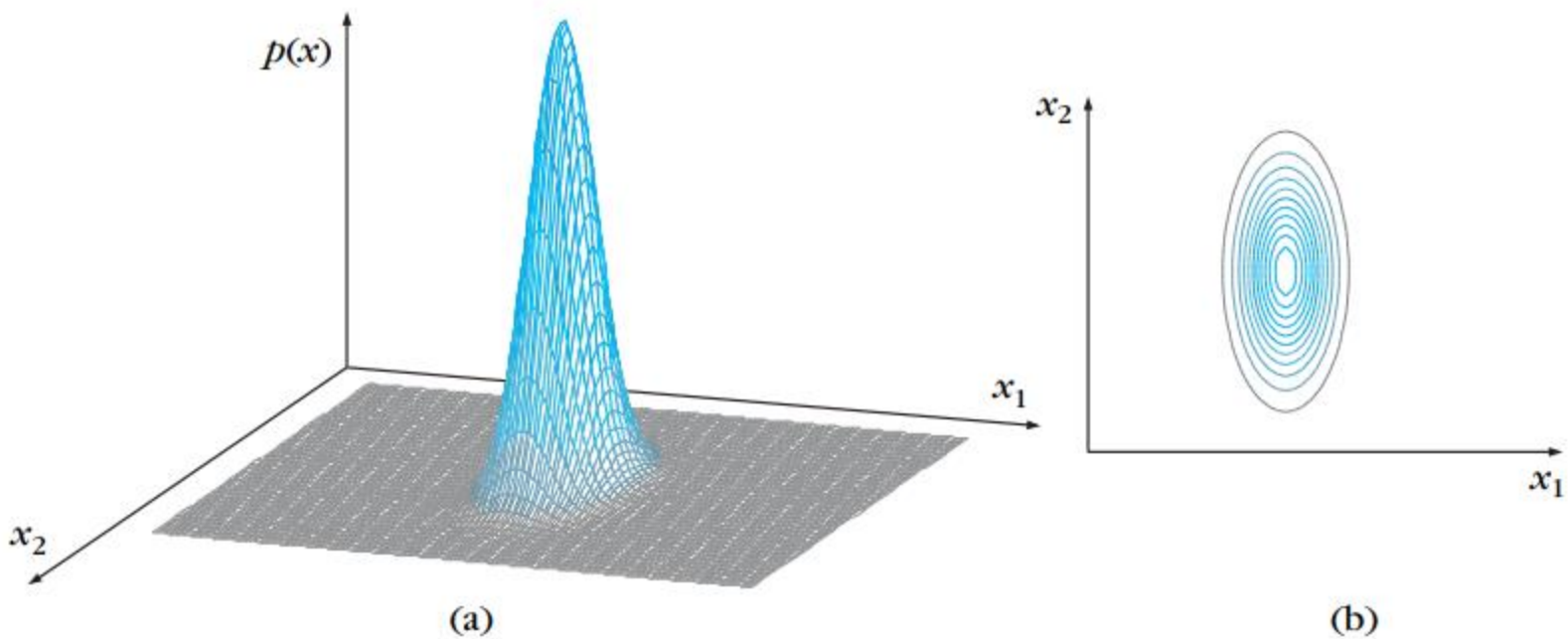
$$\Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$



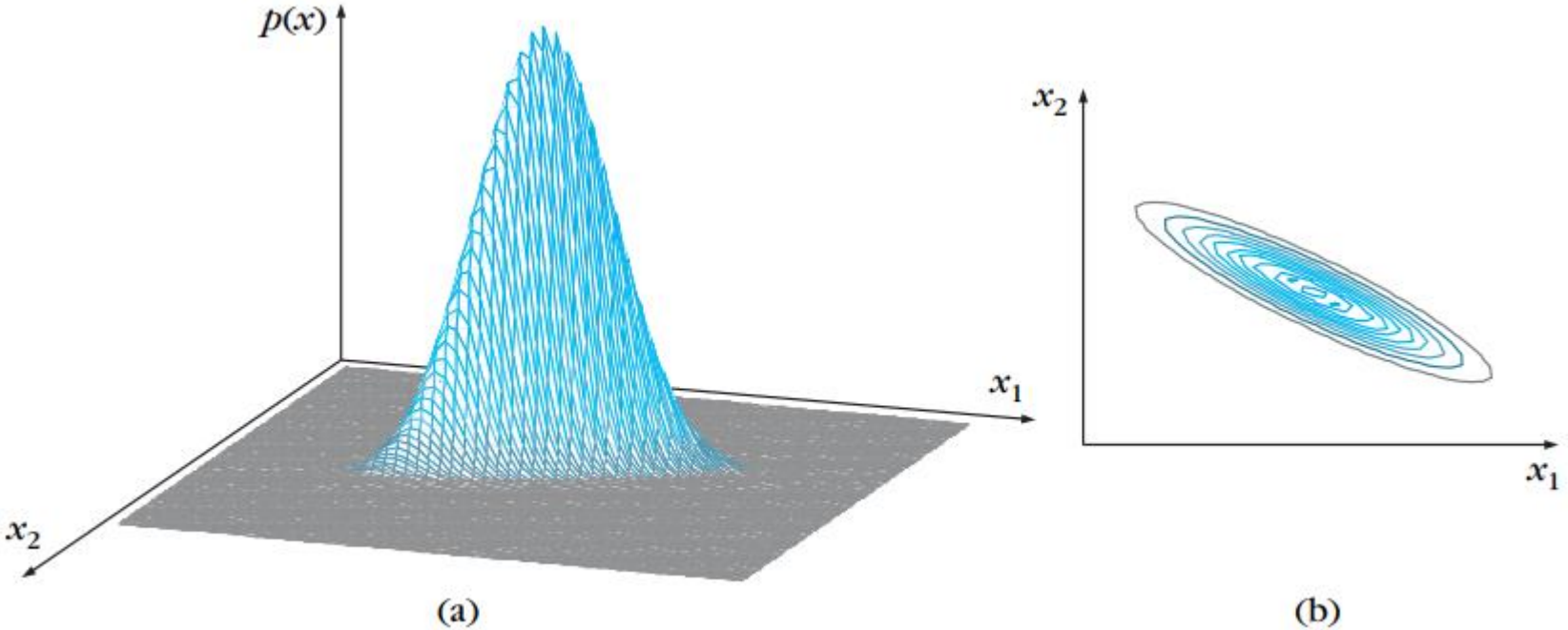
Two-dimensional Gaussian pdf and its contour. The graph has a spherical symmetry showing no preference in any direction.



Two-dimensional Gaussian pdf and its contour. Σ is diagonal with $\sigma_1^2 \gg \sigma_2^2$. The graph is elongated along the x_1 direction.



Two-dimensional Gaussian pdf and its contour. Σ is diagonal with $\sigma_2^2 \gg \sigma_1^2$. The graph is elongated along the $x_2$direction.



