ECEN 649 Pattern Recognition

Bayes Decision Theory

Ulisses Braga-Neto

ECE Department

Texas A&M University

Classification without Predictors

- This is the case where there are no measurements to base classification on.
- In such a case, the natural thing to do is to assign the most common label, that is, the one the has the highest *a-priori* probability $P(Y = i)$, for $i = 0, 1$:

$$\hat{Y} = \begin{cases} 0\,, & P(Y=0) \geq P(Y=1) \\ 1\,, & P(Y=1) > P(Y=0) \end{cases}$$

- Note that this is equivalent to assigning the label that is closest to the mean $E[Y] = P(Y = 1)$.
- This predictor has *classification error*

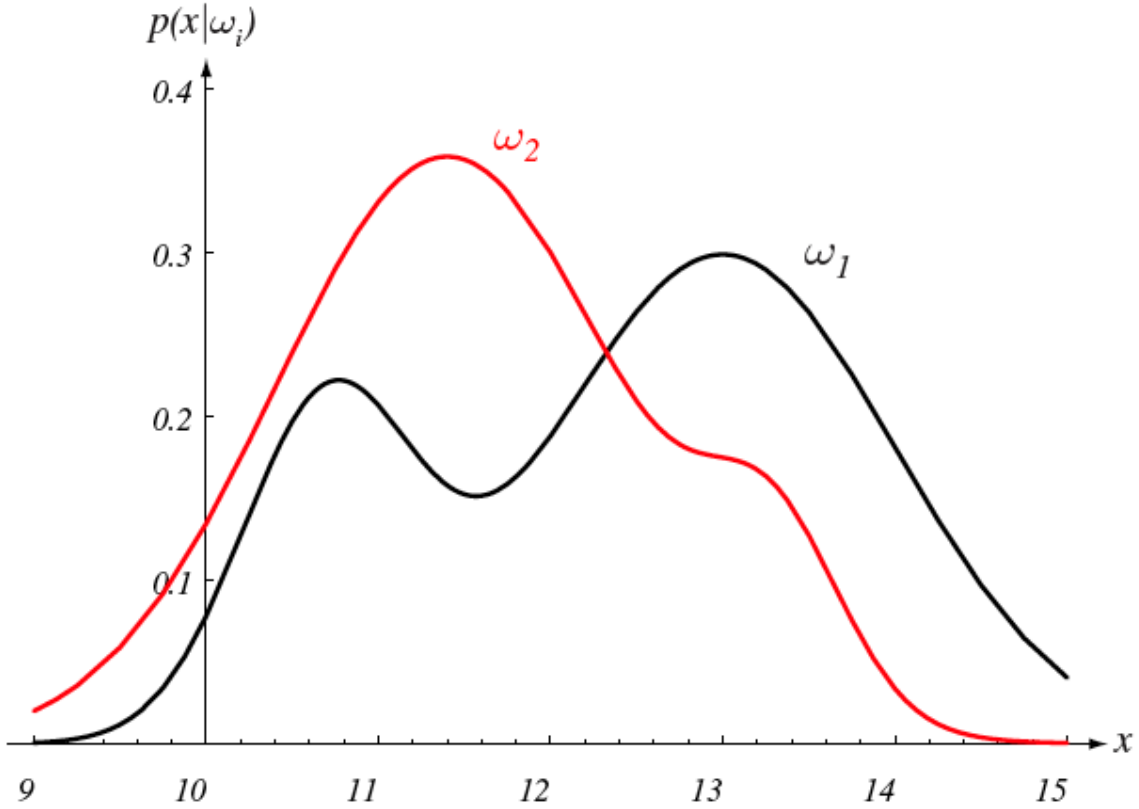
$$\epsilon = P(\hat{Y} \neq Y) = \min\{P(Y=0), P(Y=1)\}$$

Classification with Predictors

- There is something funny about the previous case: one will always call the same label.
- If one label is much more prevalent than the other, the classification error will be small, and this *may* be fine (however, think about the case of a test that always comes up negative for a rare disease).
- But if the labels are close to equally likely, the classification error will be close to 0.5 (flipping a coin).
- Luckily, this is a very rare scenario. We almost always have access to *predictor variables* $X \in R^d$ (also known as a *feature vector*) to help classification.

Class-Conditional Densities

The relative frequencies of each label as a function of predictor values are given by the *class-conditional densities* $p(x|Y = i)$, for $i = 0, 1$.



Posterior Probabilities

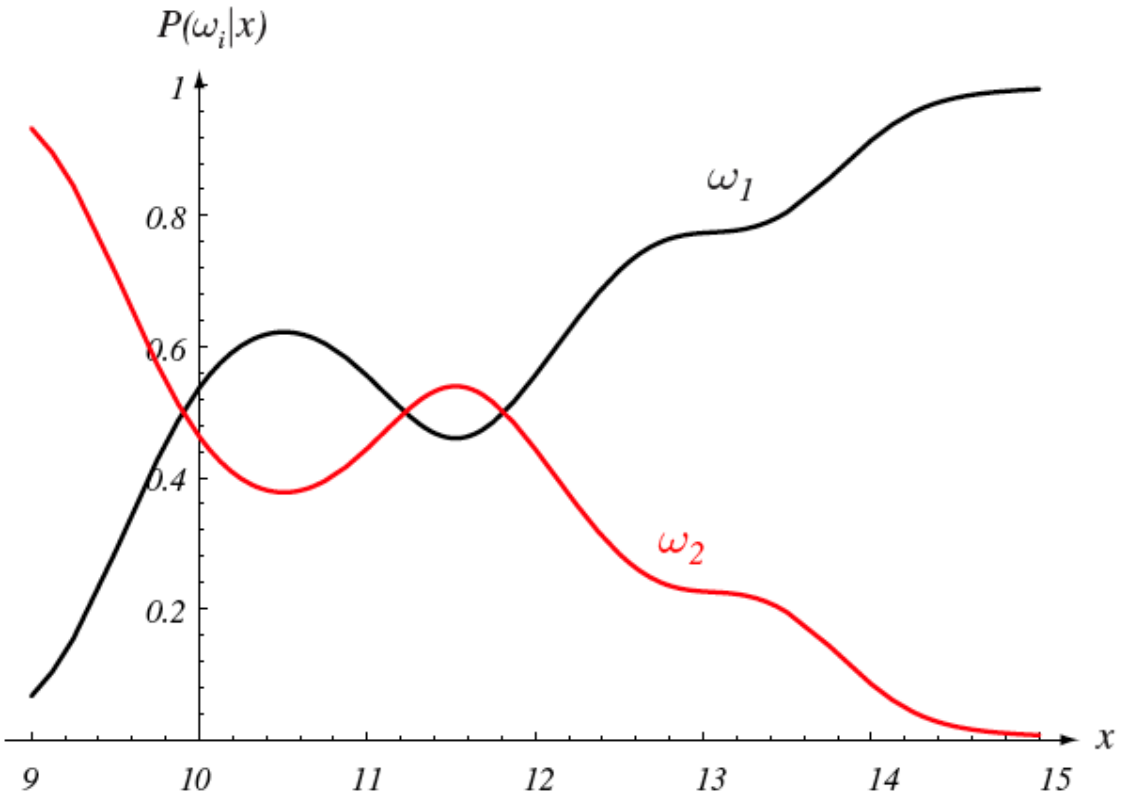
Using Bayes' theorem, we can start from the prior probabilities and class-conditional densities and find the posterior probability of $Y = i$ given that $X = x$ has been observed, for $i = 0, 1$:

$$P(Y = i|X = x) = \frac{p(x|Y = i)P(Y = i)}{p(x)}$$

$$= \frac{p(x|Y = i)P(Y = i)}{p(x|Y = 0)P(Y = 0) + p(x|Y = 1)P(Y = 1)}$$

Posterior Probabilities - II

Posterior probabilities are *not* probability densities (e.g., they do not integrate to 1) but are simply probabilities (in particular, they are always between 0 and 1).



Classifiers and Classification Error

- Formally, a *classifier* is a (measurable) function $\psi : R^d \to \{0, 1\}$ from the feature space R^d into the binary set of labels $\{0, 1\}$. Therefore, a classifier partitions the feature space into two regions.

- The *classification error* is the probability of misclassification:

$$\epsilon[\psi] = P(\psi(X) \neq Y)$$

- This is the fundamental criterion of performance in classification. The classification error is determined by the joint distribution F_{XY}, also called the *feature-label* distribution.

