Machine Learning

6 – Bayes Classifier

Avisek Gupta
Postdoctoral Fellow, IAI, TCG CREST
avisek003@gmail.com

September 02, 2022

Bayes Classifier: General Definition

Let the data be $x \in \mathbb{R}^d$ on which $\alpha_i, \ i = 1, ..., a$, actions are possible. Each action α_i depend on the classification of x to one of c classes $\omega_j, \ j = 1, ..., c$. A loss function lambda $\lambda(\alpha_i|\omega_j)$, (or, λ_{ij}) is defined to measure the effectiveness of each action α_i, given the knowledge that x was drawn from class ω_j .

Let us define the Conditional Risk of performing α_i given x as $R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|w_j)P(w_j|x)$.

We wish to define a Decision Rule $\alpha(x)$ which minimizes the overall Risk defined as,

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

Bayes Classifier: General Definition

Let the data be $x \in \mathbb{R}^d$ on which $\alpha_i, \; i = 1, ..., a$, actions are possible. Each action α_i depend on the classification of x to one of c classes $\omega_j, \; j = 1, ..., c$. A loss function lambda $\lambda(\alpha_i|\omega_j)$, (or, λ_{ij}) is defined to measure the effectiveness of each action α_i, given the knowledge that x was drawn from class ω_j.

Let us define the Conditional Risk of performing α_i given x as $R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|w_j)P(w_j|x)$.

We wish to define a Decision Rule $\alpha(x)$ which minimizes the overall Risk defined as,

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

The Bayes Classifier follows the Bayes Decision Rule: To minimize R, select $\alpha(x)$ as the action α_i for which $R(\alpha_i|x)$ is minimized.

The resulting risk R^* is called the Bayes Risk.

Bayes Classifier: Specific Cases

Case (1): Binary Classification, no data x is observed.

Two classes: w_1, w_2

Action α_i defined as: Select class $w_i,\ i = 1, 2.$

Loss $\lambda_{ij} = \lambda(\alpha_i|w_j) = \begin{cases} 0 & , i = j \\ 1 & , i \neq j \end{cases}, i, j = 1, 2\,.$

The Conditional Risks that should be minimized:

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

Bayes Classifier: Specific Cases

Case (1): Binary Classification, no data x is observed.

Two classes: w_1, w_2

Action α_i defined as: Select class w_i, $i = 1, 2.$

Loss $\lambda_{ij} = \lambda(\alpha_i|w_j) = \begin{cases} 0 & , i = j \\ 1 & , i \neq j \end{cases}, i, j = 1, 2 .$

The Conditional Risks that should be minimized:

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

Equal to 0

Assume constant functions

Bayes Classifier: Specific Cases

Case (1): Binary Classification, no data x is observed.

Two classes: w_1, w_2

Action α_i defined as: Select class w_i, $i = 1, 2.$

Loss $\lambda_{ij} = \lambda(\alpha_i|w_j) = \begin{cases} 0 & , i = j \\ 1 & , i \neq j \end{cases}, \; i, j = 1, 2 \,.$

The Conditional Risks that should be minimized:

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

Bayes Decision Rule to attain the Bayes Risk R^* : Decide class w_1 if $P(w_1) > P(w_2)$, otherwise decide class w_2 .

Bayes Classifier: Specific Cases

Case (2): Binary Classification, data x is observed and follows class-conditional distributions given by the density functions: $p(x|w_1)$, $p(x|w_2)$.

Two classes: w_1, w_2

Action α_i defined as: Select class w_i, $i = 1, 2.$

Loss $\lambda_{ij} = \lambda(\alpha_i|w_j) = \begin{cases} 0 & , i = j \\ 1 & , i \neq j \end{cases}, \; i, j = 1, 2$.

The Conditional Risks that should be minimized:

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

Bayes Classifier: Specific Cases

Case (2): Binary Classification, data x is observed and follows class-conditional distributions given by the density functions: $p(x|w_1)$, $p(x|w_2)$.

Two classes: w_1, w_2

Action α_i defined as: Select class w_i, $i = 1, 2.$

Loss $\lambda_{ij} = \lambda(\alpha_i|w_j) = \begin{cases} 0 & , i = j \\ 1 & , i \neq j \end{cases}, \; i, j = 1, 2$.

The Conditional Risks that should be minimized:

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

Bayes Decision Rule to attain the Bayes Risk R^* :

Decide class w_1 if $p(x|w_1)\, P(w_1) \; > \; p(x|w_2)\, P(w_2)$,

otherwise decide class w_2 .

OR,

Bayes Decision Rule to attain the Bayes Risk R^* :

Decide class w_1 if $P(w_1|x) \; > \; P(w_2|x)$,

otherwise decide class w_2 .

Bayes Classifier: Specific Cases

Case (2): Binary Classification, data x is observed and follows class-conditional distributions given by the density functions: $p(x|w_1)$, $p(x|w_2)$.

Two classes: w_1, w_2

Action α_i defined as: Select class w_i, $i = 1, 2.$

Loss $\lambda_{ij} = \lambda(\alpha_i|w_j) = \begin{cases} 0 & , i = j \\ 1 & , i \neq j \end{cases}, \; i, j = 1, 2$.

Probability of misclassification (general definition):

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

For the Bayes Decision Rule below:

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

The Conditional Risks that should be minimized:

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

Bayes Decision Rule to attain the Bayes Risk R^* :

Decide class w_1 if $p(x|w_1)\, P(w_1) > p(x|w_2)\, P(w_2)$,

otherwise decide class w_2 .

OR,

Bayes Decision Rule to attain the Bayes Risk R^* :

Decide class w_1 if $P(w_1|x) > P(w_2|x)$,

otherwise decide class w_2 .