Two-dimensional Gaussian pdf and its contour.

nondiagonal Σ

تابع جداکننده – سطح جداکننده

- دو کلاس i و j توسط تابع g از هم جدا می شوند.

$$g(\underline{x}) \equiv P(\omega_i|\underline{x}) - P(\omega_j|\underline{x}) = 0$$

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

$+$

$$g(\underline{x}) = 0$$

$-$

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

- در یک سمت این تابع مقدار مثبت (کلاس i) و در سمت دیگر مقدار منفی (کلاس j) است. تابع g بر روی مرز این دو ناحیه مقدار صفر را دارد.

تابع جداکننده – سطح جداکننده

- برای یافتن مرز دو کلاس باید تابع g را برابر صفر قرار داد. اما محاسبه آن مشکل است. از این رو ساده‌سازی باید صورت گیرد.
- اگر تابع f به صورت یکنواخت صعودی باشد.

$$P(\omega_i|\underline{x}) \geq P(\omega_j|\underline{x}) \Rightarrow f\left(P(\omega_i|\underline{x})\right) \geq f\left(P(\omega_j|\underline{x})\right)$$

- پس کماکان

$$g(\underline{x}) \equiv f\left(P(\omega_i|\underline{x})\right) - f\left(P(\omega_j|\underline{x})\right) = 0$$

می‌تواند یک مرز جداکننده باشد.

- تابع $f = \ln(.)$ یکنوا است. پس می‌تواند برای ساده‌سازی به کار گرفته شود.

$$g_i(\underline{x}) = ln\left(P(\omega_i|\underline{x})\right)$$

$$g_j(\underline{x}) = ln\left(P(\omega_j|\underline{x})\right)$$

تابع جداکننده – سطح جداکننده – ساده‌سازی

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

$$g_i(\underline{x}) = ln\left(P(\omega_i|\underline{x})\right) = ln\left(p(\underline{x}|\omega_i)P(\omega_i)\right)$$
$$g_j(\underline{x}) = ln\left(P(\omega_j|\underline{x})\right) = ln\left(p(\underline{x}|\omega_j)P(\omega_j)\right)$$

$$g_i(\underline{x}) = ln\left(p(\underline{x}|\omega_i)\right) + ln(P(\omega_i))$$
$$= ln\left(\frac{1}{2\pi^{l/2}|\Sigma_i|^{1/2}}\right) - \frac{1}{2}(\underline{x} - \mu_i)^T \Sigma_i{}^{-1}(\underline{x} - \mu_i) + ln(P(\omega_i))$$

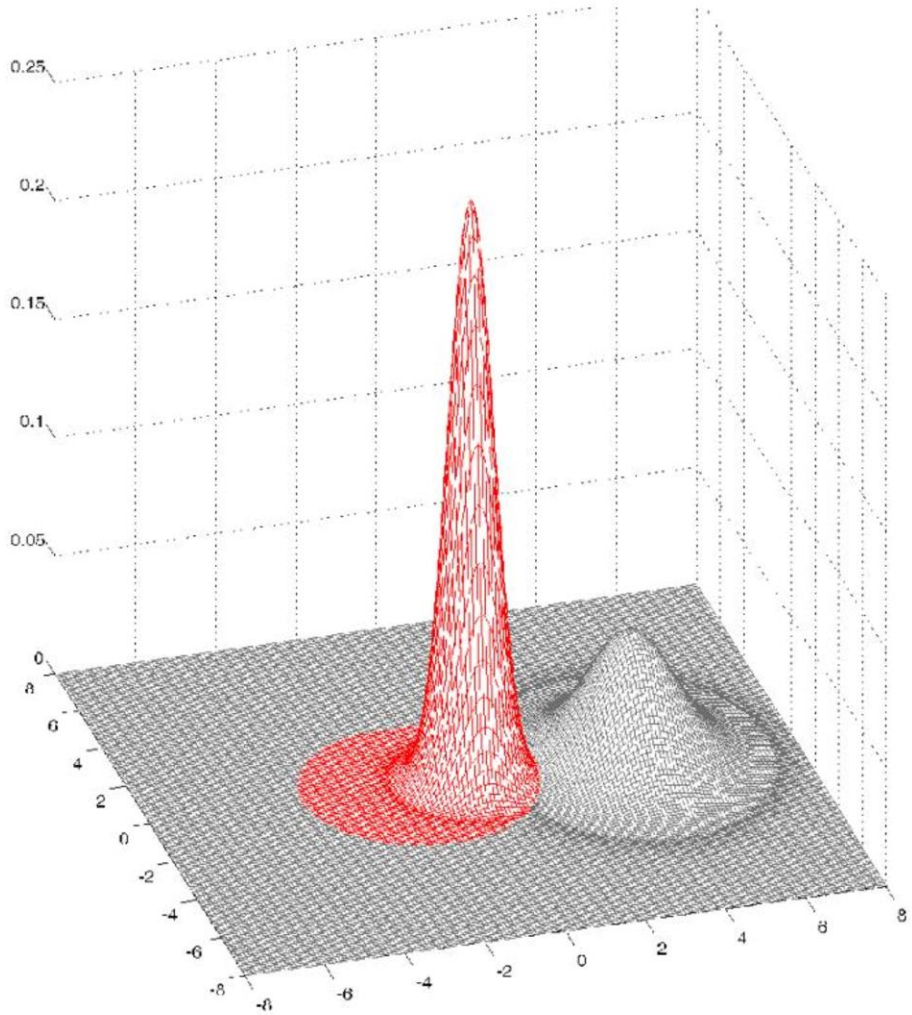
تابع جداکننده – سطح جداکننده – ساده‌سازی

