Bayesian Decision Making

Introduction

- Classification different from regression in the sense that the output will be a discrete label denoting the entity of the class.

- We will study the decision making process, by relying on probabilistic inferences.

- The Bayes Theorem is very powerful, and knowledge of it helps to design the classifier, with appropriate decision surfaces.

Bayes Decision

- It is the decision making process when all underlying probability distributions are known
- Generative model
- It is optimal given the distributions are known.
- In this discussion, we assume supervised learning paradigm.

Problem

- 2 type of fish -sea bass and salmon in a conveyer belt
- Problem: Need to recognize the type of fish.
- 2 class/ category problem

Prior probability

Let us denote the 2 classes by

- ω_1 : Salmon
- ω_2 : Sea bass

- Let us say that salmon occurs more likely than sea bass. (This information is known as the prior and is generally estimated by experimentation.)

Mathematically , $P(\omega_1) > P(\omega_2)$

Estimation of prior probabilities by frequentist approach

- Assume that out of 8000 fish used for training, we observed that 6000 were salmon, 2000 sea bass

- Accordingly we have

$P(\omega_1) = 0.75$ prior probability of salmon

$P(\omega_2) = 0.25$ prior probability of sea bass

Simple Decision making based on prior knowledge

Assign unknown fish as salmon ω_1 , if

$$P(\omega_1) > P(\omega_2)$$

else

assign it to sea bass ω_2.

Issues with using only prior knowledge

- Decision rule is flawed. Salmon will always be favored and assigned to a test fish.
- Such an approach will not work in practice.
- Sole dependence on prior probability for making decisions is not a good idea

Need for features

- **Solution**: Look for additional information to describe the sea bass and salmon.
- Key idea is to look for discriminative features.
- Features like length, width, color, life span, texture may describe salmon and sea bass.

Feature vector

- Assume that we have *d* features.

- Let set of features describing a fish be represented by a *d* dimensional feature vector **x**

Class Conditional Density

- For each class/ category of fish, we can associate the *d* features to come from a probability distribution function.
- This pdf is referred to as ‘class conditional density’.
- The nature of the features are continuous.
- Note that, for a set of *d* features describing the fish, we work on a *d* dimensional probability distribution.

- For the time being , let us work on how to improve our decision process by incorporating a single feature x.

- Later, we extend the framework for d dimensional features, and also for classes greater than 2.

Class conditional Density

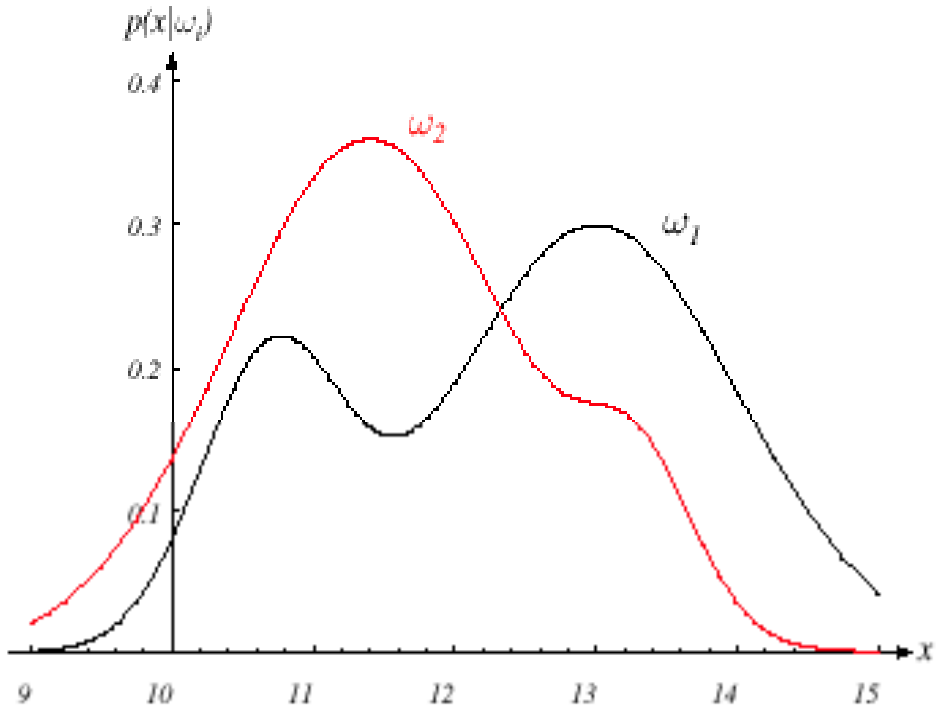


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayes Theorem

- Bayes Theorem :