Conditional Classification Error

- The *conditional classification error* is the error at a particular observed value of X: $P(\psi(X) \neq Y|X = x)$.
- Notice that

$$\begin{aligned} P(\psi(X) \neq Y|X = x) &= P(\psi(X) = 0, Y = 1|X = x) \\ &\quad + P(\psi(X) = 1, Y = 0|X = x) \\ &= I_{\psi(x)=0}\, P(Y = 1|X = x) + I_{\psi(x)=1} P(Y = 0|X = x) \\ &= I_{\psi(x)=0}\, \eta(x) + I_{\psi(x)=1}(1 - \eta(x)) \end{aligned}$$

where the posterior probability function $\eta : R^d \to [0, 1]$,

$$\eta(x) = P(Y = 1|X = x)$$

plays a very important role in the sequel.

Conditional Classification Error - II

- The classification error is the "average" conditional classification error:

$$\begin{aligned}\epsilon[\psi] &= P(\psi(X) \neq Y) \\ &= \int_{x \in R^d} P(\psi(X) \neq Y | X = x) p(x)\, dx \\ &= E[P(\psi(X) \neq Y | X)]\end{aligned}$$

Therefore, knowing the error at each point $x \in R^d$ of the feature space, plus the "weight" $p(x)$, is enough to determine the overall classification error.

Classification Error

- Using the previous formulas, one can further develop the classification error as:

$$\epsilon[\psi] = \int_{x \in R^d} P(\psi(X) \neq Y | X = x) p(x)\, dx$$

$$= \int_{x \in R^d} \left(I_{\psi(x)=0}\, \eta(x) \; + \; I_{\psi(x)=1} (1 - \eta(x)) \right) p(x)\, dx$$

$$= \int_{\{x | \psi(x)=0\}} \eta(x) p(x)\, dx \; + \int_{\{x | \psi(x)=1\}} (1 - \eta(x)) p(x)\, dx$$

Classification Error - II

- Now, from Bayes theorem,

$$\begin{aligned} \eta(x)p(x) &= p(x \mid Y=1)P(Y=1) \\ (1-\eta(x))p(x) &= p(x \mid Y=0)P(Y=0) \end{aligned}$$

- Replacing these into the previous formula yields an alternative equation for the classification error:

$$\begin{aligned} \epsilon[\psi] = &\int_{\{x|\psi(x)=0\}} p(x \mid Y=1)P(Y=1)\,dx \\ &+ \int_{\{x|\psi(x)=1\}} p(x \mid Y=0)P(Y=0)\,dx \end{aligned}$$

Class-Specific Error Rates

- We can rewrite the previous equation as:

$$\epsilon[\psi] = (1-c)\epsilon^0[\psi] + c\epsilon^1[\psi]$$

where $c = P(Y=1)$, and

$$\epsilon^0[\psi] = \int_{\{x|\psi(x)=1\}} p(x \mid Y=0)\, dx$$

$$\epsilon^1[\psi] = \int_{\{x|\psi(x)=0\}} p(x \mid Y=1)\, dx$$

are the *class-specific* error rates. Given ψ, these error rates do not depend on the prior probabilities c and $1-c$, while the overall error $\epsilon[\psi]$ clearly does.

Testing Error Rates

- Suppose ψ is used as a *test* to distinguish "positive" cases (class 1) from "negative" cases (class 0).
- Then $\epsilon^0[\psi]$ and $\epsilon^1[\psi]$ are called the test's *false positive* and *false negative* error rates, respectively.
- One also defines the test's *sensitivity* and *specificity* as

$$\text{sensitivity} = 1 - \epsilon^1[\psi] = \int_{\{x|\psi(x)=1\}} p(x \mid Y = 1)\, dx$$

$$\text{specificity} = 1 - \epsilon^0[\psi] = \int_{\{x|\psi(x)=0\}} p(x \mid Y = 0)\, dx$$

Optimal Classification

- Is there a best classifier for a given problem, i.e., a classifier with the minimum classification error possible? And how does one pick it?
- We said that the best predictor, in the MSE sense, of Y given $X = x$ is the conditional probability $E[Y|X = x]$ (or $E[Y]$ when there are no predictors). When $Y \in \{0, 1\}$, we have

$$E[Y|X = x] = P(Y = 1|X = x) = \eta(x)$$

- This will not do in this case, because $\eta(x)$ can assume any value in the interval $[0, 1]$, and a classifier can only take values 0 or 1.

Optimal Classification - II

- However, with the restriction $\psi(X) \in \{0, 1\}$, we have

$$\epsilon[\psi] = P(\psi(X) \neq Y) = E\left(|\psi(X) - Y|\right) = E\left(|\psi(X) - Y|^2\right) .$$

Therefore, a classifier that minimizes the classification error also minimizes the MSE.

- (DGL Theorem 2.1) The classifier with minimal error is

$$\psi^*(x) = \arg\max_i P(Y = i|X = x) = I_{\eta(x) > \frac{1}{2}} .$$

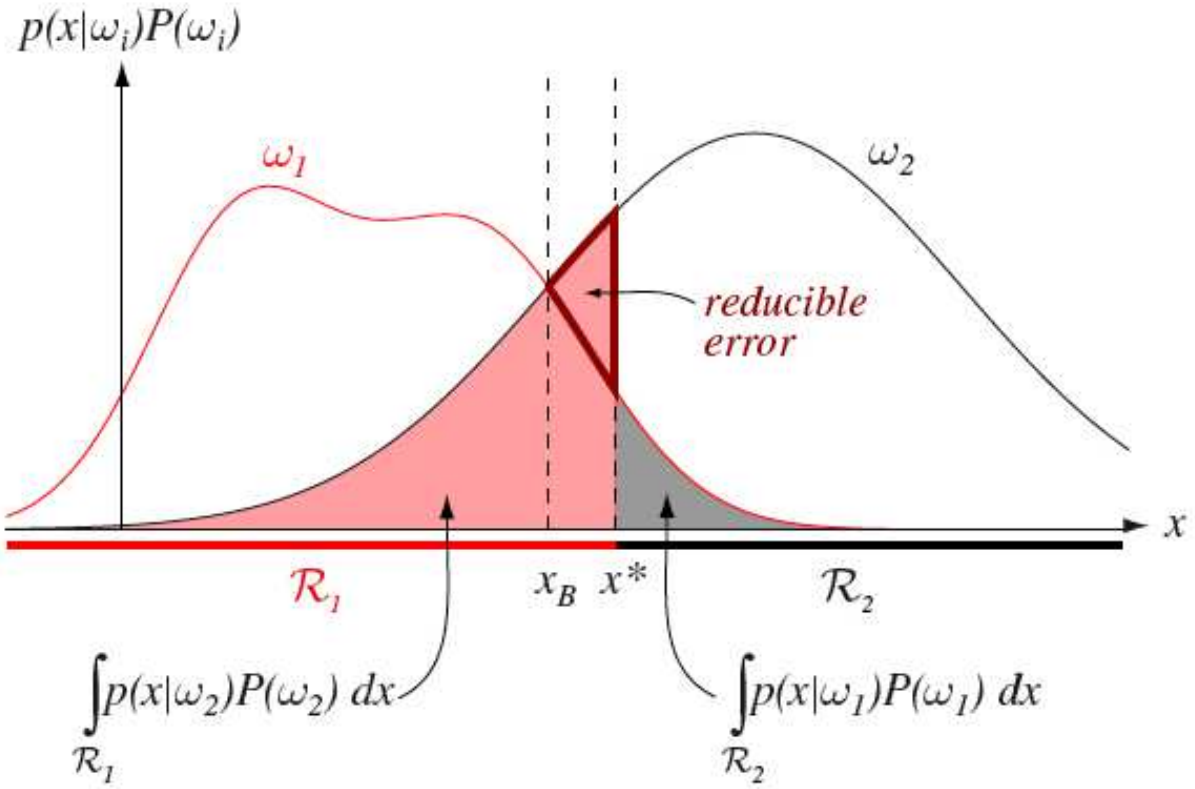
This is the MAP (Maximum A-Posteriori) classifier, more commonly known as the *Bayes classifier*.