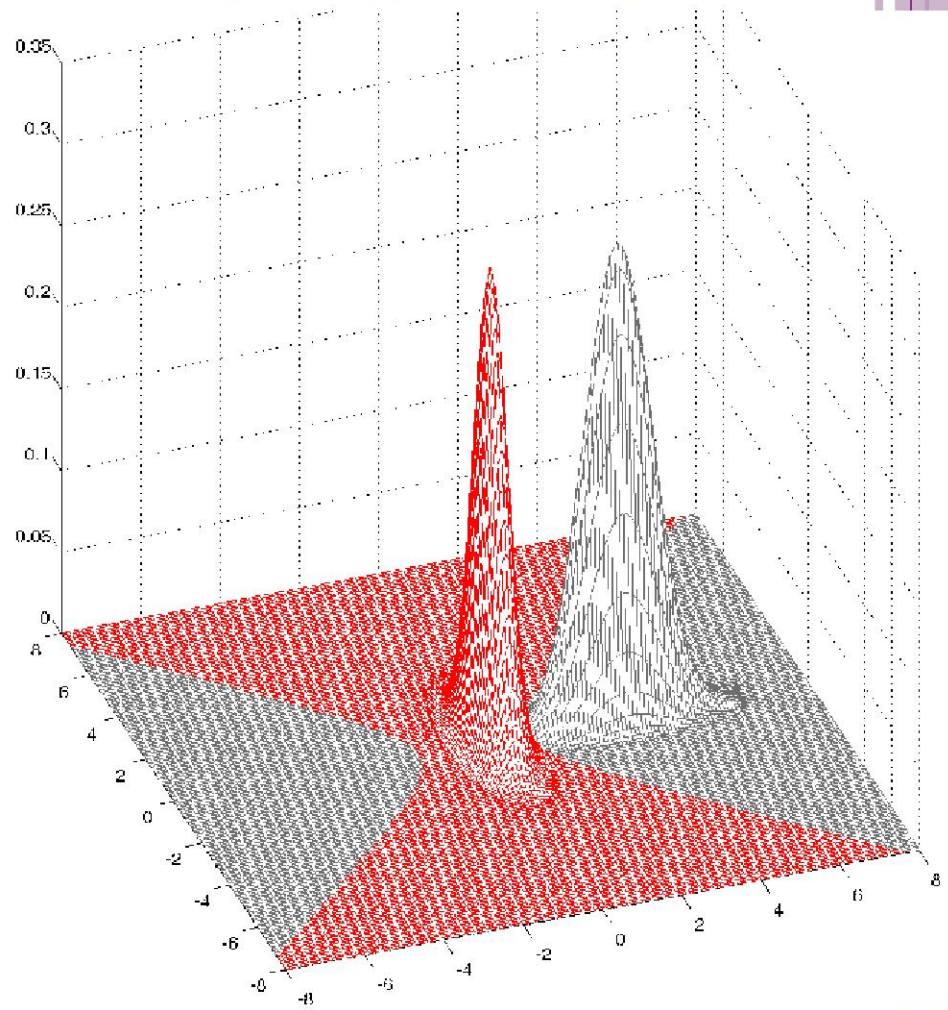
$$\left(\underline{x}-\mu_i\right)^T {\Sigma_i}^{-1}\left(\underline{x}-\mu_i\right)=$$
$$\underline{x}^T {\Sigma_i}^{-1}\underline{x}-\mu_i^T {\Sigma_i}^{-1}\underline{x}-\underline{x}^T {\Sigma_i}^{-1}\mu_i+\mu_i^T {\Sigma_i}^{-1}\mu_i=$$
$$\underline{x}^T {\Sigma_i}^{-1}\underline{x}-2\mu_i^T {\Sigma_i}^{-1}\underline{x}+\mu_i^T {\Sigma_i}^{-1}\mu_i$$

$$g_i(\underline{x})=ln\left(p(\underline{x}|\omega_i)\right)+ln\left(P(\omega_i)\right)$$
$$=ln\left(\frac{1}{2\pi^{l/2}|\Sigma_i|^{1/2}}\right)+\frac{1}{2}\underline{x}^T {\Sigma_i}^{-1}\underline{x}-\mu_i^T {\Sigma_i}^{-1}\underline{x}+\frac{1}{2}\mu_i^T {\Sigma_i}^{-1}\mu_i+ln\left(P(\omega_i)\right)$$

- تنها عامل درجه دوم در این رابطه ${\underline{x}_*}^T {\Sigma_*}^{-1}\underline{x}_*$ است.

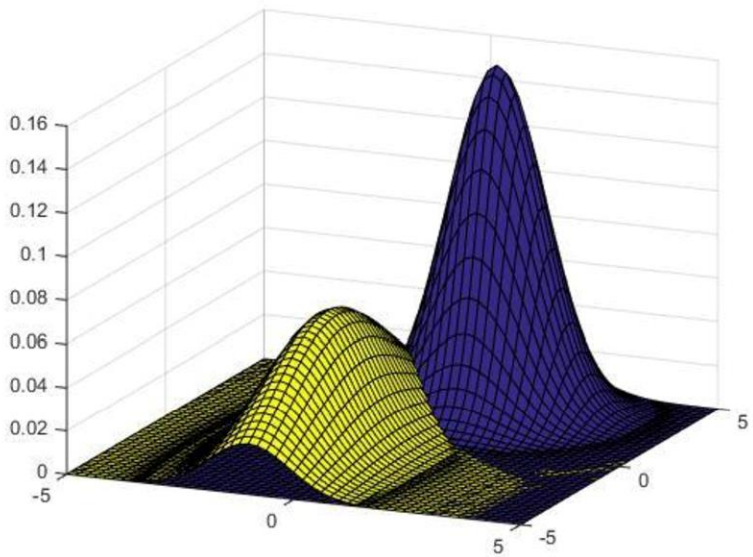
در حالت کلی سطح جداکننده یک معادله درجه دوم است

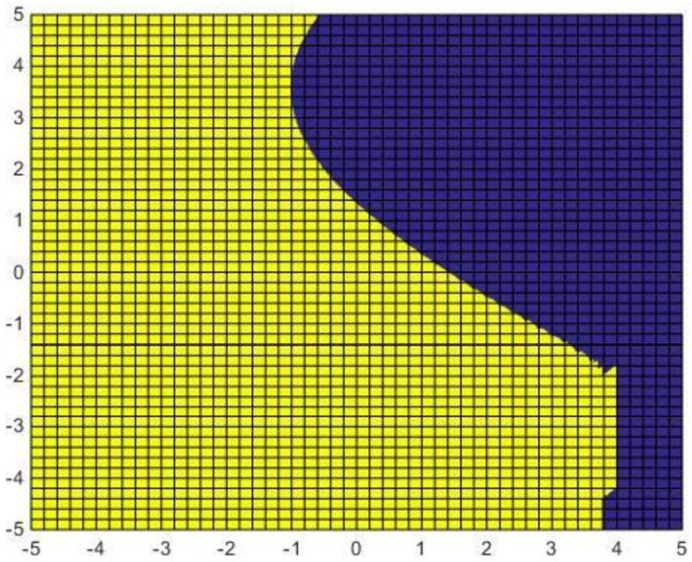




مرز تصمیم (توزیع نرمال چند متغیره)

Sigma1 = [.8 0; 0 8];
Sigma2 = [1 0; 0 1];





ماتریس کواریانس دو تابع چگالی یکسان باشد: $\Sigma_i = \Sigma_j$

$$g_i(\underline{x}) = ln\left(p(\underline{x}|\omega_i)\right) + ln\left(P(\omega_i)\right)$$

$$= ln\left(\frac{1}{2\pi^{l/2}|\Sigma_i|^{1/2}}\right) + \frac{1}{2}\underline{x}^T\Sigma_i{}^{-1}\underline{x} - \mu_i^T\Sigma_i{}^{-1}\underline{x} + \frac{1}{2}\mu_i^T\Sigma_i{}^{-1}\mu_i + ln\left(P(\omega_i)\right)$$

$$g_i(\underline{x}) - g_j(\underline{x}) = W_i\underline{x} + C_i - W_j\underline{x} - C_j$$

$$= W\underline{x} + C$$

$$W = -\mu_i^T\Sigma_i{}^{-1} + \mu_j^T\Sigma_j{}^{-1}$$

اگر ماتریس کواریانس کلاس‌ها باهم برابر باشد سطح تصمیم خط است.