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j) P(\omega_j)}{P(x)}$$

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

In the case of two categories

$$P(x) = \sum_{j=1}^{j=2} P(x \mid \omega_j) P(\omega_j)$$

Posterior probability plot for the two classes

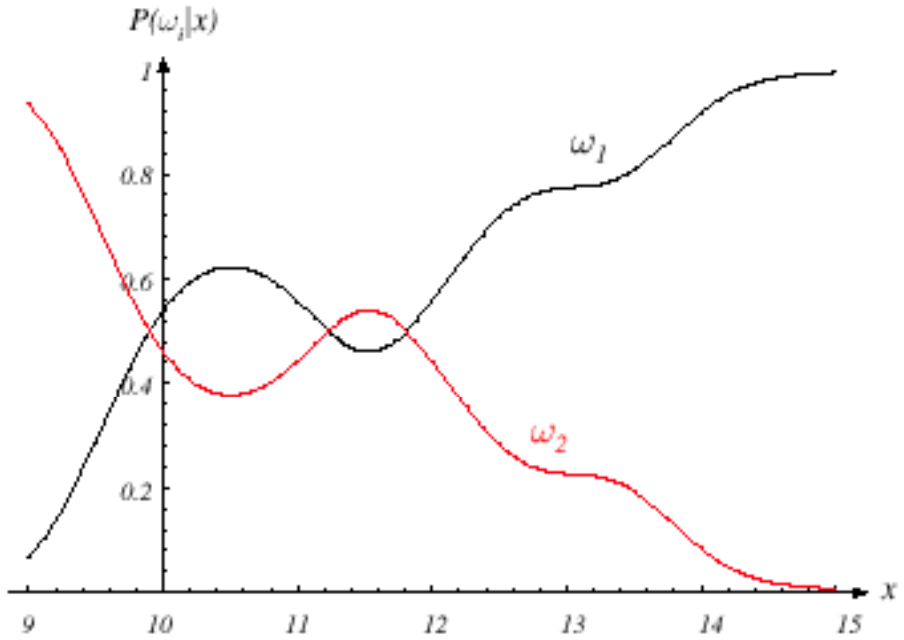


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.

Decision based on posterior probabilities

- Decision given the posterior probabilities

x is an observation for which:

if $P(\omega_1 | x) > P(\omega_2 | x)$ True state of nature $= \omega_1$
if $P(\omega_1 | x) < P(\omega_2 | x)$ True state of nature $= \omega_2$

Therefore: whenever we observe a particular x, the probability of error is :

$$P(\text{error} | x) = P(\omega_1 | x) \text{ if we decide } \omega_2$$

$$P(\text{error} | x) = P(\omega_2 | x) \text{ if we decide } \omega_1$$

Decision based on posterior probabilities

- Minimizing the probability of error

Decide ω_1 *if* $P(\omega_1 | x) > P(\omega_2 | x)$; otherwise decide ω_2

Therefore:

$$P(\text{error} | x) = \min [P(\omega_1 | x), P(\omega_2 | x)]$$

(Bayes decision)

Probability of error

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

- We want $P(error \mid x)$ to be as small as possible for every value of x

The Bayes classifier scheme strives to achieve that

Bayesian classification framework for high dimensional features and more classes

Let $\{\omega_1, \omega_2, \ldots, \omega_C\}$ be the set of "*C*" states of nature (or "categories" / "classes")

Assume , that for an unknown pattern, a *d* dimensional feature vector **x** is constructed:

From Bayes rule

$$P(\omega_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_j) P(\omega_j)}{P(\mathbf{x})}$$

We compute the posterior probability of the pattern with respect to each of the "*c*" classes.

In the decision making step, we assign the pattern to the class for which the posterior probability is greatest.