Optimal Classification - III

- By Bayes theorem, we have, equivalently,

$$\psi^*(x) = \arg\max_i p(x|Y=i)P(Y=i)$$

- Graphical interpretation:



Bayes Error

- The error of the Bayes classifier $\epsilon^* = \epsilon[\psi^*]$ is a fundamental quantity in PR, known as the *Bayes error*.
- Note that the Bayes classifier is given by

$$\psi^*(x) = \begin{cases} 1\,, & \eta(x) > 1-\eta(x) \quad (\Leftrightarrow \eta(x) > \frac{1}{2}) \\ 0\,, & \eta(x) \leq 1-\eta(x) \quad (\Leftrightarrow \eta(x) \leq \frac{1}{2}) \end{cases}$$

- Therefore

$$\epsilon^* = \int_{\{x|\eta(x)<1-\eta(x)\}} \eta(x)p(x)\,dx + \int_{\{x|\eta(x)\geq 1-\eta(x)\}} (1-\eta(x))p(x)\,dx$$

$$= E[\min\{\eta(X), 1-\eta(X)\}]$$

Bayes Error - II

- Using the identity

$$\min\{a, 1-a\} = \frac{1}{2} - \frac{1}{2}|2a-1|\,, \quad 0 \le a \le 1$$

It follows that

$$\epsilon^* = \frac{1}{2} - \frac{1}{2}E[|2\eta(X)-1|]$$

- In particular, we always have $\epsilon^* \le \frac{1}{2}$.

Example

Example

Example

Example

Bayes Decision Theory

- Suppose that upon observing $X = x$ one takes an *action* $\alpha(x)$ in a finite set of a possible actions

$$\alpha(x) \in \{\alpha_0, \alpha_1, \ldots, \alpha_{a-1}\}$$

- Suppose there are c *states of nature* (i.e., classes), $Y \in \{0, 1, \ldots, c-1\}$. Each action incurs a *loss*

$$\lambda_{ij} = \text{ cost of taking action } \alpha_i \text{ when true state of nature is } j$$

- Action i may be simply deciding that the true state of nature is i, but we may have $a > c$, in which case one of the extra actions might be *rejecting* to make a decision.
- The losses indicate, for example, the *cost* of making incorrect decisions.

Bayes Decision Theory - II

- The expected loss upon observing $X = x$ is

$$R[\alpha(x) = \alpha_i] = \sum_{j=0}^{c-1} \lambda_{ij} P(Y = j | X = x)$$

  This is called the *conditional risk* given $X = x$.
- The *overall risk* is given by

$$R = E[R(\alpha(X))] = \int_{x \in R^d} R(\alpha(x)) p(x)\, dx$$

- To minimize R, we select $\alpha(x) = \alpha_i$ such that $R[\alpha(x) = \alpha_i]$ is minimum, *at each value* $x \in R^d$. This optimal strategy is called the *Bayes decision rule*, with corresponding optimal *Bayes risk* R^*.

Bayes Decision Theory - III

- In the special case that $a = c = 2$, that is, there are two classes and two actions, we have

$$\begin{aligned} R[\alpha(x) = \alpha_0] &= \lambda_{00} P(Y = 0|X = x) \\ &\quad + \lambda_{01} P(Y = 1|X = x) \quad (I) \\ R[\alpha(x) = \alpha_1] &= \lambda_{10} P(Y = 0|X = x) \\ &\quad + \lambda_{11} P(Y = 1|X = x) \quad (II) \end{aligned}$$

- We decide for action α_0 if $(II) > (I)$, that is, if

$$(\lambda_{10} - \lambda_{00}) P(Y = 0|X = x) > (\lambda_{01} - \lambda_{11}) P(Y = 1|X = x)$$

Bayes Decision Theory - IV

- Applying Bayes theorem (and assuming that $\lambda_{10} > \lambda_{00}$) allows us to write

$$\frac{p(x|Y=0)}{p(x|Y=1)} > \frac{\lambda_{01}-\lambda_{11}}{\lambda_{10}-\lambda_{00}} \frac{P(Y=1)}{P(Y=0)}$$

that is, we decide for action α_0 if the *likelihood ratio* on the left is larger than the given threshold.

- The case where

$$\lambda_{00} = \lambda_{11} = 0$$
$$\lambda_{10} = \lambda_{01} = 1$$

is called the *0-1 loss case*. This is the case that we had considered before, if action α_i is simply deciding that the state of nature is i, for $i = 0, 1$.

Discriminant Functions

- A classifier can be specified through a set of *discriminant functions* $\{g_i(x)|i = 0, 1, \ldots, c-1\}$ as:

$$\psi(x) \;=\; \arg\max_i \; g_i(x)$$

- The $i-th$ *decision region* is determined by

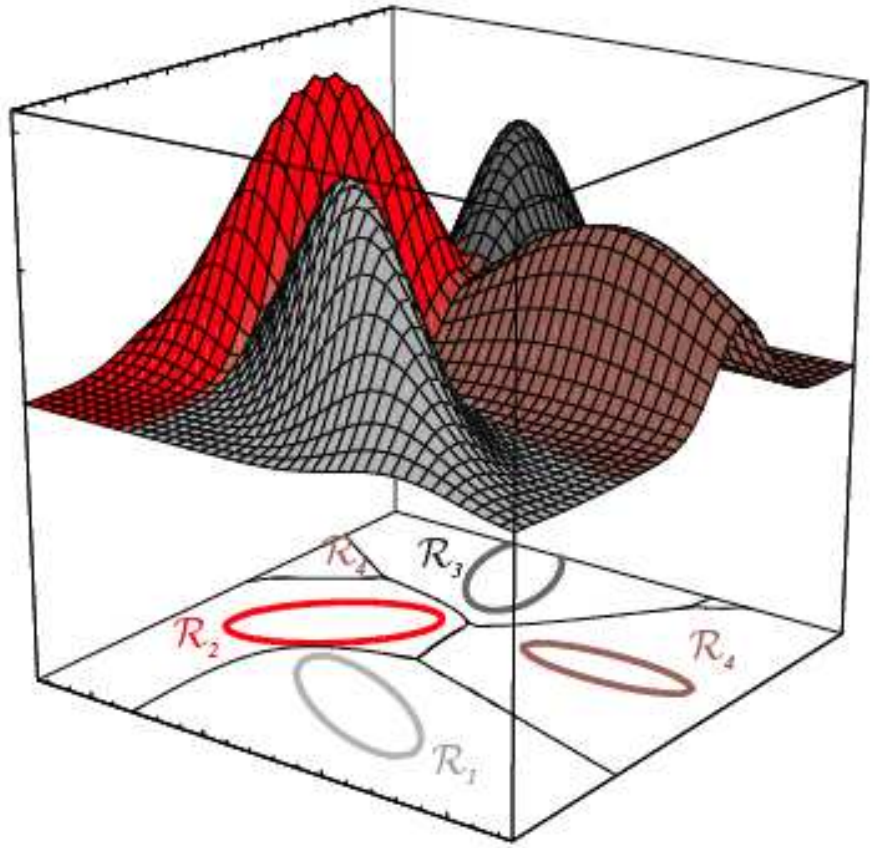
$$g_i(x) > g_j(x)\,, \quad \text{for all } i \neq j$$

  The loci of ties among largest discriminant functions determine the *decision surfaces*.

- A set of discriminant functions determines a unique classifier, but the converse is not true: the same classifier can be determined by multiple sets of discriminant functions.

Discriminant Functions - II

Graphical Example:



Discriminant Functions - III

- For the Bayes classifier, we have the following equivalent sets of discriminant functions:

$$
\begin{aligned}
g_i(x) &= -R[\alpha(x)]\,, \quad i = 0, 1, \ldots, c-1 \\
g_i(x) &= P(Y = i | X = x)\,, \quad i = 0, 1, \ldots, c-1 \\
g_i(x) &= p(x|Y = i)P(Y = i)\,, \quad i = 0, 1, \ldots, c-1
\end{aligned}
$$

- Monotonic transformations to the discriminant functions do not alter the classifier.
- For example, it is often useful to take logs and represent the Bayes classifier through the discriminant functions

$$
g_i(x) = \ln p(x|Y = i) + \ln P(Y = i)\,, \quad i = 0, 1, \ldots, c-1
$$

Discriminant Functions - IV

- In the two-category case, we can define a single discriminant function

$$g(x) = g_1(x) - g_0(x)$$

In which case the classifies is determined by

$$g(x) > 0 \Rightarrow \psi(x) = 1$$
$$g(x) \leq 0 \Rightarrow \psi(x) = 0$$

- For example, for the Bayes classifier

$$g(x) = P(Y = 1|X = x) - P(Y = 0|X = x)$$
$$g(x) = \ln \frac{p(x|Y = 1)}{p(x|Y = 0)} + \ln \frac{p(Y = 1)}{p(Y = 0)}$$

Gaussian Model

- Consider the case where the class-conditional densities are multivariate Gaussian densities:

$$p(x|Y=i) \sim N_d(\mu_i, \Sigma_i)\,, \quad i = 0, 1, \ldots, c-1$$

In other words,

$$p(x|Y=i) \;=\; (2\pi)^{-\frac{d}{2}} |\Sigma_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right)$$

for $i = 0, 1, \ldots, c-1$.