ماتریس کواریانس یکسان و قطری باشد و واریانس ویژگی‌ها برابر باشد.

$$\boldsymbol{\Sigma_i = \Sigma_j = \sigma I}$$

$$g_i(\underline{x}) = ln\left(p(\underline{x}|\omega_i)\right) + ln\left(P(\omega_i)\right)$$

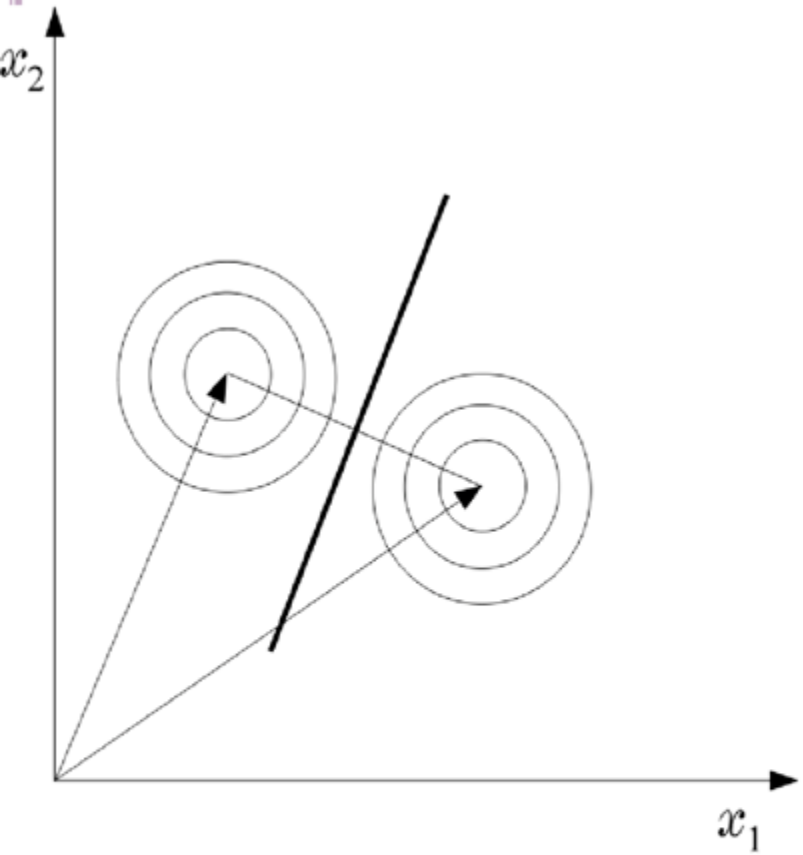
$$= ln\left(\frac{1}{2\pi^{l/2}\sigma_i}\right) + \frac{1}{2\sigma_i^2}\underline{x}^T\underline{x} - \frac{1}{\sigma_i^2}\mu_i^T\underline{x} + \frac{1}{2\sigma_i^2}\mu_i^T\mu_i + ln\left(P(\omega_i)\right)$$

$$g_i(\underline{x}) - g_j(\underline{x}) = W_i\underline{x} + C_i \ - W_j\underline{x} - C_j$$

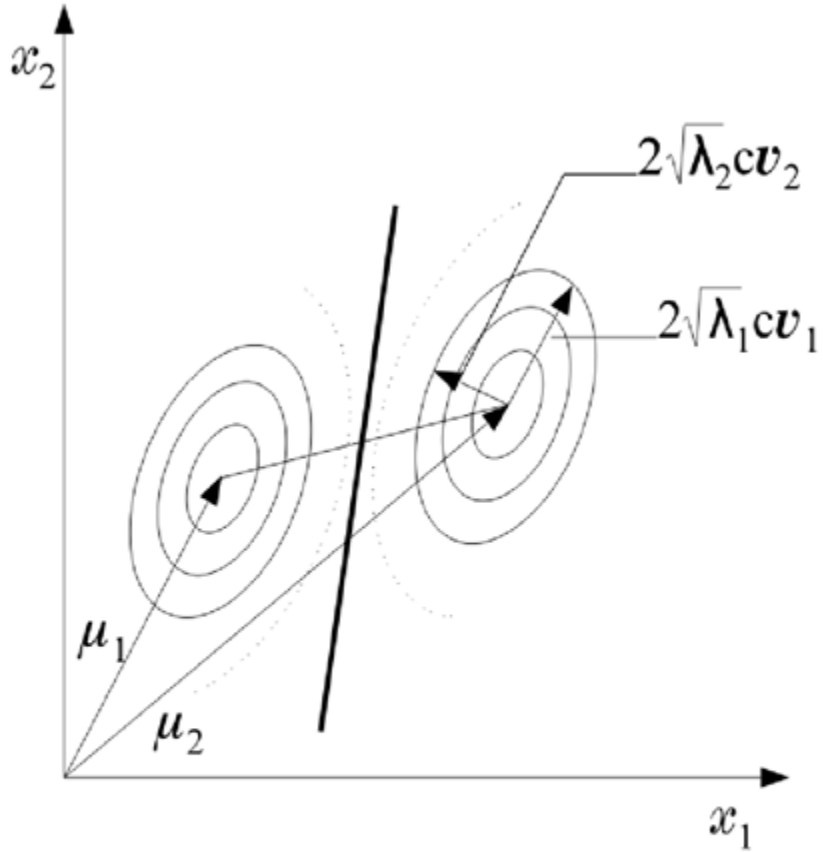
$$= W\underline{x} + C$$

$$W = -\frac{1}{\sigma^2}\left(\mu_i^T - \mu_j^T\right)$$

در شرایط بالا سطح تصمیم، خطی است که بر خط گذرنده از میانگین کلاس‌ها عمود است.



(a)



(b)