Bayes Classifier: Specific Cases

Case (3): c-class Classification, data x is observed and follows class-conditional distributions given by the density functions: $p(x|w_j)$, $j = 1, ..., c$.

c number of classes: $w_1, ..., w_c$

Action α_i defined as: Select class w_i

Loss: $\lambda_{ij} = \begin{cases} 0 & , i = j \\ 1 & , i \neq j \end{cases} \quad i, j = 1, ..., c$

The Conditional Risks that should be minimized:

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda_{ij} P(w_j|x) = \sum_{j \neq i} P(w_j|x) = 1 - P(w_i|x)$$

Bayes Classifier: Specific Cases

Case (3): c-class Classification, data x is observed and follows class-conditional distributions given by the density functions: $p(x|w_j)$, $j = 1, \ldots, c$.

c number of classes: $w_1, \ldots, w_c$

Action α_i defined as: Select class w_i

Loss: $\lambda_{ij} = \begin{cases} 0 & , i = j \\ 1 & , i \neq j \end{cases} \quad i, j = 1, \ldots, c$

The Conditional Risks that should be minimized:

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda_{ij} P(w_j|x) = \sum_{j \neq i} P(w_j|x) = 1 - P(w_i|x)$$

Bayes Decision Rule to attain the Bayes Risk R^* :

Decide class w_i where $P(w_i|x) > P(w_j|x) \; \forall j \neq i$.

OR,

Decide class w_i where $p(x|w_i) \, P(w_i) > p(x|w_j) \, P(w_j) \; \forall j \neq i$.

Bayes Classifier: Specific Cases

Case (3): c-class Classification, data x is observed and follows class-conditional distributions given by the density functions: $p(x|w_j)$, $j = 1, ..., c$.

c number of classes: $w_1, ..., w_c$

Action α_i defined as: Select class w_i

Loss: $\lambda_{ij} = \begin{cases} 0 & , i = j \\ 1 & , i \neq j \end{cases} \quad i, j = 1, ..., c$

Probability of misclassification (general definition):

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

For the Bayes Decision Rule below:

$P(error|x) = 1 - P(w_i|x)$, where $P(w_i|x) > P(w_j|x) \; \forall j \neq i$

The Conditional Risks that should be minimized:

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda_{ij} P(w_j|x) = \sum_{j \neq i} P(w_j|x) = 1 - P(w_i|x)$$

Bayes Decision Rule to attain the Bayes Risk R^* :

Decide class w_i where $P(w_i|x) > P(w_j|x) \; \forall j \neq i$.

OR,

Decide class w_i where $p(x|w_i)\,P(w_i) > p(x|w_j)\,P(w_j) \; \forall j \neq i$.

Gaussian Distributions

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

The distribution is completely specific by the parameters μ and σ, where

the expected value of x is $\mu = \int_{-\infty}^{+\infty} x\, p(x)\, dx$, and

the expected squared deviation of x from μ is $\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2\, p(x)\, dx$.

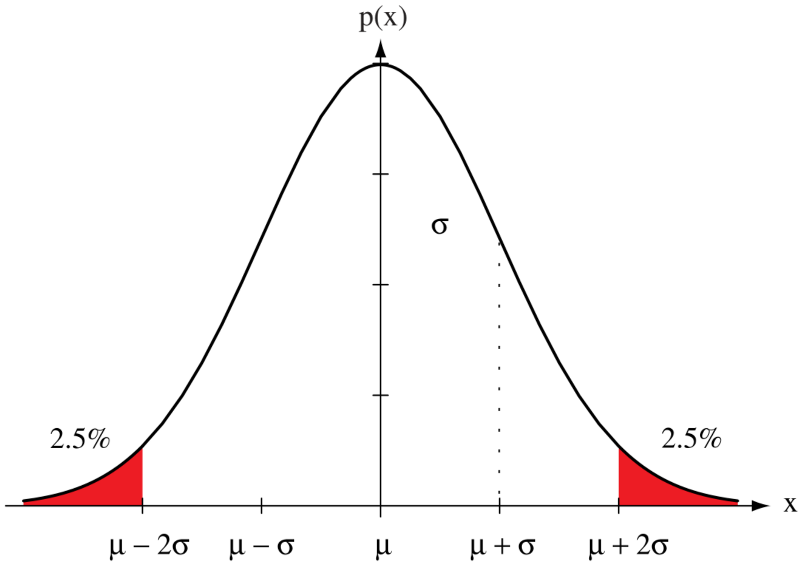


*Image Source: Duda Hart Stork - Pattern Classification

Gaussian Distributions

Multivariate Gaussian Density:

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

Here $\mathbf{x}, \mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ is symmetric and positive semi-definite.