- This is a case of great interest in engineering and science.

Gaussian Model - II

- The Bayes classifier is specified through the discriminant functions

$$
\begin{aligned}
g_i(x) &= \ln p(x|Y=i) + \ln P(Y=i) \\
&= -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - \underbrace{\frac{d}{2}\ln 2\pi}_{=\text{cte (drop)}} \\
&\quad - \frac{1}{2}\ln|\Sigma_i| + \ln P(Y=i)
\end{aligned}
$$

for $i = 0, 1, \ldots, c-1$.

Gaussian Model - Nearest Mean Classifier

- Case 1: Equal spherical covariance matrices.
- In this case,

$$\Sigma_i = \sigma^2 I_d \;\Rightarrow\; \Sigma_i^{-1} = \frac{1}{\sigma^2} I_d\,, \quad i = 0, 1, \ldots, c-1$$

The constant term $\ln |\Sigma_i| = -2d \ln \sigma$ can be dropped, leading to

$$g_i(x) \;=\; -\frac{1}{2}\frac{||x - \mu_i||^2}{\sigma^2} \;+\; \ln P(Y = i)$$

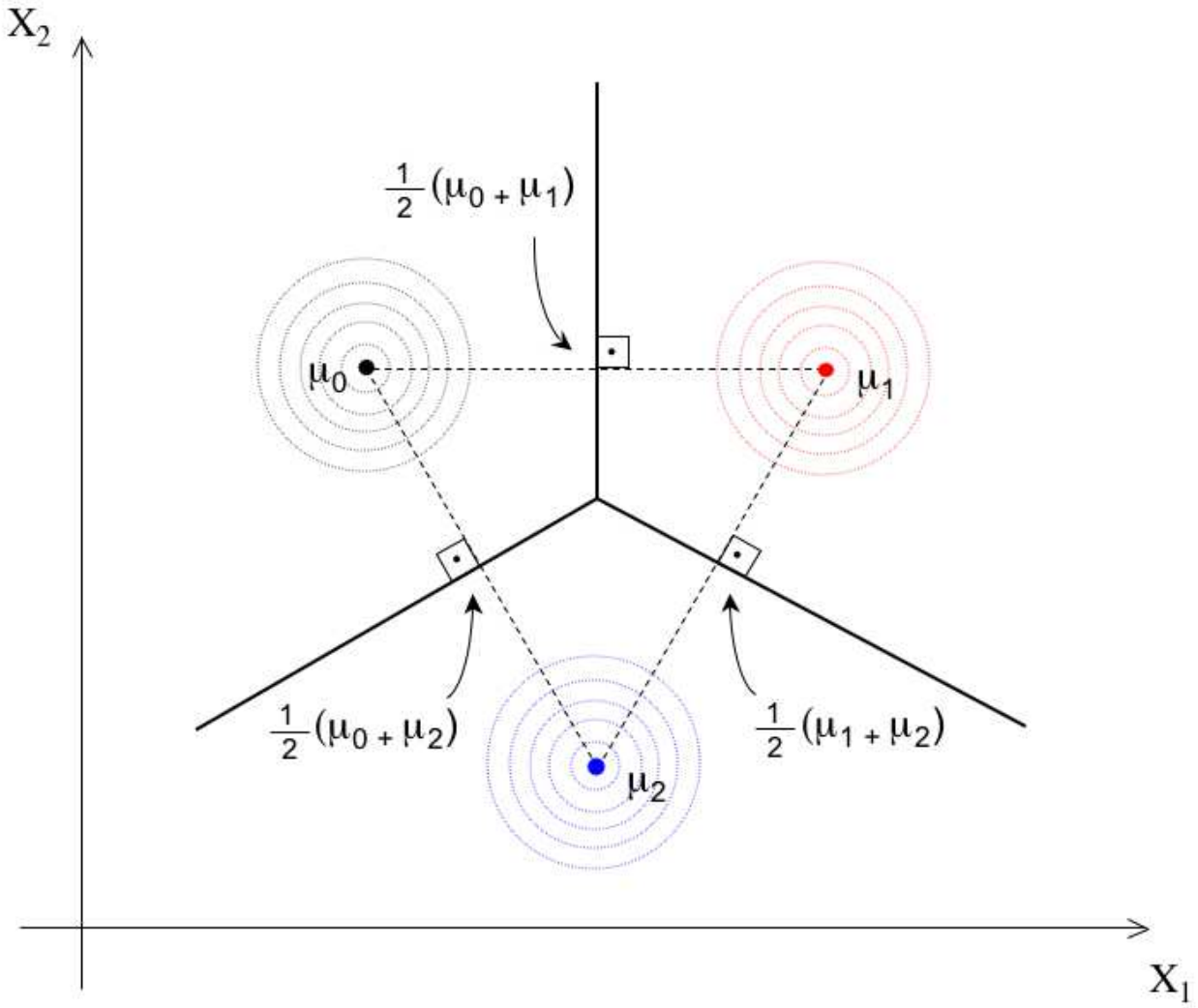
- If the classes are equally-likely, more terms can be dropped and we obtain:

$$g_i(x) \;=\; -||x - \mu_i||^2$$

This is called the optimal *Nearest-Mean Classifier*.

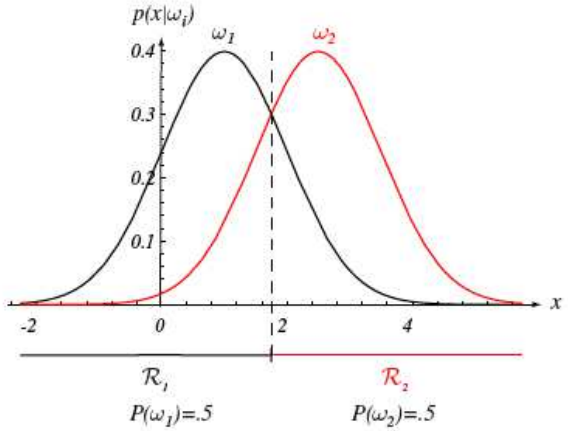
Gaussian Model - Nearest Mean Classifier

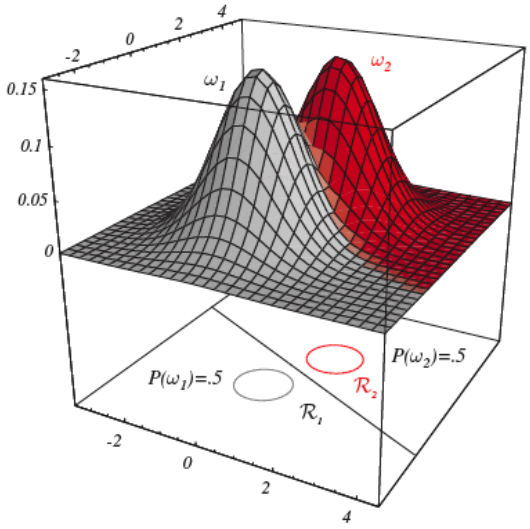
Example of optimal NMC (This is also known as a *Voronoi diagram*):

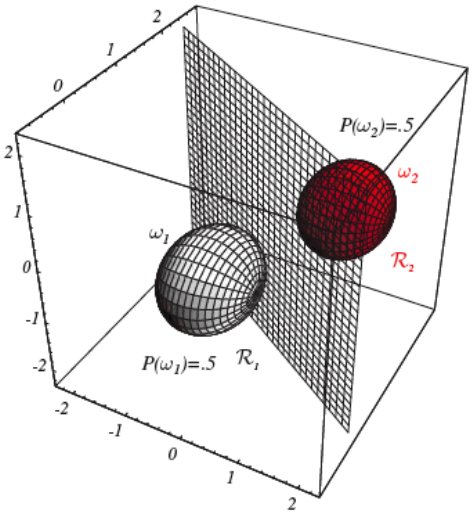


Gaussian Model - Nearest Mean Classifier

More Examples of NMC:







Gaussian Model - Linear Discriminant

- Case 2: Equal arbitrary covariance matrices.