$\Sigma = \sigma^2 I$

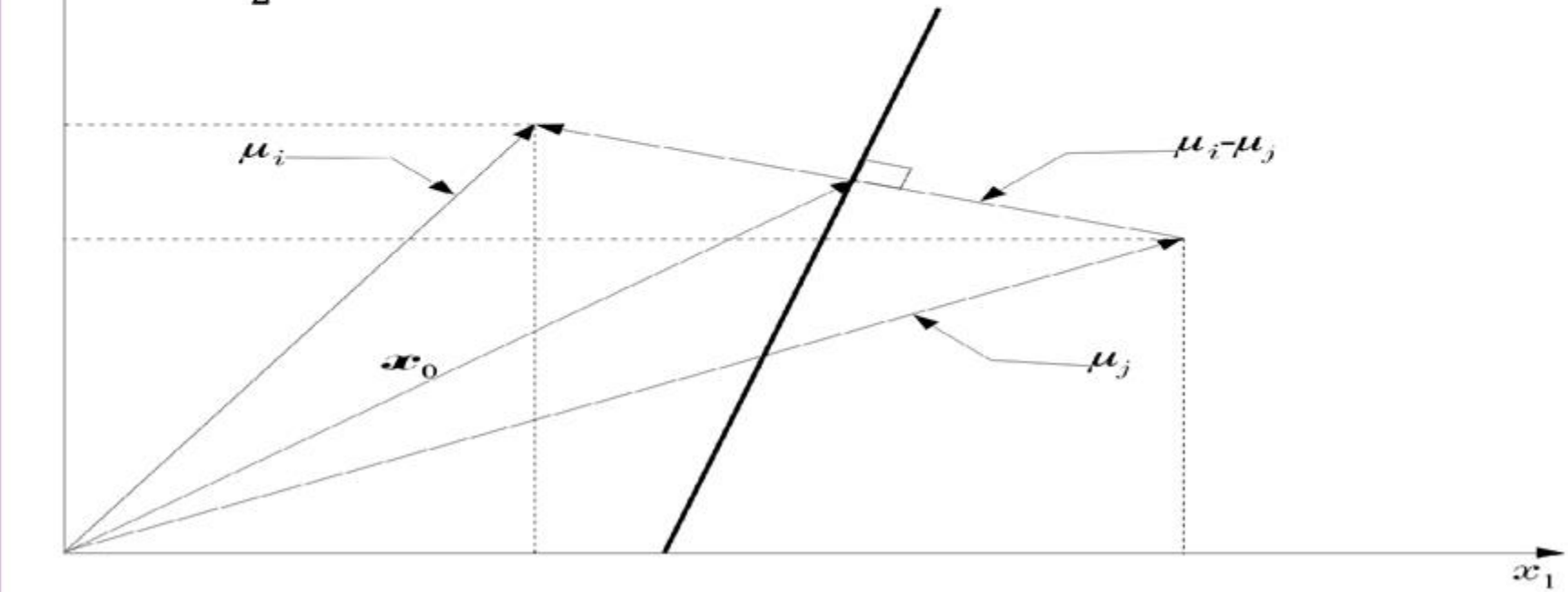
EUCLIDEAN DISTANCE

$\Sigma \neq \sigma^2 I$

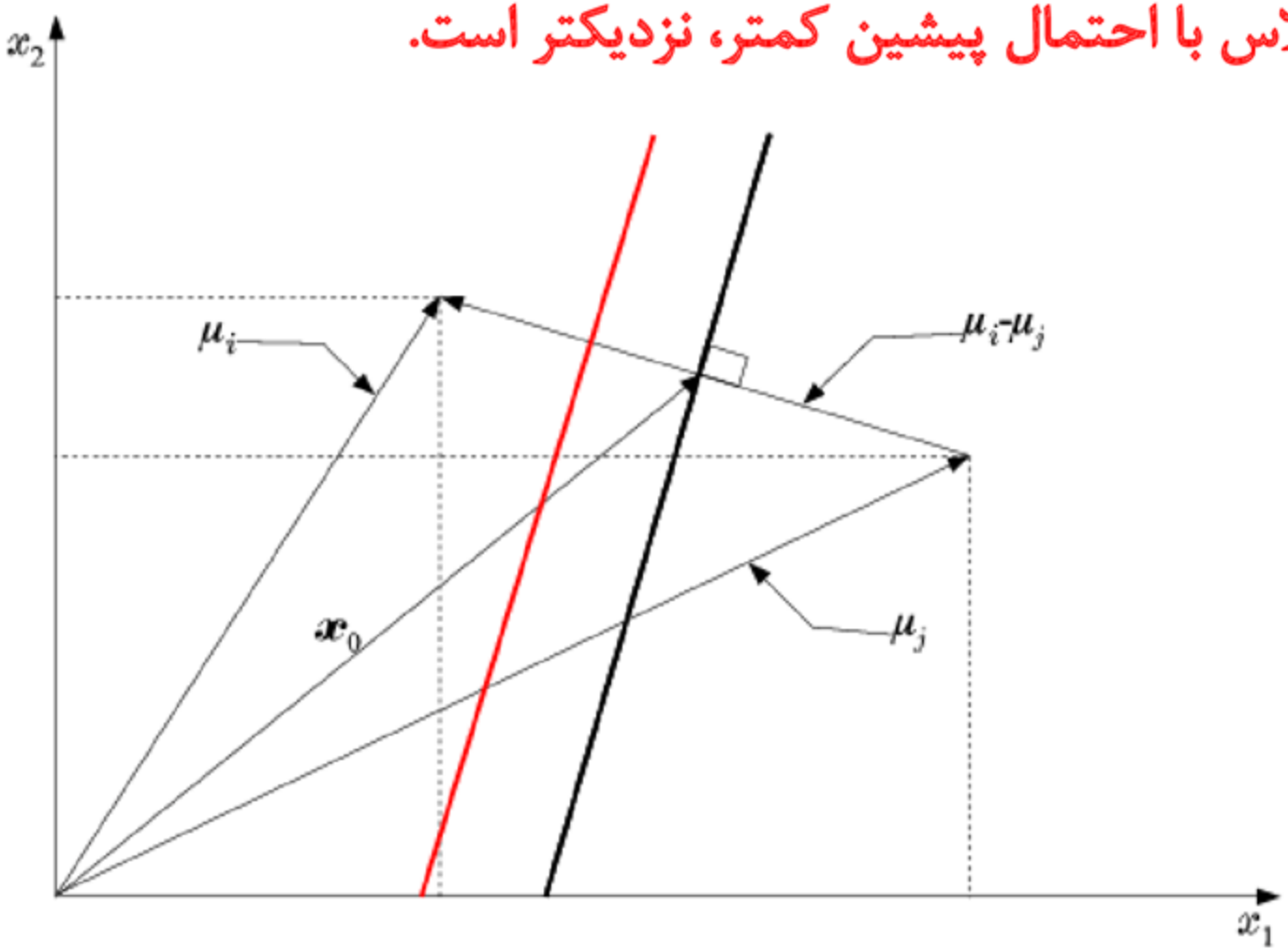
MAHALANOBIS DISTANCE

اگر $\boldsymbol{\Sigma} = \boldsymbol{\sigma}^2 \boldsymbol{I}$ باشد و $p(\omega_1) = p(\omega_2)$ آنگاه عمودمنصف گذرنده از میانگین‌ها، مرز جداکننده است.