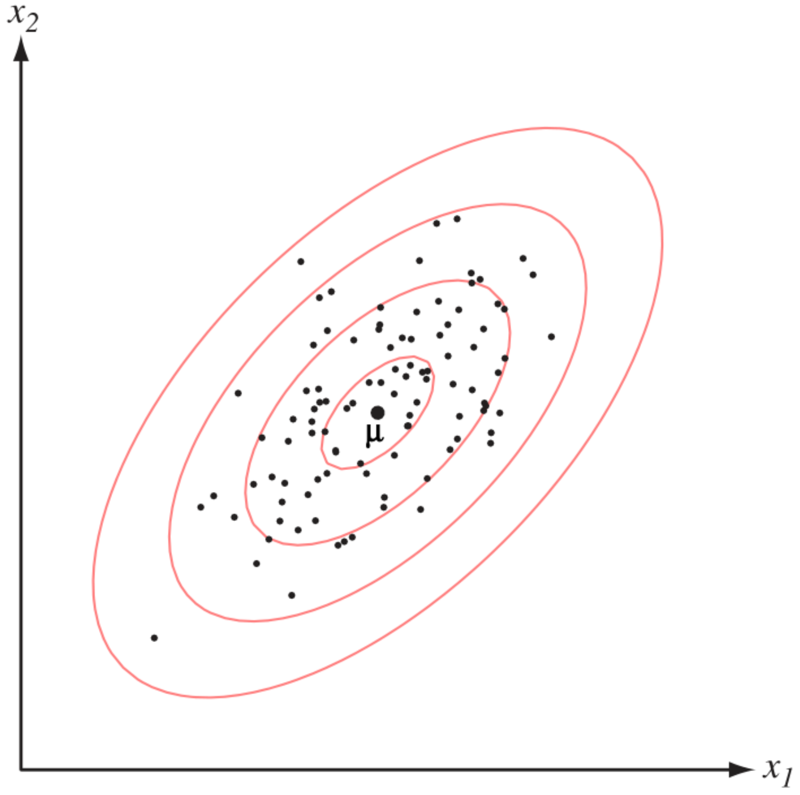


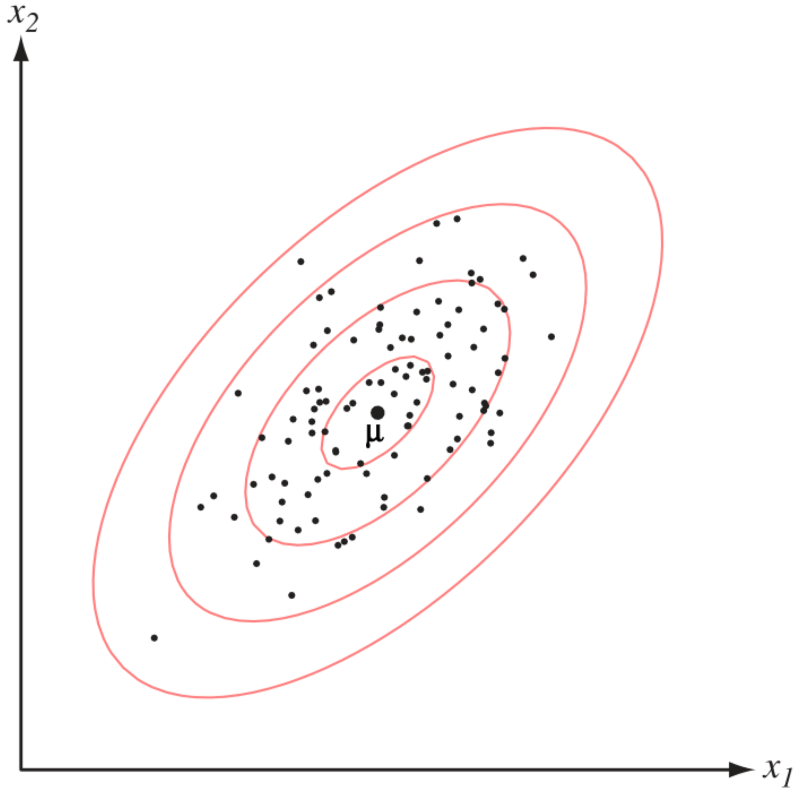
*Image Source: Duda Hart Stork - Pattern Classification

Gaussian Distributions

Multivariate Gaussian Density:

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

Here $\mathbf{x}, \mu \in \mathbb{R}^d,\ \Sigma \in \mathbb{R}^{d\times d}$ is symmetric and positive semi-definite.



Loci of points of constant density: $(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)$

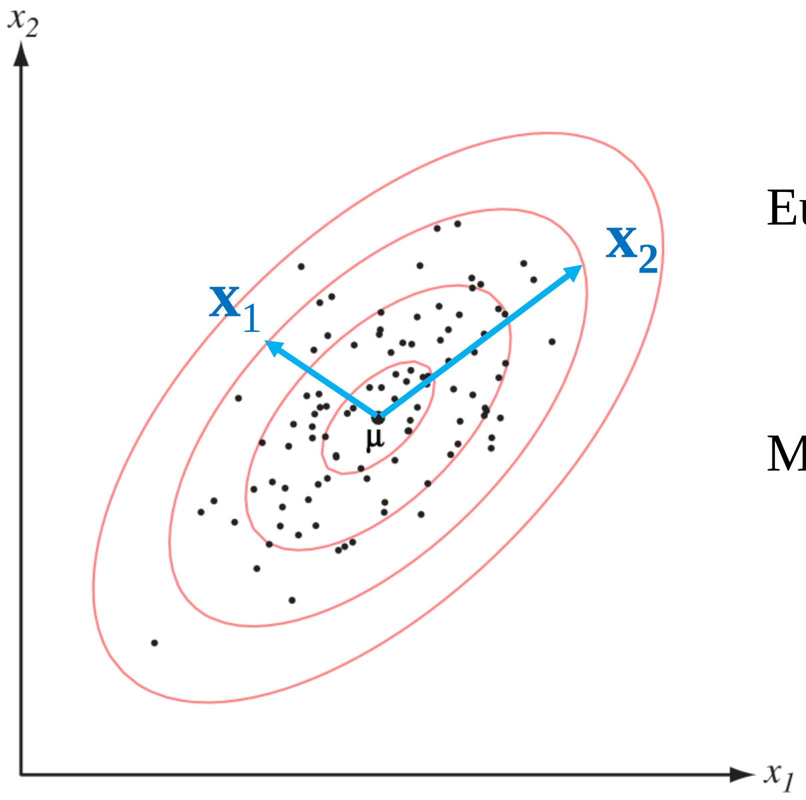
*Image Source: Duda Hart Stork - Pattern Classification

Gaussian Distributions

Multivariate Gaussian Density:

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

Here $\mathbf{x}, \mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$ is symmetric and positive semi-definite.



Euclidean distance:

$$||x_1 - \mu||^2 < ||x_2 - \mu||^2$$

Mahalanobis distance:

$$\Delta^2(x_1, \mu) = \Delta^2(x_2, \mu)$$

Mahalanobis Distance: $\Delta^2 = (\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)$

*Image Source: Duda Hart Stork - Pattern Classification

Binary Classification with Gaussian class-conditional densities

Recall the Bayes Decision Rule to attain the Bayes Risk R^*:

Decide class w_1 if $p(x|w_1)\,P(w_1) > p(x|w_2)\,P(w_2)$,

otherwise decide class w_2.

Writing in terms of a discriminant function $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$,

where $g_i(\mathbf{x}) = p(\mathbf{x}|w_i)\,P(w_i)$.

Decide class w_1 if $g(\mathbf{x}) > 0$, otherwise decide class w_2.

If $p(\mathbf{x}|w_i) \sim N(\mu, \Sigma_i)$, then

Binary Classification with Gaussian class-conditional densities

Recall the Bayes Decision Rule to attain the Bayes Risk R^*:

Decide class w_1 if $p(x|w_1)\,P(w_1) > p(x|w_2)\,P(w_2)$,

otherwise decide class w_2.