- In this case,

$$\Sigma_i = \Sigma\,, \quad i = 0, 1, \ldots, c-1$$

Once again, the constant term $\ln|\Sigma_i| = \ln|\Sigma|$ can be dropped, resulting in

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i) + \ln P(Y=i)$$

Gaussian Model - Linear Discriminant

- This is apparently a quadratic discriminant; however the quadratic term is $x^T\Sigma^{-1}x$, which is constant and can be dropped, leading to a *linear discriminant*:

$$g_i(x) = a_i^T x + b_i$$

where

$$a_i = \Sigma^{-1}\mu_i$$

$$b_i = -\frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \ln P(Y = i)$$

Gaussian Model - Linear Discriminant

- In the case $c = 2$, the single discriminant is given by:

$$g(x) = g_1(x) - g_0(x) = a^T x + b$$

where

$$a = \Sigma^{-1}(\mu_1 - \mu_0)$$

$$b = -\frac{1}{2}(\mu_1 - \mu_0)^T \Sigma^{-1}(\mu_1 + \mu_0) + \ln \frac{P(Y=1)}{P(Y=0)}$$

- Clearly, the equation $g(x) = 0$ defines a hyperplane decision boundary.

Gaussian Model - Linear Discriminant

- If the classes are equally-likely, then the hyperplane is given by:

$$g(x) = a^T(x - x_0) = 0$$

where

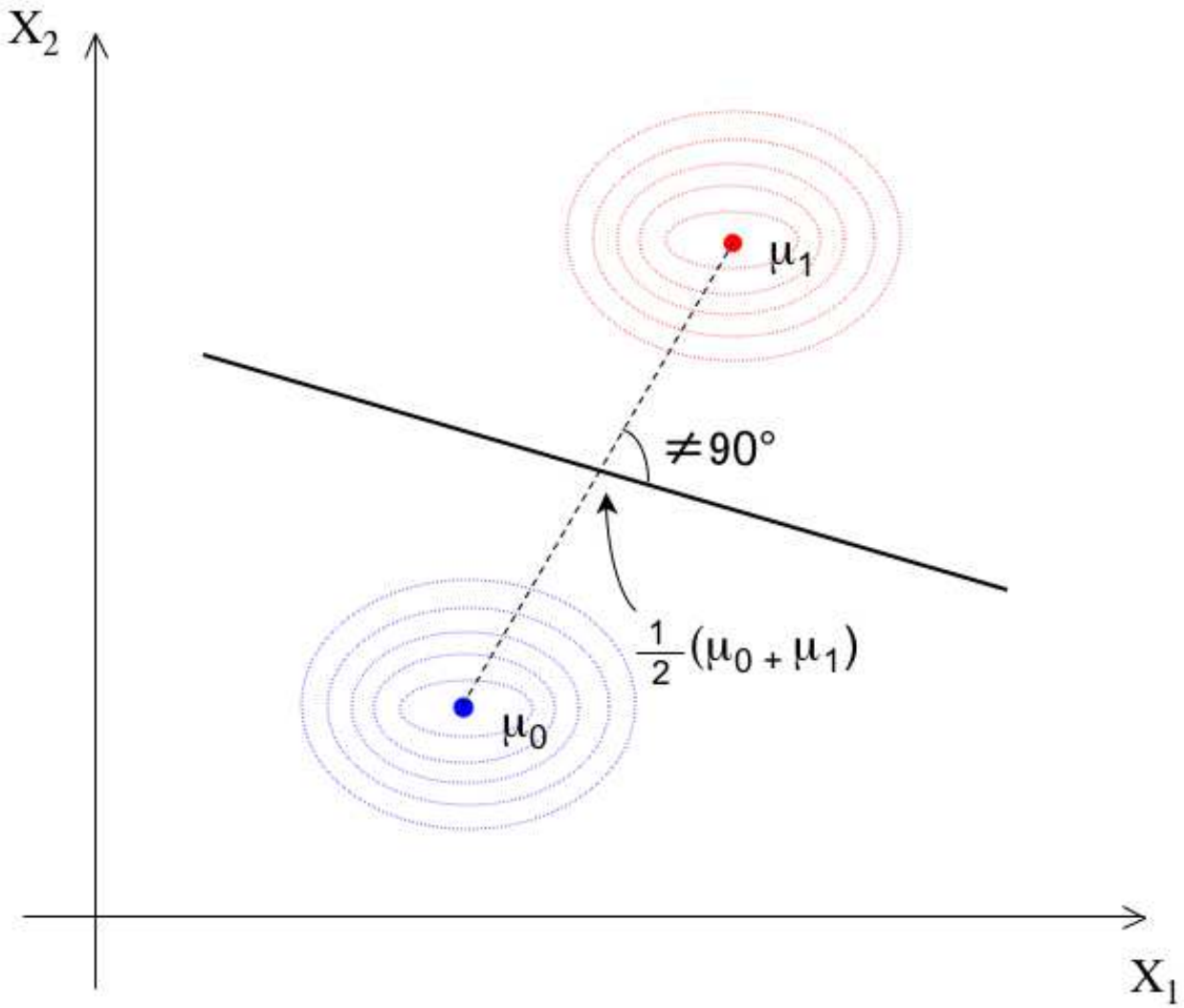
$$a = \Sigma^{-1}(\mu_1 - \mu_0)$$

$$x_0 = \frac{1}{2}(\mu_1 + \mu_0)$$

that is, the hyperplane must pass through the midpoint between the means, but it is *not* in general perpendicular to the axis joining the means (as was true for the nearest-mean classifier in case 1), because $\Sigma^{-1}(\mu_1 - \mu_0)$ is not in general parallel to $(\mu_1 - \mu_0)$, unless the latter is an eigenvector of Σ^{-1} (and thus of Σ).

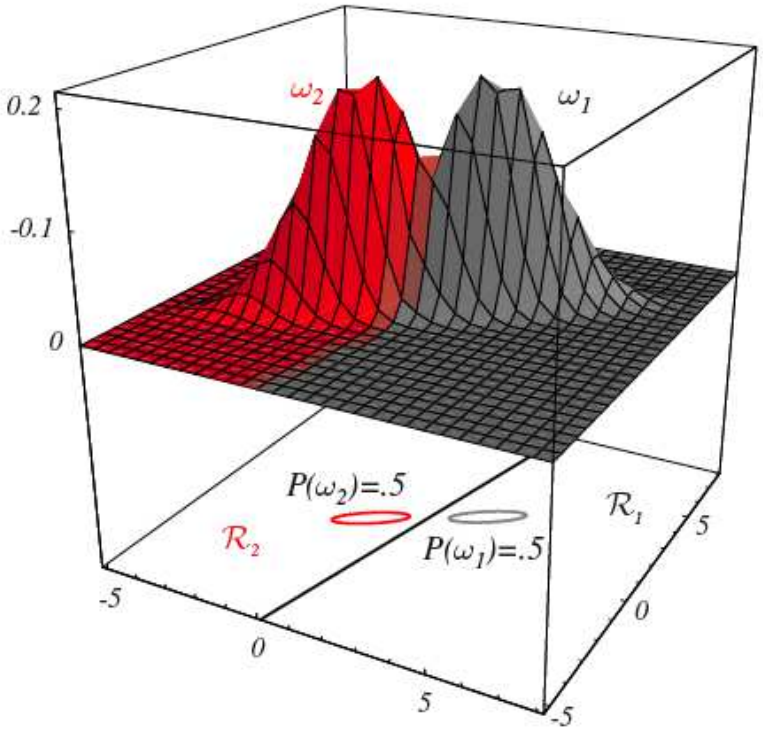
Gaussian Model - Linear Discriminant

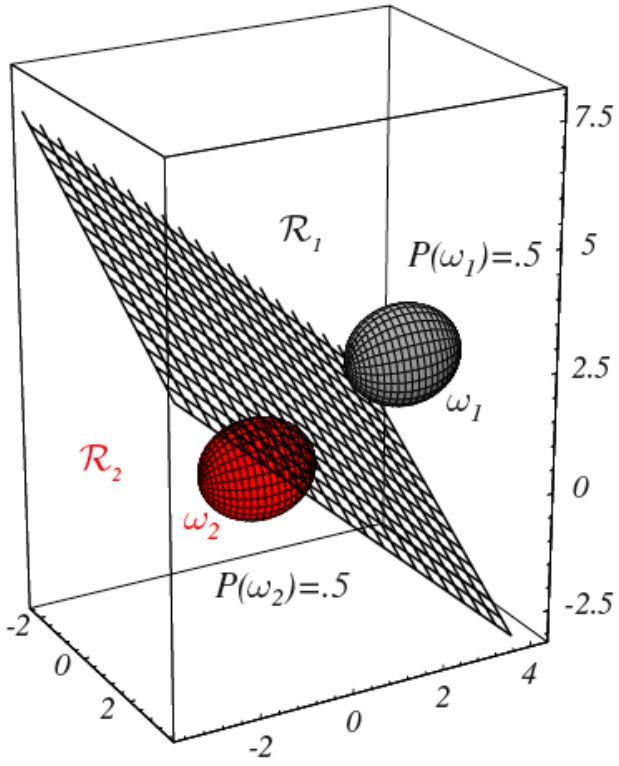
Example of Optimal Linear Classifier:



Gaussian Model - Linear Discriminant

More Examples of Optimal Linear Classifiers:





Gaussian Model - Linear Discriminant

- Using the properties of the Gaussian distribution,

$$g(X) \mid Y=0 \;\sim\; a^T X + b \mid Y=0 \;\sim\; \mathcal{N}(a^T\mu_0 + b, a^T\Sigma a)$$

$$g(X) \mid Y=1 \;\sim\; a^T X + b \mid Y=1 \;\sim\; \mathcal{N}(a^T\mu_1 + b, a^T\Sigma a)$$

- It follows that

$$\epsilon^0[\psi^*] \;=\; P(g(X) > 0 \mid Y = 0) \;=\; \Phi\left(\frac{a^T\mu_0 + b}{\sqrt{a^T\Sigma a}}\right)$$

$$\epsilon^1[\psi^*] \;=\; P(g(X) \leq 0 \mid Y = 1) \;=\; \Phi\left(-\frac{a^T\mu_1 + b}{\sqrt{a^T\Sigma a}}\right)$$

where $\Phi(x)$ is the CDF of a $\mathcal{N}(0,1)$ distribution.

Gaussian Model - Linear Discriminant

- Substituting the values of a and b leads to

$$\epsilon^0[\psi^*] = \Phi\left(\frac{-\frac{1}{2}\,\delta^2 + k}{\delta}\right) \quad \text{and} \quad \epsilon^1[\psi^*] = \Phi\left(\frac{-\frac{1}{2}\,\delta^2 - k}{\delta}\right)$$

where $k = \ln(P(Y=1)/P(Y=0))$ and

$$\delta = \sqrt{(\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)}$$

is the *Mahalanobis distance* between the populations.

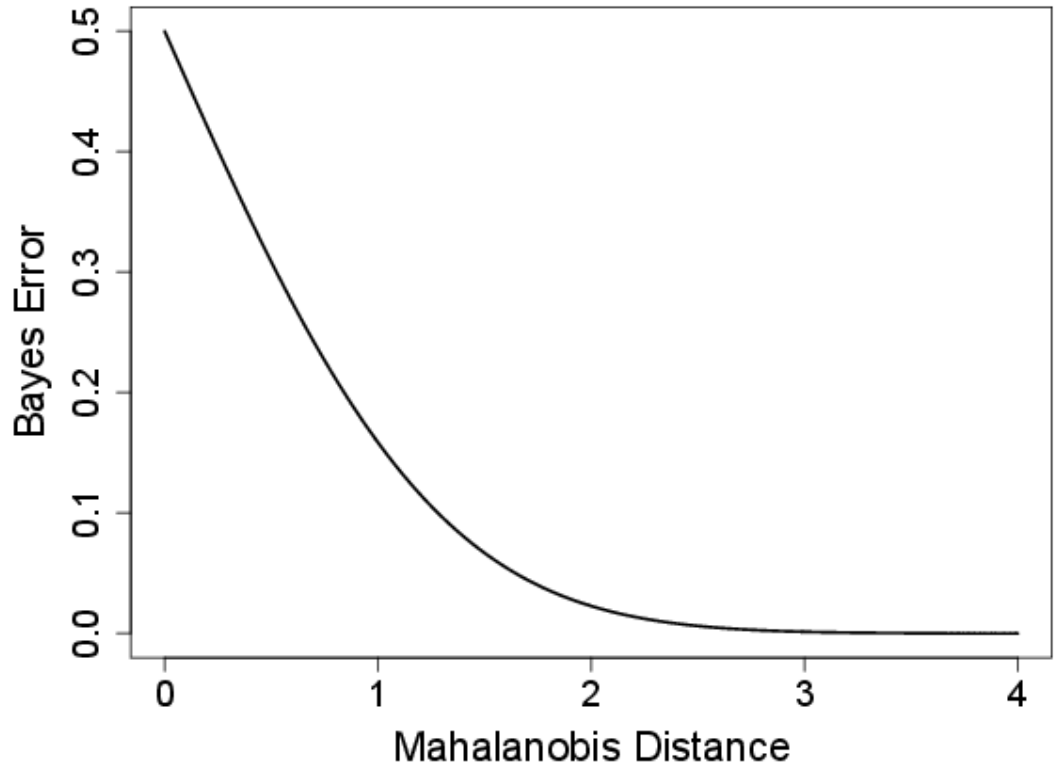
- The Bayes error is given by

$$\begin{aligned}\epsilon^* &= (1-c)\epsilon^0[\psi^*] + c\,\epsilon^1[\psi^*] \\ &= (1-c)\Phi\left(\frac{-\frac{1}{2}\,\delta^2 + k}{\delta}\right) + c\,\Phi\left(\frac{-\frac{1}{2}\,\delta^2 - k}{\delta}\right)\end{aligned}$$

Gaussian Model - Linear Discriminant

- In the case $P(Y=0) = P(Y=1) = 1/2$, the previous expressions reduce to:

$$\epsilon^0[\psi^*] = \epsilon^1[\psi^*] = \epsilon^* = \Phi\left(-\frac{\delta}{2}\right)$$



Gaussian Model - Quadratic Discriminant

- Case 3: Distinct arbitrary covariance matrices.

- In this case, the discriminant functions are fully quadratic

$$g_i(x) = x^T A_i x + b_i^T x + c_i$$

where

$$\begin{aligned} A_i &= -\frac{1}{2}\Sigma_i^{-1} \\ b_i &= \Sigma_i^{-1}\mu_i \\ c_i &= -\frac{1}{2}\mu_i^T\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(Y=i) \end{aligned}$$

Gaussian Model - Quadratic Discriminant

- In the case $c = 2$, the single discriminant is given by:

$$g(x) = g_1(x) - g_0(x) = x^T A x + b^T x + c = 0$$

where

$$A = -\frac{1}{2}\left(\Sigma_1^{-1} - \Sigma_0^{-1}\right)$$

$$b = \Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0$$

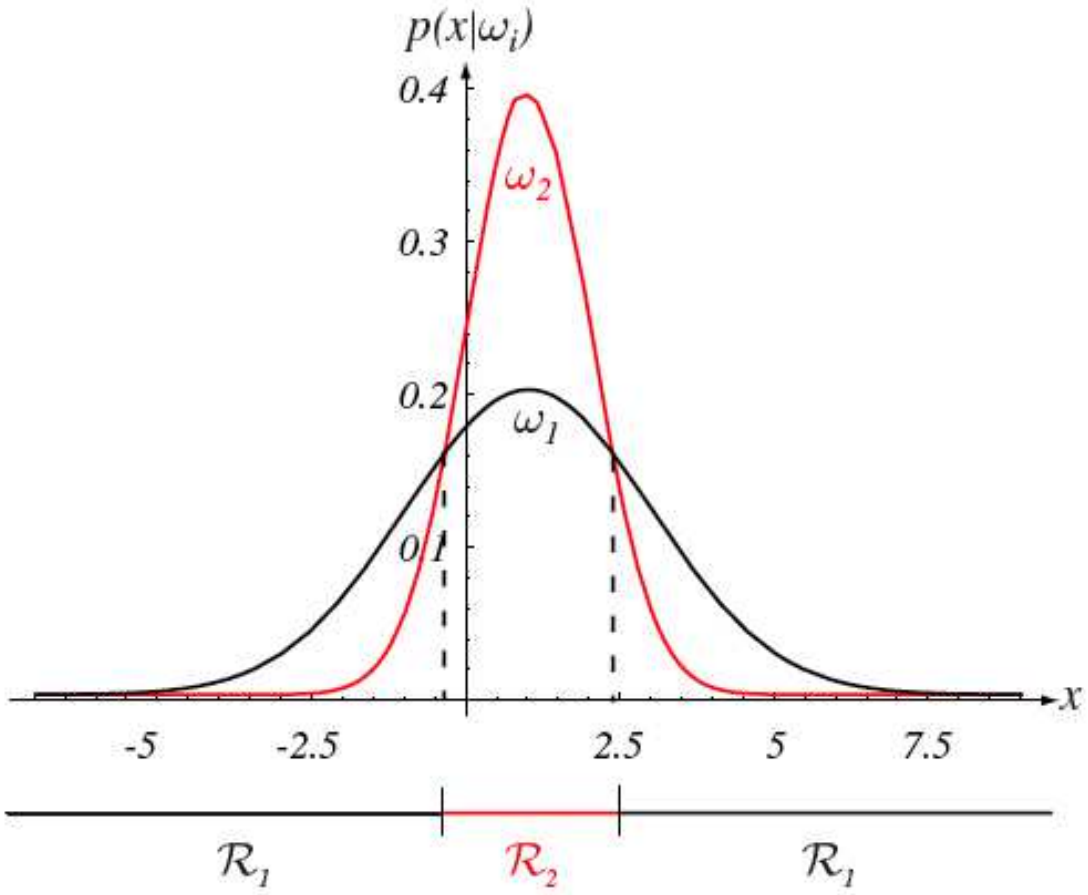
$$c = -\frac{1}{2}\left(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0\right) - \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_0|} + \ln\frac{P(Y=1)}{P(Y=0)}$$

Gaussian Model - Quadratic Discriminant

- The corresponding decision surfaces are *hyperquadrics*, which can be exactly one of the following:
  - hyperplanes
  - pairs of parallel hyperplanes
  - pairs of intersecting hyperplanes
  - hyperspheres
  - hyperellipsoids
  - hyperparaboloids
  - hyperhyperboloids

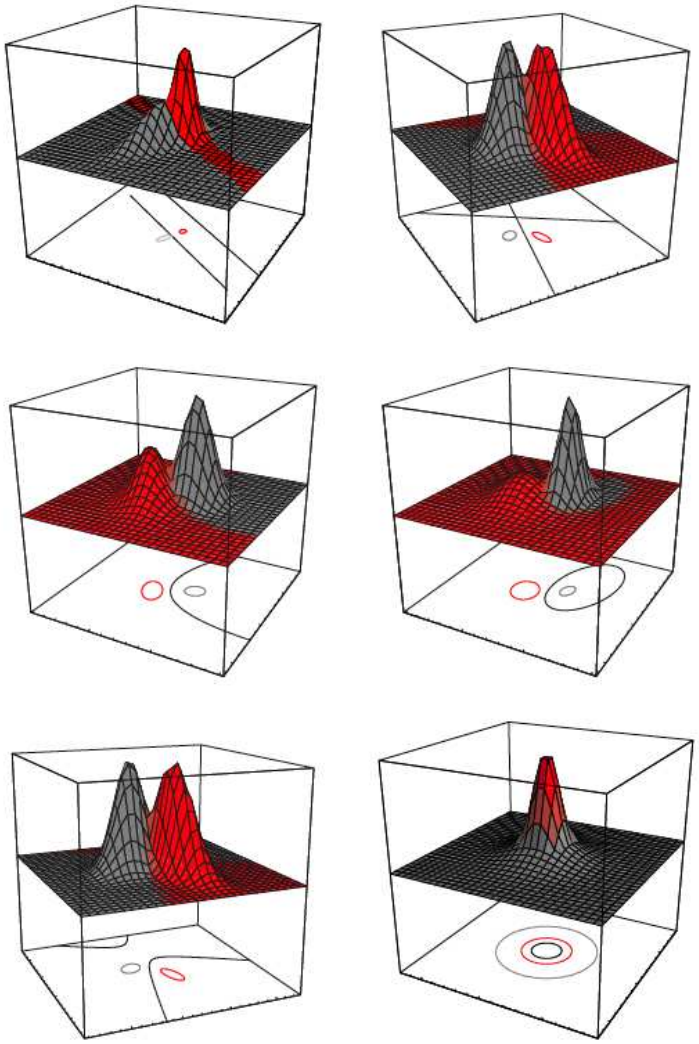
Gaussian Model - Quadratic Discriminant

Example of Quadratic Discriminant in 1-D case:



Gaussian Model - Quadratic Discriminant

More Examples of Quadratic Discriminants:



Alternative Distance Measures

- The Bayes error

$$\epsilon^* = E[\min\{\eta(x), 1 - \eta(x)\}]$$

is a *functional* of the feature-label distribution, $\epsilon^* = \Phi(F_{XY})$ which measures the discriminatory content in F_{XY}.

- It measures the quality of the distribution F_{XY} (that is, assuming Y fixed, the quality of the feature vector X) for pattern recognition.

- Equivalently, it provides a measure of *distance* between the classes.

Alternative Distance Measures - II

- Over the years, other measures of discrimination, i.e., other functionals of F_{XY} have been proposed.
- These *alternative distance measures* are related to the Bayes error, though usually not through a one-to-one relationship.
- They may be useful in feature selection, in proofs, and in the general understanding of why the Bayes error (and thus the difficulty of classification) is large or small. They are also of historical significance.

Kolmogorov's Variational Distance

- Kolmogorov's Variational Distance is given by

$$\delta_{\mathrm{KO}} = \frac{1}{2}E[|\eta(x) - (1-\eta(x))|] = \frac{1}{2}E[|2\eta(x) - 1|]$$

Clearly,

$$\epsilon^* = \frac{1}{2} - \frac{1}{2}E[|2\eta(x) - 1|] = \frac{1}{2} - \delta_{\mathrm{KO}}$$

- This is therefore a "true" distance: the larger it is, the further apart are the classes. It is of limited interest though because it is linearly related to the Bayes error.

Nearest-Neighbor Distance

- The Nearest-Neighbor Distance is given by

$$\epsilon_{\mathrm{NN}} \;=\; E[2\eta(x)(1-\eta(x))]$$

- This is the asymptotic error of the nearest-neighbor classification rule (Cover-Hart Theorem, proved later).
- The following *distribution-free* inequalities hold:

$$\epsilon^* \;\leq\; \epsilon_{\mathrm{NN}} \;\leq\; 2\epsilon^*(1-\epsilon^*) \;\leq\; 2\epsilon^*$$

The first inequality follows directly from the fact that

$$\min\{\eta, 1-\eta\} \;\leq\; 2\eta(1-\eta)\,, \qquad \text{for all } 0 \leq \eta \leq 1$$

The second inequality requires just a bit more work (see DGL. p. 22), while the third inequality is straightforward.