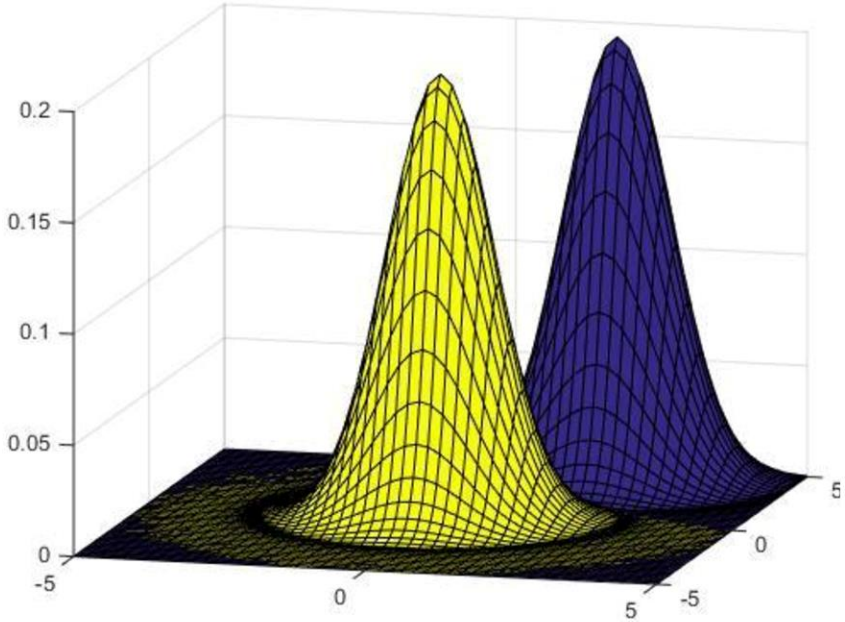
- $x_0 = \frac{1}{2}(\mu_1 + \mu_2)$

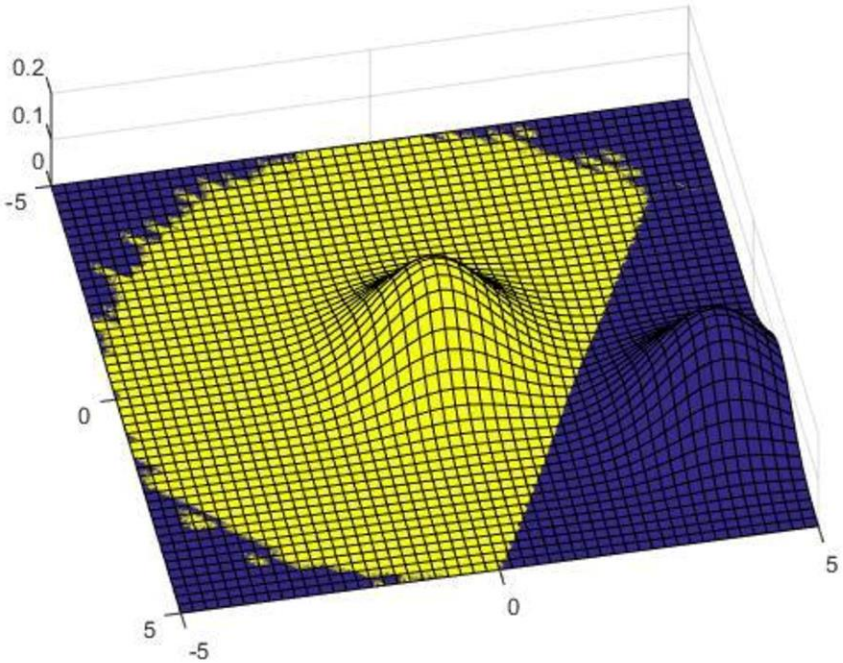


اگر $\Sigma = \sigma^2 I$ باشد و $p(\omega_1) \neq p(\omega_2)$ آنگاه خط جداکننده به کلاس با احتمال پیشین کمتر، نزدیکتر است.



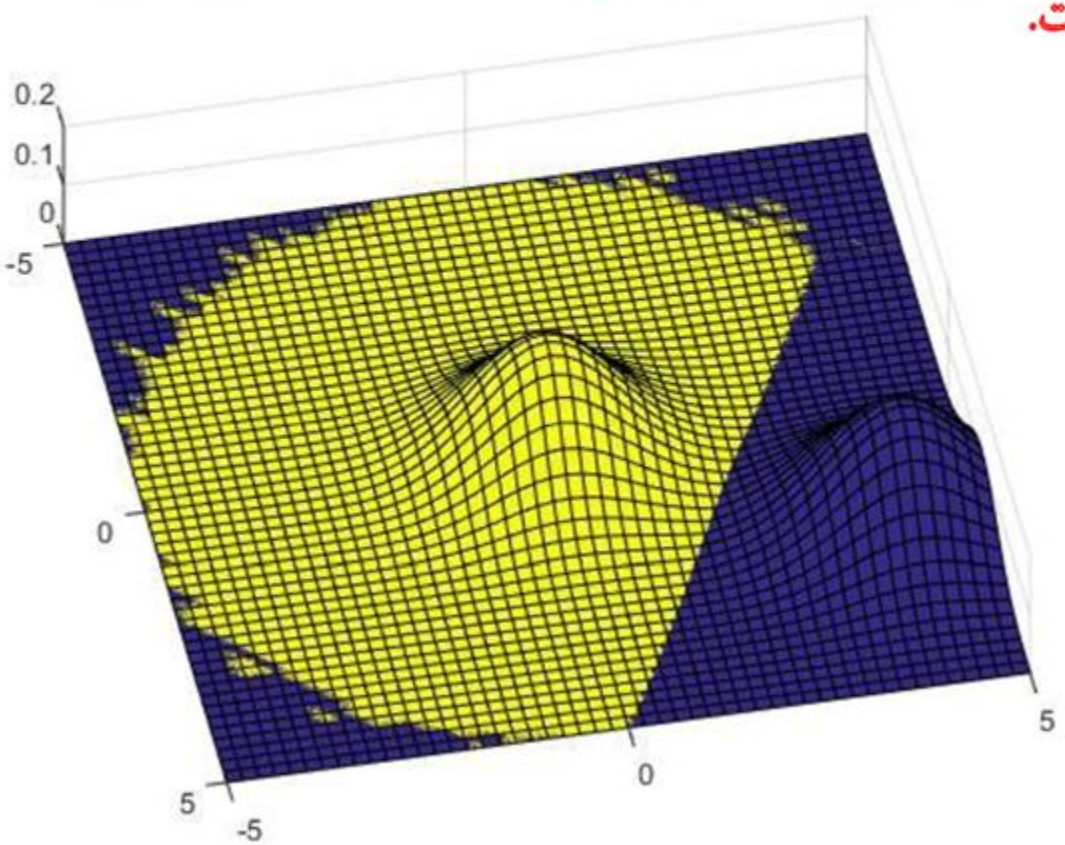
مرز تصمیم در یک مساله دو متغیره با احتمال پیشین برابر و ماتریس کواریانس $\Sigma = \sigma^2 I$