Writing in terms of a discriminant function $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$,

where $g_i(\mathbf{x}) = p(\mathbf{x}|w_i)\,P(w_i)$.

Decide class w_1 if $g(\mathbf{x}) > 0$, otherwise decide class w_2.

If $p(\mathbf{x}|w_i) \sim N(\mu, \Sigma_i)$, then

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{i}})^T \Sigma_i^{-1} (\mathbf{x} - \mu_{\mathbf{i}}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

Binary Classification with Gaussian class-conditional densities

Recall the Bayes Decision Rule to attain the Bayes Risk R^*:

Decide class w_1 if $p(x|w_1)\,P(w_1) \;>\; p(x|w_2)\,P(w_2)$,

otherwise decide class w_2 .

Writing in terms of a discriminant function $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$,

where $g_i(\mathbf{x}) = p(\mathbf{x}|w_i)\,P(w_i)$.

Decide class w_1 if $g(\mathbf{x}) > 0$, otherwise decide class w_2 .

If $p(\mathbf{x}|w_i) \sim N(\mu, \Sigma_i)$, then

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_\mathbf{i})^T \Sigma_i^{-1} (\mathbf{x} - \mu_\mathbf{i}) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(w_i)$$

Case (1):

$$\Sigma_i = \sigma^2 I$$

Case (2):

$$\Sigma_i = \Sigma$$

Case (3):

Σ_i is an arbitrary symmetric psd matrix

Binary Classification with Gaussian class-conditional densities

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{i}})^T \Sigma_i^{-1} (\mathbf{x} - \mu_{\mathbf{i}}) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(w_i)$$

Case (1): $\Sigma_i = \sigma^2 I$

Binary Classification with Gaussian class-conditional densities

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{i}})^T \Sigma_i^{-1} (\mathbf{x} - \mu_{\mathbf{i}}) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(w_i)$$

Case (1): $\Sigma_i = \sigma^2 I$

The discriminant function reduces to:

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln\, P(\omega_i)$$

Binary Classification with Gaussian class-conditional densities

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^T \Sigma_i^{-1}(\mathbf{x}-\mu_\mathbf{i}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

Case (1): $\Sigma_i = \sigma^2 I$

The discriminant function reduces to:

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x}-\boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln\, P(\omega_i)$$

Expanding the norm:

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}[\mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu}_i^t\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\mu}_i] + \ln\, P(\omega_i)$$

Equivalently, this can be written as,

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0},$$ where,

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i$$, and, $$w_{i0} = \frac{-1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln\, P(\omega_i).$$

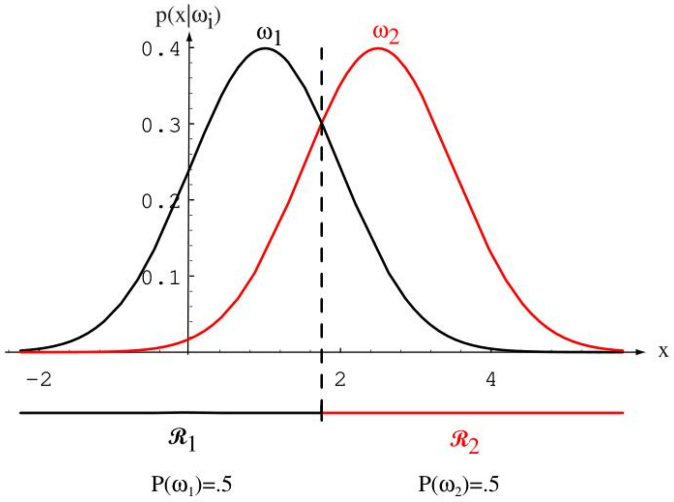
Binary Classification with Gaussian class-conditional densities

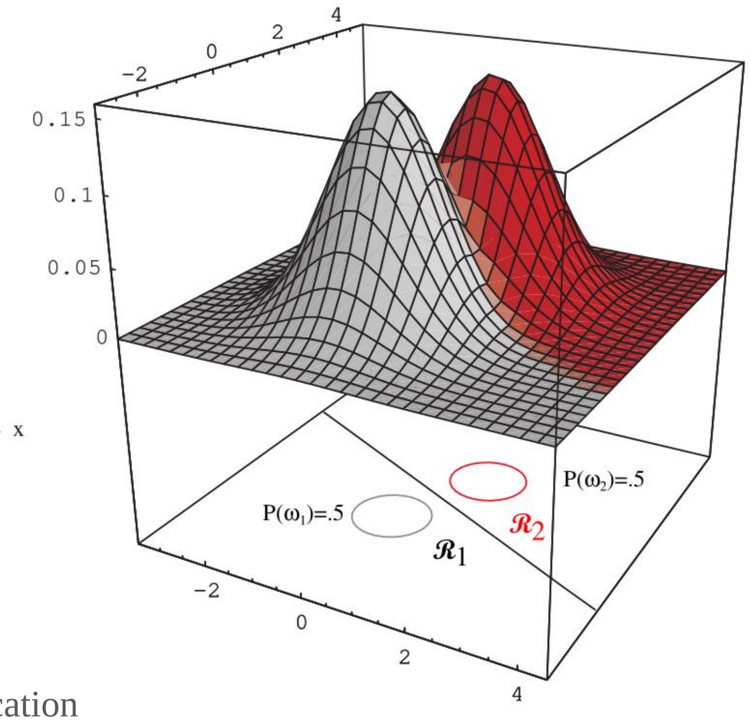
$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{i}})^T \Sigma_i^{-1}(\mathbf{x}-\mu_{\mathbf{i}}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

Case (1): $\Sigma_i = \sigma^2 I$

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0},$$ where,

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i$$, and, $$w_{i0} = \frac{-1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln\ P(\omega_i).$$