Nearest-Neighbor Distance - II

- Since $0 \leq 2\eta(x)(1-\eta(x)) \leq \frac{1}{2}$, for all x, we have

$$0 \leq \epsilon_{\mathrm{NN}} \leq \frac{1}{2}$$

- Furthermore, it is clear that

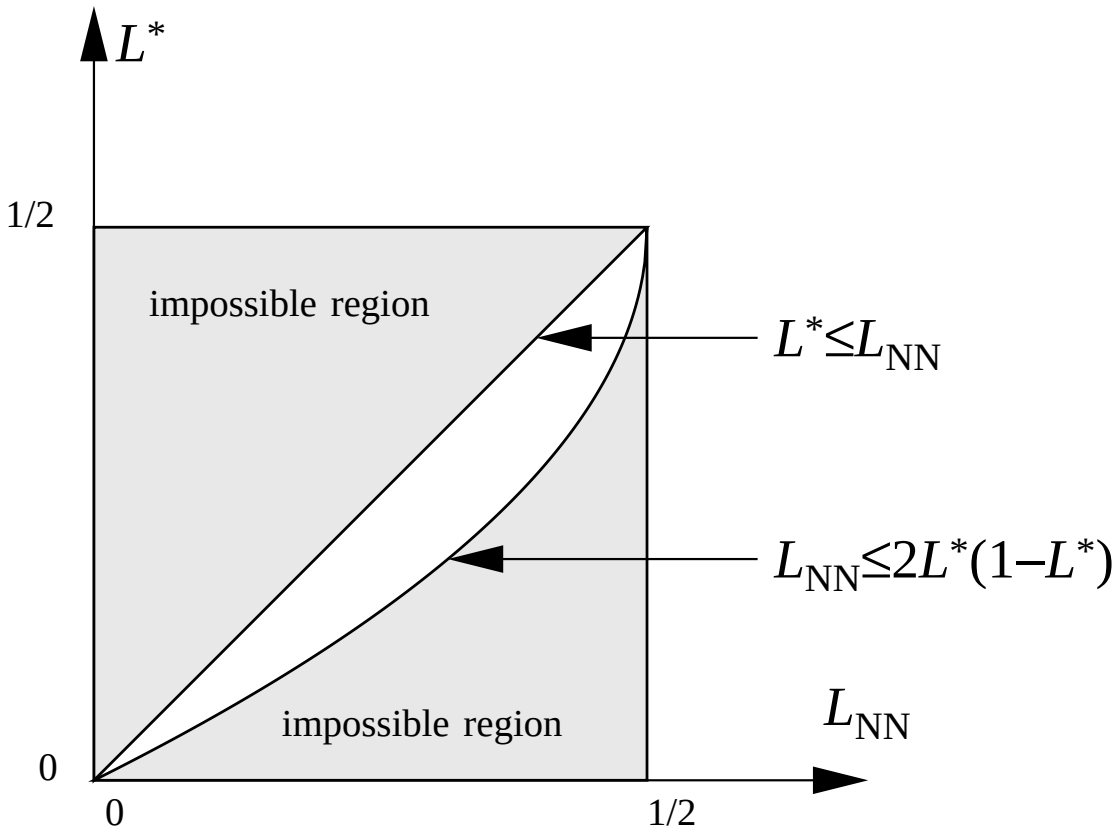
$$\epsilon_{\mathrm{NN}} = 0 \;\Leftrightarrow\; \eta(x) \in \{0,1\} \text{ w.prob.1} \;\Leftrightarrow\; \epsilon^* = 0$$

$$\epsilon_{\mathrm{NN}} = \frac{1}{2} \;\Leftrightarrow\; \eta(x) = \frac{1}{2} \text{ w.prob.1} \;\Leftrightarrow\; \epsilon^* = \frac{1}{2}$$

so ϵ_{NN} carries plenty of information about ϵ^*.

Nearest-Neighbor Distance - III

Graphical interpretation of bounds:



Mahalanobis Distance

- Consider the first and second moments of the class-conditional distributions; for $i = 0, 1$,

$$\mu_i = E[X \mid Y = i]$$
$$\Sigma_i = E[(X - \mu_i)(X - \mu_i)^T \mid Y = i]$$

and let $\Sigma = P(Y = 0)\Sigma_0 + P(Y = 1)\Sigma_1$ be the *pooled* covariance matrix.

- The *Mahalanobis distance*, defined previously as

$$\delta = \sqrt{(\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)}$$

is a an alternative distance measure that is related to the Bayes error.

Mahalanobis Distance - II

- If $\Sigma_0 = \Sigma_1 = \sigma^2 I$, then

$$\delta = \frac{||\mu_1 - \mu_0||}{\sigma}$$

showing that δ is the distance between the means, "normalized" by the common variance.

- If δ is large, the classes are well separated, and the Bayes error should be small. This is shown next.

- (Devijver and Kitler, 1982) For all F_{XY} such that $E[||X^2||] < \infty$, we have

$$\epsilon^* \leq \epsilon_{\text{NN}} \leq \frac{2c(1-c)}{1 + c(1-c)\delta^2}$$

where $c = P(Y = 1)$. Thus, if δ is large, ϵ^* must be small.

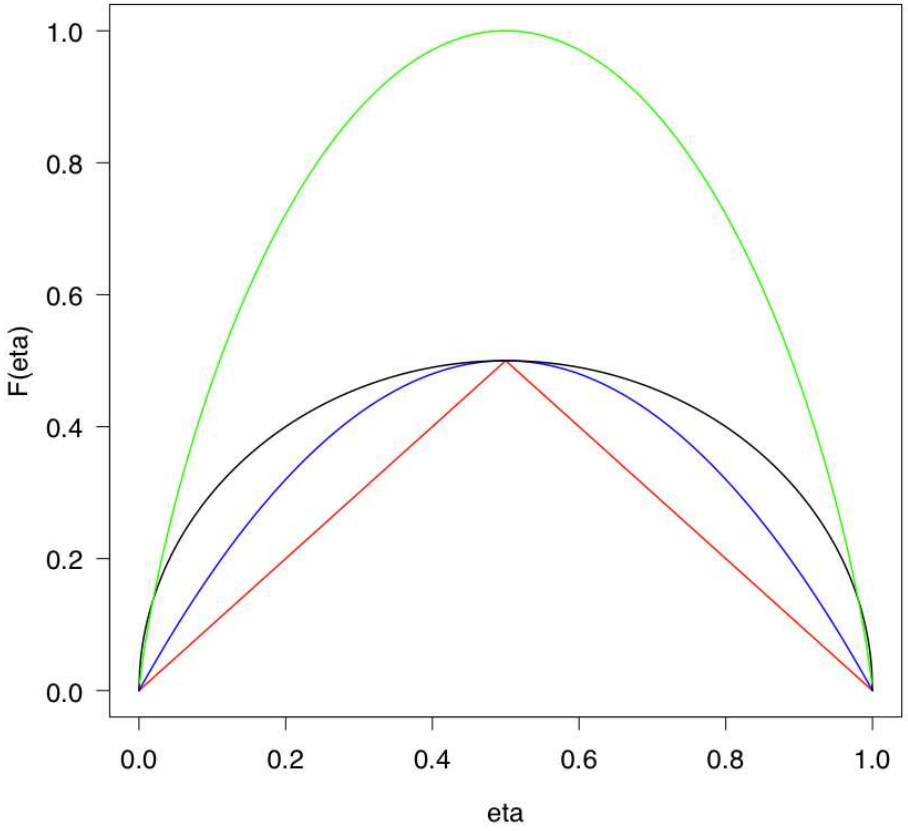
F-errors

- Given any concave function $F : [0, 1] \to [0, \infty)$, we define the F-error for the distribution F_{XY} as

$$d_F(X, Y) = E[F(\eta(X))]$$

- Examples of F-errors:
  - Bayes Error: $F(u) = \min(u, 1 - u)$
  - Nearest-Neighbor Error: $F(u) = 2u(1 - u)$
  - Matsushita Error: $F(u) = \sqrt{u(1 - u)}$
  - Chernoff Error: $F(u) = u^\alpha (1 - u)^{1-\alpha}$, for $0 < \alpha < 1$
  - Conditional Entropy: $F(u) = \mathcal{H}(\eta(x), 1 - \eta(x)) = -u \log_2 u - (1 - u) \log_2 (1 - u)$

F-errors - II



$\min\{\eta, 1-\eta\}$

$2\eta(1-\eta)$

$\sqrt{\eta(1-\eta)}$

$\mathcal{H}(\eta, 1-\eta)$

These plots show that $\epsilon^* \leq \epsilon_{\mathrm{NN}} \leq \rho \stackrel{\Delta}{=} E\left[\sqrt{\eta(x)(1-\eta(x))}\right]$

as well as $\epsilon^* \leq \epsilon_{\mathrm{NN}} \leq \mathcal{E} \stackrel{\Delta}{=} E[\mathcal{H}(\eta(x), 1-\eta(x))]$, but not $\rho \leq \mathcal{E}$ or $\mathcal{E} \leq \rho$.

F-errors - III

Some properties of F-errors:

- If $F(u) \geq \min(u, 1-u)$ for all $u \in [0,1]$ then $d_F \geq \epsilon^*$.
- $d_F \geq 0$, and If $F(u) = 0$ only at $u = 0, 1$, then $d_F = 0 \Leftrightarrow \epsilon^* = 0$.
- If $F(u)$ reaches a maximum at $u = \frac{1}{2}$ then d_F is maximum $\Leftrightarrow \epsilon^* = \frac{1}{2}$.

F-errors - IV

- F-error Information-Loss Property (DGL Theorem 3.3). Let $X' = t(X)$, where $t : R^d \to R^k$ is any measurable transformation. Then

$$d_F(X', Y) \geq d_F(X, Y)$$

with equality if t is invertible. In particular, an F-error error never decreases upon feature set transformation.

- Since the Bayes error is an F-error, it has all the properties mentioned above.