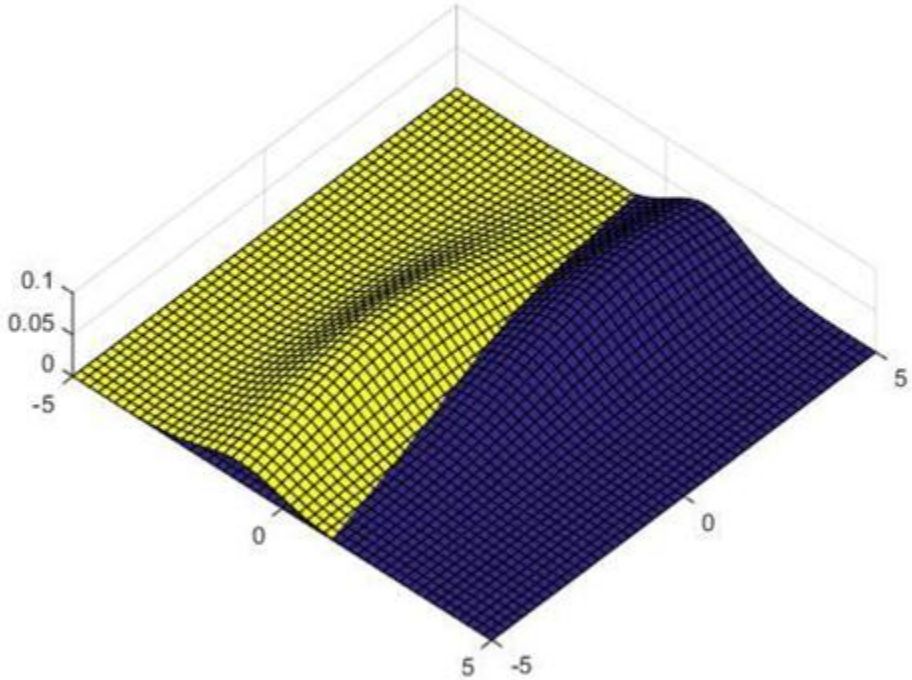


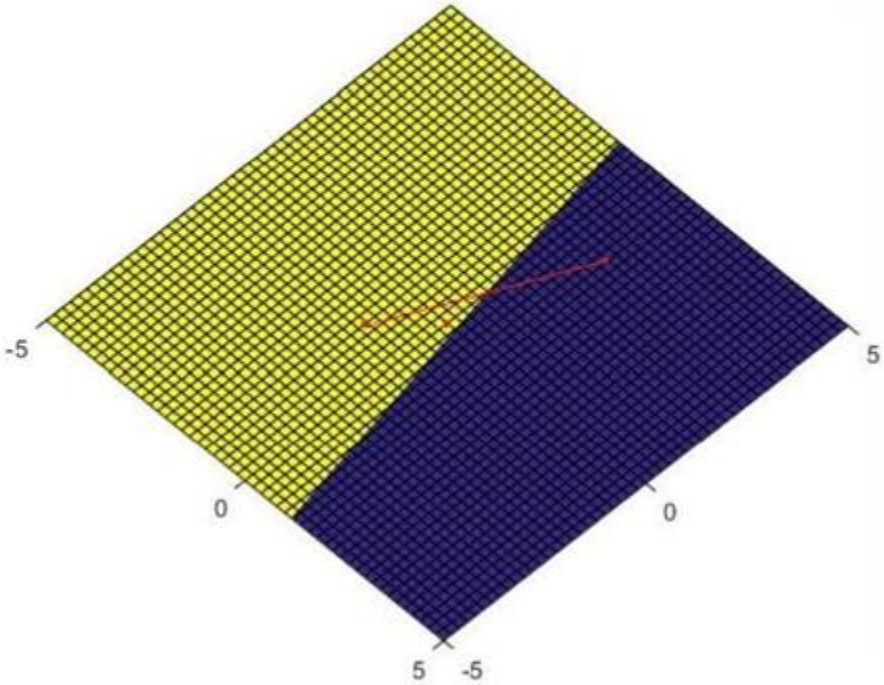
مرز تصمیم در یک مساله دو متغیره با احتمال پیشین برابر و ماتریس کواریانس $\Sigma = \sigma^2 I$

فاصله اقلیدسی تا مرکز هر کلاس می‌تواند معیار تعلق باشد. داده به کلاسی تعلق دارد که به مرکز آن نزدیکتر است.



مرز تصمیم در یک مساله دو متغیره با احتمال پیشین برابر و ماتریس کواریانس $\Sigma \neq \sigma^2 I$ و یکسان برای همه کلاس‌ها



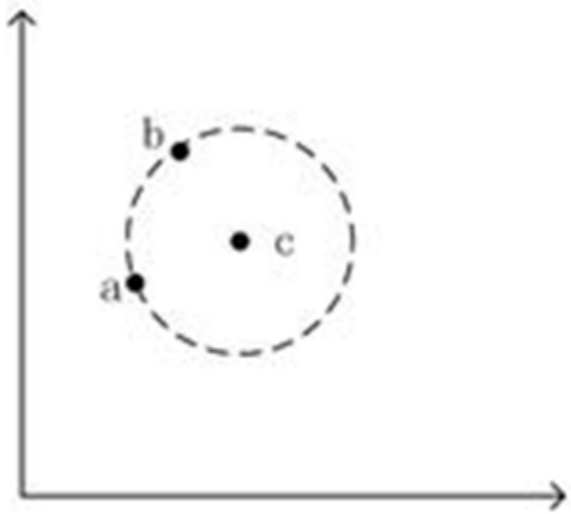


- مرز یک خط راست است. فاصله ماهالانوبیس تا مرکز هر کلاس می‌تواند معیار تعلق باشد.

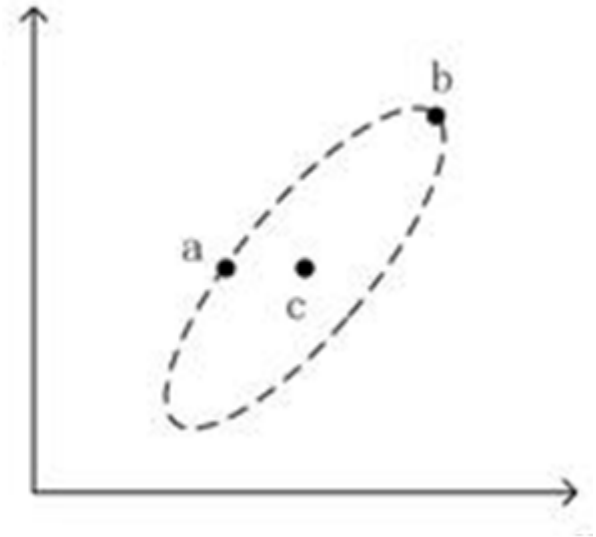
مرز تصمیم
جمع‌بندی

- اگر ماتریس کواریانس دو کلاس دلخواه باشد؛ مرز تصمیم منحنی می‌شود.
- اگر ماتریس کواریانس دو کلاس یکسان باشد، مرز تصمیم خط خواهد بود (فاصله ماهالانوبیس).
- اگر ماتریس کواریانس دو کلاس یکسان و برابر σI باشد، مرز تصمیم بر خط گذرنده از میانگین دو کلاس عمود خواهد بود (فاصله اقلیدسی).
- اگر ماتریس کواریانس دو کلاس یکسان و برابر σI و احتمال پیشین دو کلاس برابر باشد، مرز تصمیم عمود منصف خط گذرنده از میانگین دو کلاس خواهد بود. اگر احتمال پیشین برابر نباشد به کلاس با احتمال کمتر نزدیکتر است.

فاصله اقلیدسی و ماهالانوبیس



a and b at the same
Euclidean distance from c



a and b at the same
Mahalanobis distance from c

$$(b-c)^T \Sigma^{-1} (b-c)$$

مثال

Given ω_1, ω_2 : $P(\omega_1) = P(\omega_2)$ and $p(\underline{x}|\omega_1) = N(\underline{\mu}_1, \Sigma)$,

$$p(\underline{x}|\omega_2) = N(\underline{\mu}_2, \Sigma),\ \underline{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix},\ \underline{\mu}_2 = \begin{bmatrix} 3 \\ 3 \end{bmatrix},\ \Sigma = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$$

classify the vector $\underline{x} = \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix}$ using Bayesian classification:

- $\Sigma^{-1} = \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix}$
- Compute Mahalanobis d_m from μ_1, μ_2 : $d^2{}_{m,1} = [1.0, \quad 2.2]$