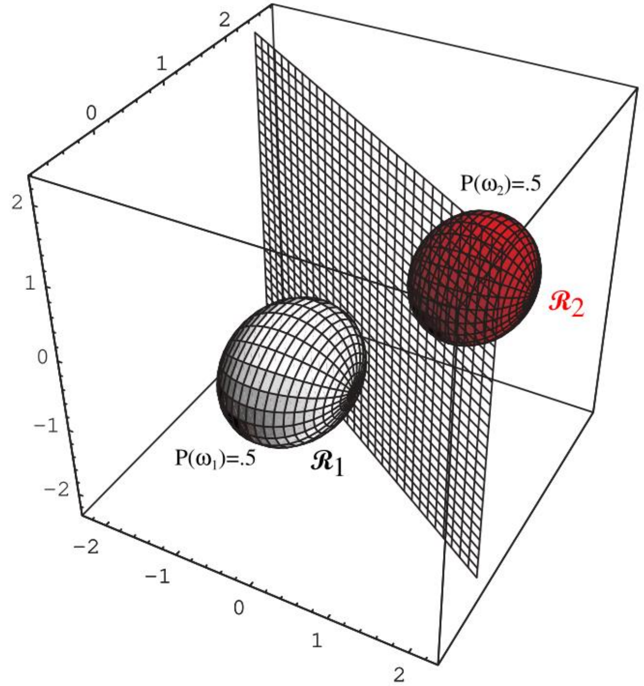


*Image Sources: Duda Hart Stork - Pattern Classification

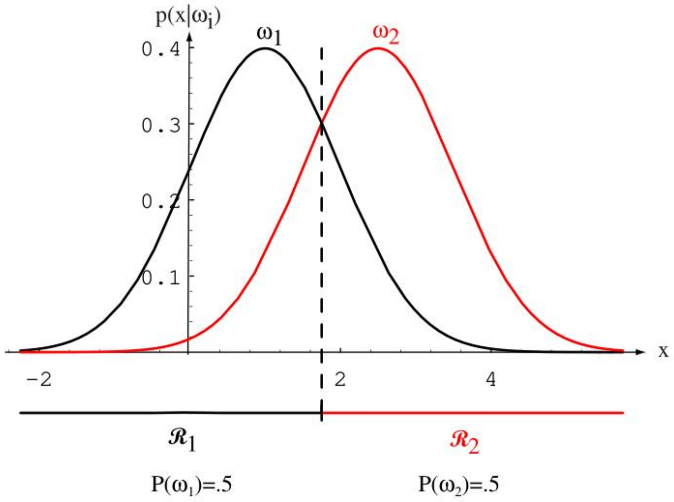
Binary Classification with Gaussian class-conditional densities

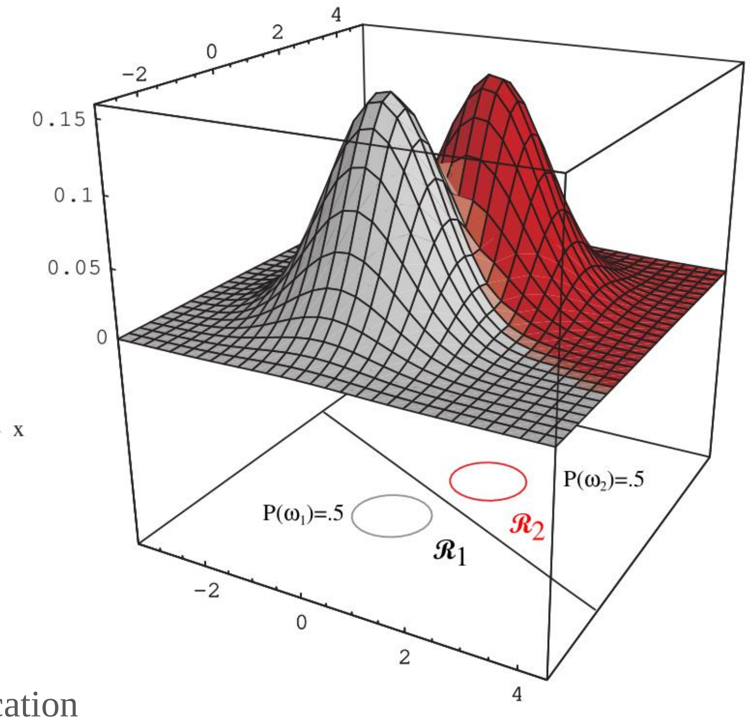
$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{i}})^T\Sigma_i^{-1}(\mathbf{x}-\mu_{\mathbf{i}}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

Case (1): $\Sigma_i = \sigma^2 I$

Considering $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$, the decision boundary: $\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0) = 0$

where, $\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$, and, $\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j).$





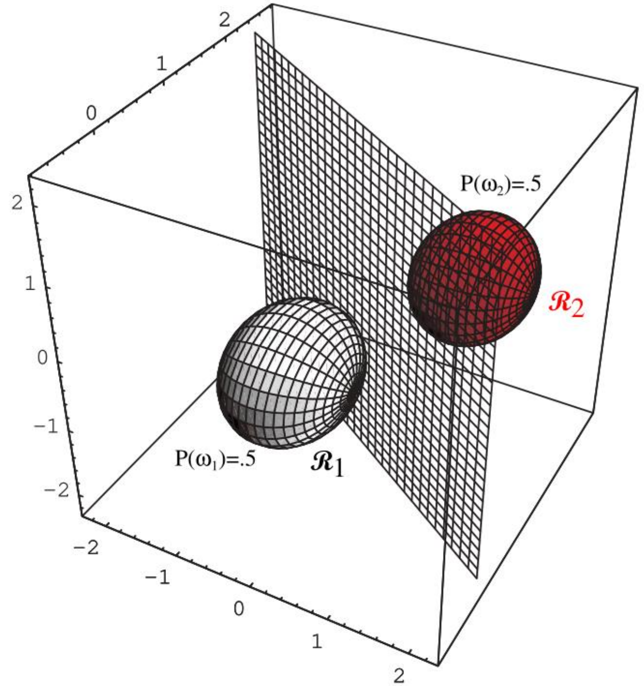


*Image Sources: Duda Hart Stork - Pattern Classification

Binary Classification with Gaussian class-conditional densities

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{i}})^T \Sigma_i^{-1} (\mathbf{x}-\mu_{\mathbf{i}}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

$P(w_1) = P(w_2)$?
$P(w_1) \neq P(w_2)$?