Bayesian classification framework for high dimensional features and more classes

$$\omega_{test} = \arg\max_j P(\omega_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_j) P(\omega_j)}{P(\mathbf{x})} \quad j = 1,2....C$$

$$P(\mathbf{x}) = \sum_{j=1}^{C} p(\mathbf{x} \mid \omega_j) P(\omega_j)$$

Evidence acts as a normalization factor ….term same for all Classes.

ω_{test} is the class for which the posterior probability is highest.

The pattern is assigned to this class.

Risk minimization framework

Let $\{\omega_1, \omega_2, \ldots, \omega_C\}$ be the set of "*C*" states of nature (or "categories")

Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of possible "*a*" actions

Let $\lambda(\alpha_i \mid \omega_j)$ be the loss incurred for taking action α_i when the state of nature is ω_j

Risk minimization framework

The expected loss: $$R(\alpha_i) = \sum_{j=1}^{C} \lambda(\alpha_i \mid \omega_j) P(\omega_j)$$

Given an observation with vector **x**, the conditional risk is:

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

At every **x**, a decision is made: $\alpha(\boldsymbol{x})$, by minimizing the expected loss.

Our final goal is to minimize the total risk over all **x**.

$$\int R(\alpha(\mathbf{x}) \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

Reading

- Sections 2.1, 2.2 :

  Duda, Hart , Stork : Pattern Classification

Bayesian Decision Making

Risk minimization framework

Let $\{\omega_1, \omega_2, \ldots, \omega_C\}$ be the set of "*C*" states of nature (or "categories")

Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of possible "*a*" actions

Let $\lambda(\alpha_i \mid \omega_j)$ be the loss incurred for taking action α_i when the state of nature is ω_j

Risk minimization framework

The expected loss: $$R(\alpha_i) = \sum_{j=1}^{C} \lambda(\alpha_i \mid \omega_j) P(\omega_j)$$

Given an observation with vector **x**, the conditional risk is:

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

At every **x**, a decision is made: $\alpha(\boldsymbol{x})$, by minimizing the expected loss.

Our final goal is to minimize the total risk over all **x**.

$$\int R(\alpha(\mathbf{x}) \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

Risk minimization framework

- Two-category classification

  α_1 : *deciding* ω_1

  α_2 : *deciding* ω_2

  $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$ *is* loss incurred for deciding ω_i when the true state of nature is ω_j

Conditional risk:

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

Risk minimization framework

Our rule is the following:

if $R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$

action α_1: "decide ω_1" is taken

This results in the equivalent rule :

Decide ω_1 if:

$$(\lambda_{21} - \lambda_{11})\, p(\mathbf{x} \mid \omega_1)\, P(\omega_1) > (\lambda_{12} - \lambda_{22})\, p(\mathbf{x} \mid \omega_2)\, P(\omega_2)$$

and decide ω_2 otherwise

Likelihood ratio

The preceding rule is equivalent to the following rule:

$$if \ \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

Then take action α_1 (decide ω_1)

Otherwise take action α_2 (decide ω_2)

Likelihood ratio

- Regions of decision and zero-one loss function, therefore:

$$Let \quad \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda$$

$$\text{then decide } \omega_1 \text{ if : } \frac{p(\mathbf{x}\,|\,\omega_1)}{p(\mathbf{x}\,|\,\omega_2)} > \theta_\lambda$$