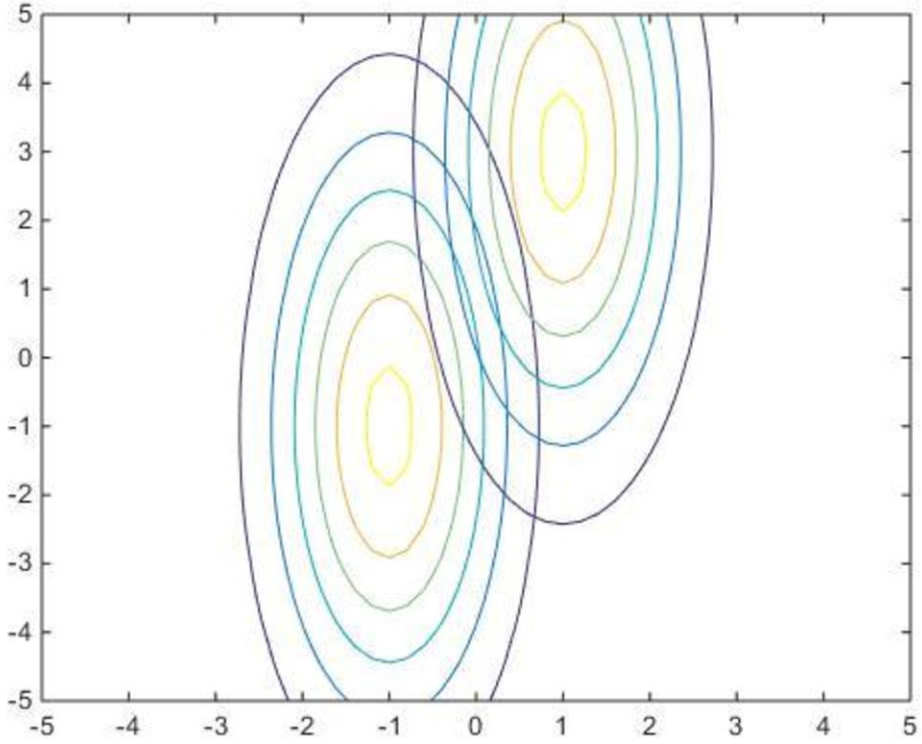
$$\Sigma^{-1}\begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix} = 2.952,\ d^2{}_{m,2} = [-2.0, \quad -0.8]\,\Sigma^{-1}\begin{bmatrix} -2.0 \\ -0.8 \end{bmatrix} = 3.672$$

- Classify $\underline{x} \rightarrow \omega_1$.

Observe that $d_{E,2} < d_{E,1}$

واریانس متفاوت برای ویژگی‌ها

بیضی های هم‌مرکز نقاط باتابع چگالی احتمال برابر را نشان می‌دهند



شناسایی الگو
فصل دوم
تخمین پارامترهای تابع چگالی احتمال

Textbook: Pattern Recognition, S. Theodoridis , K. Koutroumbas, Fourth Edition, 2009.

تخمین پارامترهای تابع چگالی احتمال

- تا اینجا فرض شد که تابع چگالی احتمال به همراه پارامترهای آن شناخته شده است.
- در بسیاری از موارد این شرط صادق نیست.
- در بسیاری از مسایل، پارامترهای تابع چگالی احتمال، باید از داده‌های موجود (داده‌های آموزشی) تخمین زده شود.
- به عبارت دیگر نوع توزیع معلوم است؛ به طور مثال توزیع نرمال. اما پارامترهای آن مانند میانگین و انحراف معیار باید از داده‌های آموزش تعیین شود.

روش‌های تخمین پارامترهای تابع چگالی احتمال

- Maximum Likelihood Estimation (MLE)
  - تخمین درست‌نمایی بیشینه
- Maximum a Posteriori probability estimation (MAP)
  - تخمین بیشینه احتمال پسین

تفاوت آمار و احتمال

- وقتی از احتمال صحبت می‌کنیم، منظور این است که میزان احتمال توزیع همه نمونه‌ها در کل را می‌دانیم.
  - مثال

یک جعبه شامل ۳ سیب و ۷ پرتغال است. محاسبه احتمال توزیع سیب در این جعبه برابر ۰.۳ و پرتغال ۰.۷ خواهد بود.

- در آمار، فقط به بخشی از جهان دسترسی داریم. می‌خواهیم از این بخش در دسترس استفاده کنیم و توزیع احتمال نمونه‌ها را در کل تخمین بزنیم.
- مشخصه‌های توزیع واقعی کلی، را پارامتر گویند و با θ نمایش می‌دهند.
- مشخصه‌های آماری را با $\hat{\theta}$ نمایش می‌دهند.

MAXIMUM LIKELIHOOD ESTIMATION
تخمین درست‌نمایی بیشینه

- در این حالت تابع توزیع چگالی معین است، اما پارامترهای آن معلوم نیست. توضیح بالا به صورت زیر نمایش داده می‌شود:

$$p(x|\omega_i;\theta)$$

- یک مجموعه نمونه $X=\{x_1,x_2,x_3,\dots,x_N\}$ در دسترس است که مستقل از هم در فرآیند انتخاب هستند. داریم:

$$p(X;\theta)=p(x_1,x_2,x_3,\dots,x_n;\theta)=\prod_{k=1}^{n}p(x_k;\theta)$$

- رابطه بالا را تابع درست‌نمایی پارامتر θ با توجه به مجموعه داده X می‌نامند.

MAXIMUM LIKELIHOOD METHOD
روش درست‌نمایی بیشینه

- روش درست‌نمایی بیشینه، پارامتر θ را به صورتی تخمین می‌زند که تابع درست‌نمایی، مقدار بیشینه خود را به دست آورد.

$$\hat{\theta}_{ML} = arg \max_{\theta} \prod_{k=1}^{N} p(x_k; \theta)$$

- از آنجا که تابع لگاریتم یکنوا است، می‌توان برای یافتن نطقه بیشینه از تابع بالا لگاریتم گرفت و مشتق آن را برابر صفر قرار داد.
- برای توزیع نرمال، در نهایت به مقادیر زیر می‌رسیم:

$$\hat{\sigma}_{ML}^2 = \frac{1}{N} \sum_{k=1}^{N} (x_k - \mu)^2$$

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_{k=1}^{N} x_k$$

Maximum a Posteriori probability estimation (MAP)
تخمین بیشینه احتمال پسین

- در این حالت پارامترها، θ، به عنوان متغیر تصادفی در نظر گرفته می‌شوند که مقادیر آنها در شرایطی که تعدادی نمونه

$$x_1, x_2, x_3, \dots, x_N$$

اتفاق افتاده است باید محاسبه شوند. به عبارتی به دنبال مقدار زیر هستیم:

$$p(\theta|X)$$

- با توجه به قانون بیز داریم:

$$p(\theta)p(X|\theta) = p(X)p(\theta|X)$$

$$p(\theta|X) = \frac{p(\theta)p(X|\theta)}{p(X)}$$

Maximum a Posteriori probability estimation (MAP)
تخمین بیشینه احتمال پسین

$$p(\theta|X) = \frac{p(\theta)p(X|\theta)}{p(X)}$$

- تخمین بیشینه احتمال پسین به دنبال بیشینه کردن مقدار بالا است. برای یافتن این نقطه مشتق بالا را برابر صفر قرار می‌دهیم.

$$\hat{\theta}_{MAP}: \frac{\partial}{\partial\theta} p(\theta|X) = 0$$

$$\frac{\partial}{\partial\theta} p(\theta)p(X|\theta)=0$$