Case (1): $\Sigma_i = \sigma^2 I$

Considering $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$, the decision boundary: $\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0) = 0$

where, $\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$, and, $\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j).$

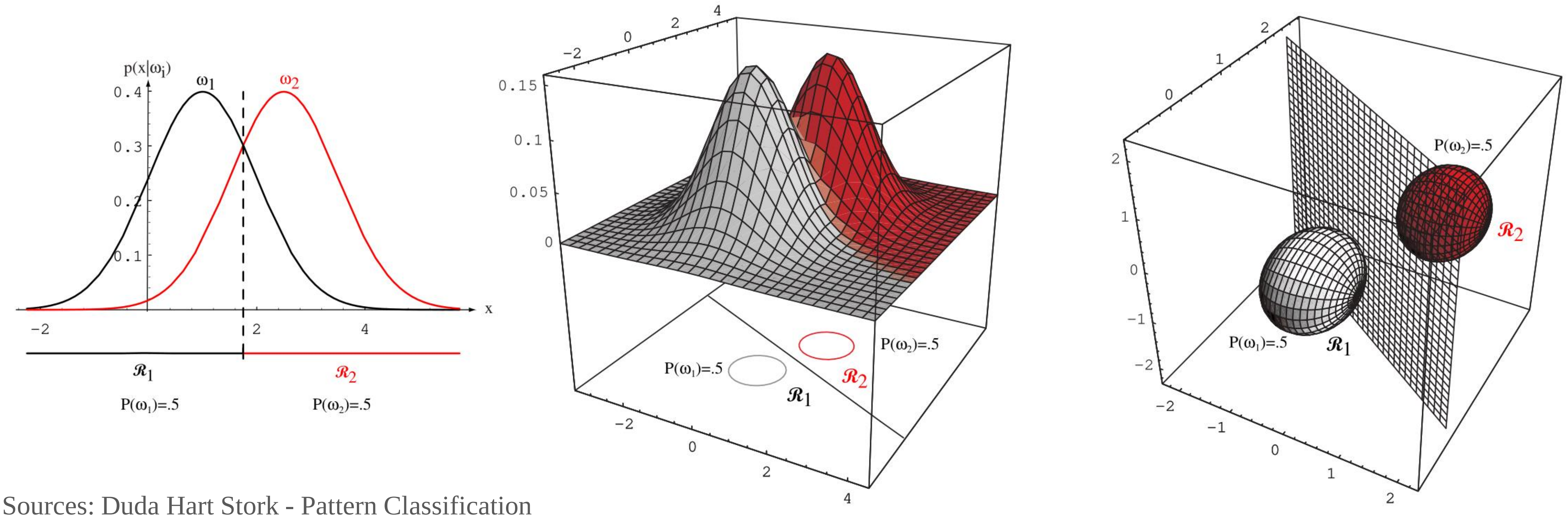


*Image Sources: Duda Hart Stork - Pattern Classification

Binary Classification with Gaussian class-conditional densities

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^T\Sigma_i^{-1}(\mathbf{x}-\mu_\mathbf{i}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

Case (2): $\Sigma_i = \Sigma$

$g_i(\mathbf{x})$ is reduced to $g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln\, P(\omega_i)$

If all priors are equal, $g_i(\mathbf{x}) = -\frac{1}{2}\Delta^2$. Therefore assign $\mathbf{x}$ to the class to which the Mahalanobis distance is minimum.

Expanding $(\mathbf{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$, we get $g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0},$

where, $\mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$, and, $w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln\, P(\omega_i).$

The decision boundary between two classes: $\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0) = 0,$

where, $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$. and,

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j).$$

Binary Classification with Gaussian class-conditional densities

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^T\Sigma_i^{-1}(\mathbf{x}-\mu_\mathbf{i}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

Case (2): $\Sigma_i = \Sigma$

The decision boundary between two classes: $\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0) = 0,$

where, $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$, and, $\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j).$

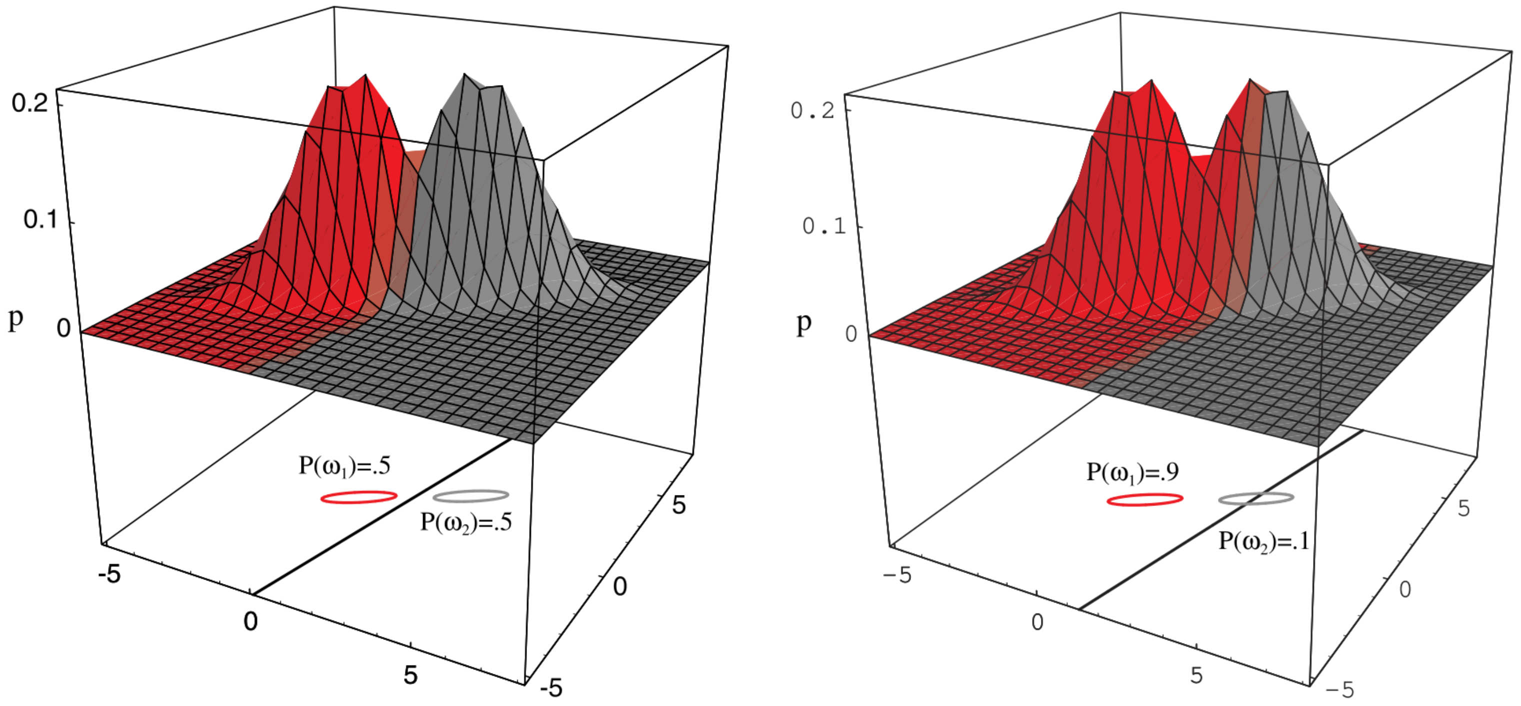


*Image Sources: Duda Hart Stork - Pattern Classification

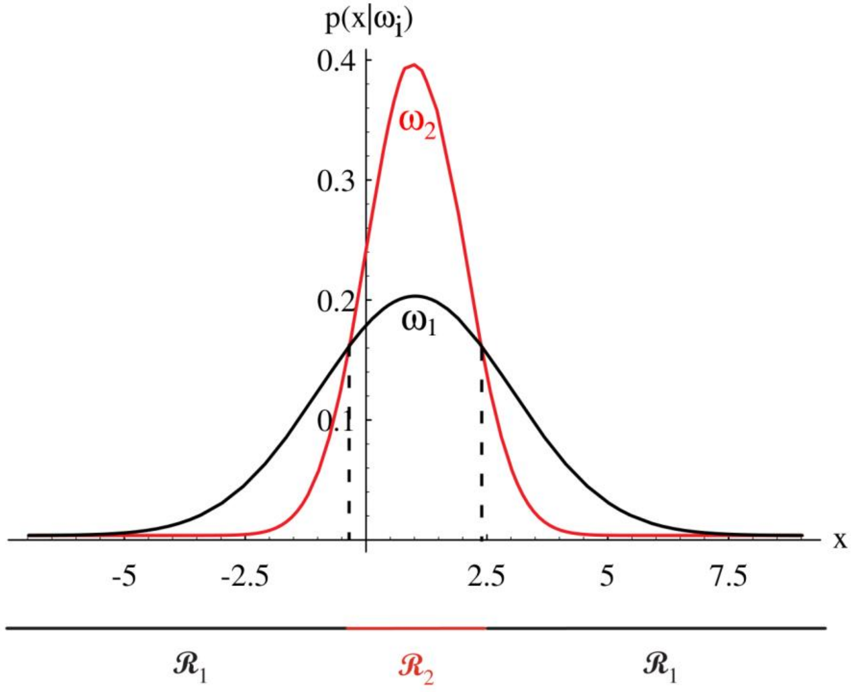
Binary Classification with Gaussian class-conditional densities

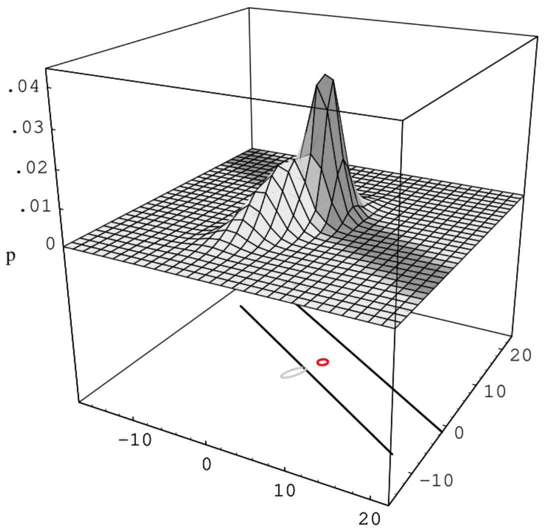
$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^T\Sigma_i^{-1}(\mathbf{x}-\mu_\mathbf{i}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

Case (3): Σ_i is an arbitrary symmetric psd matrix

$g_i(\mathbf{x})$ is reduced to $g_i(\mathbf{x}) = \mathbf{x}^t\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^t\mathbf{x} + w_{i0},$

where, $\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1},$ and, $\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$, and, $w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i).$





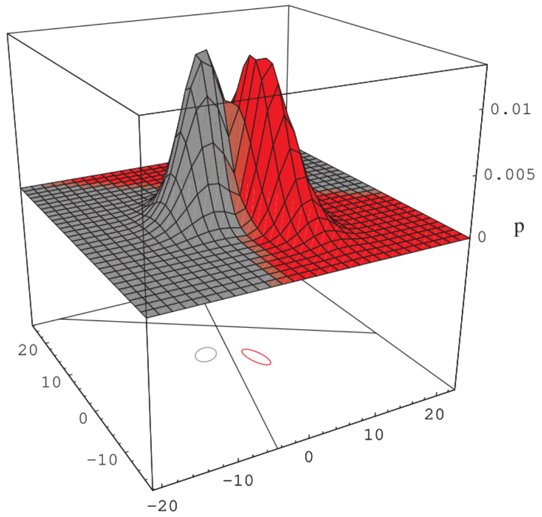


*Image Sources: Duda Hart Stork - Pattern Classification

Binary Classification with Gaussian class-conditional densities

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{i}})^T \Sigma_i^{-1} (\mathbf{x} - \mu_{\mathbf{i}}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

Case (3): Σ_i is an arbitrary symmetric psd matrix

$g_i(\mathbf{x})$ is reduced to $g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0},$

where, $\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1},$ and, $\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$, and, $w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i).$

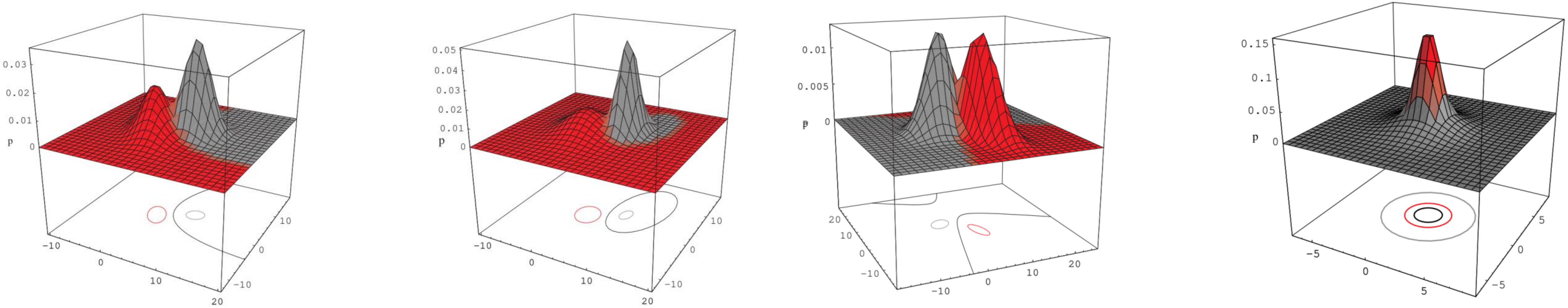


*Image Sources: Duda Hart Stork - Pattern Classification

Binary Classification with Gaussian class-conditional densities

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{i}})^T \Sigma_i^{-1}(\mathbf{x}-\mu_{\mathbf{i}}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

Case (3): Σ_i is an arbitrary symmetric psd matrix

$g_i(\mathbf{x})$ is reduced to $g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0},$

where, $\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1},$ and, $\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$, and, $w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i).$

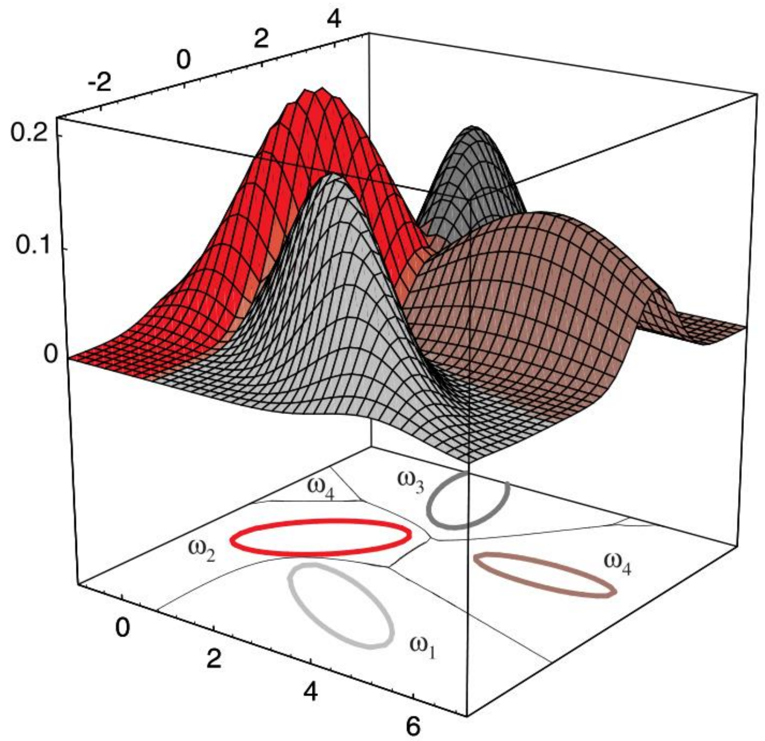


*Image Sources: Duda Hart Stork - Pattern Classification

Discriminative vs. Generative Models

Bayes Decision Rule to attain the Bayes Risk R^*:

Decide class w_i where

$$P(w_i|x) > P(w_j|x) \; \forall j \neq i$$

Classify based on posterior probabilities

Decide class w_i where

$$p(x|w_i)\, P(w_i) > p(x|w_j)\, P(w_j) \; \forall j \neq i$$

Classify based on class-conditional densities and prior probabilities

Discriminative vs. Generative Models

Bayes Decision Rule to attain the Bayes Risk R^*:

Decide class w_i where

$$P(w_i|x) > P(w_j|x) \; \forall j \neq i$$

Classify based on posterior probabilities

Decide class w_i where

$$p(x|w_i)\,P(w_i) > p(x|w_j)\,P(w_j) \; \forall j \neq i$$

Classify based on class-conditional densities and prior probabilities

What if all distributions are unknown?

Discriminative vs. Generative Models

Bayes Decision Rule to attain the Bayes Risk R^*:

Decide class w_i where

$$P(w_i|x) > P(w_j|x) \; \forall j \neq i$$

Estimate posterior probabilities

Decide class w_i where

$$p(x|w_i)\,P(w_i) > p(x|w_j)\,P(w_j) \; \forall j \neq i$$

Estimate (i) class-conditional densities and (ii) prior probabilities

Discriminative vs. Generative Models

Bayes Decision Rule to attain the Bayes Risk R^*:

Decide class w_i where

$$P(w_i|x) > P(w_j|x) \; \forall j \neq i$$

Estimate posterior probabilities

Discriminative Methods:

- Logistic Regression
- k-Nearest Neighbours
- Multi-Layered Perceptrons
- Support Vector Machines
- Random Forests
- …

Decide class w_i where

$$p(x|w_i)\,P(w_i) > p(x|w_j)\,P(w_j) \; \forall j \neq i$$

Estimate (i) class-conditional densities and (ii) prior probabilities

Generative Methods:

- Naive Bayes Classifier
- Hidden Markov Models
- Variational Autoencoders
- Generative Adversarial Networks
- …