Likelihood ratio

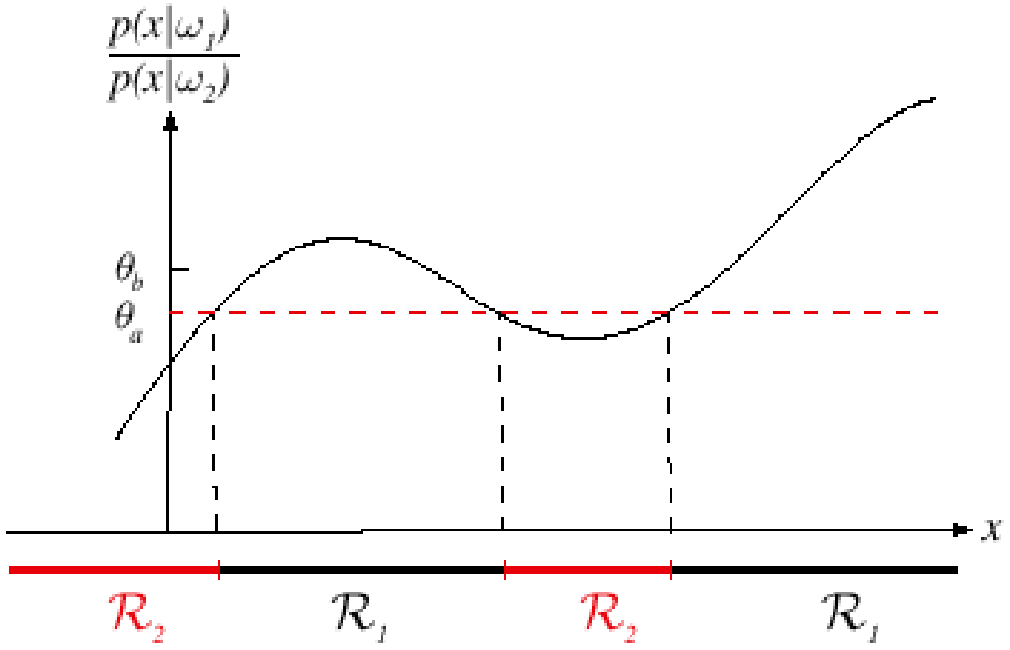


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Likelihood ratio

$$\lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$\text{then } \theta_\lambda = \frac{P(\omega_2)}{P(\omega_1)} = \theta_a$$

$$\text{if } \lambda = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix} \text{ then } \theta_\lambda = \frac{2P(\omega_2)}{P(\omega_1)} = \theta_b$$

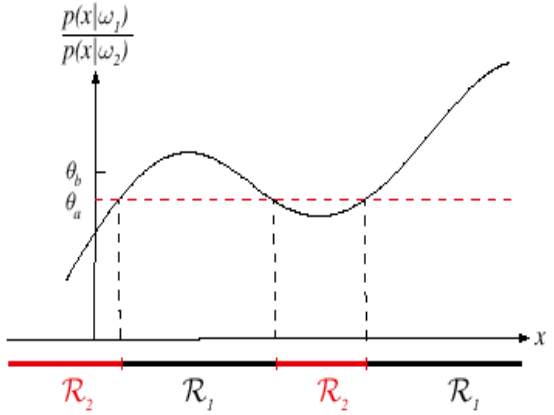


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- If loss function penalizes miscategorizing ω_2 as ω_1 more than converse we get larger threshold θ_b and hence $\boldsymbol{R}_1$ becomes smaller

Likelihood ratio

Optimal decision property

"If the likelihood ratio exceeds a threshold value independent of the input pattern **x**, we can take optimal actions"

Zero-one loss function

- Actions are decisions on classes

  If action α_i is taken and the true state of nature is ω_j
  then: the decision is correct if $i = j$ and in error if $i \neq j$

- Seek a decision rule that minimizes the *probability of error* which is the *error rate*

Zero-one loss function

- Introduction of the zero-one loss function:

$$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

Therefore, the conditional risk is: All errors are equally costly.

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

$$= \sum_{j \neq i} P(\omega_j \mid \mathbf{x}) = 1 - P(\omega_i \mid \mathbf{x})$$

"The risk corresponding to this loss function is the average probability error"

Zero-one loss function

- Minimize the risk requires maximizing $P(\omega_i \mid \boldsymbol{x})$

  (since $R(\alpha_i \mid \boldsymbol{x}) = 1 - P(\omega_i \mid \boldsymbol{x})$)

- For Minimum error rate

  – Decide ω_i *if* $P(\omega_i \mid \boldsymbol{x}) > P(\omega_j \mid \boldsymbol{x}) \ \forall \ j \neq i$

Bayesian Decision Making

Discriminant functions, Normal Distribution

Classifiers, Discriminant Functions and Decision Surfaces

- The multi-category case
  - Set of discriminant functions $g_i(\mathbf{x}), i = 1,\ldots, c$
  - The classifier assigns a feature vector $\mathbf{x}$ to class ω_i

    if:

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

Generic view of a pattern classifier

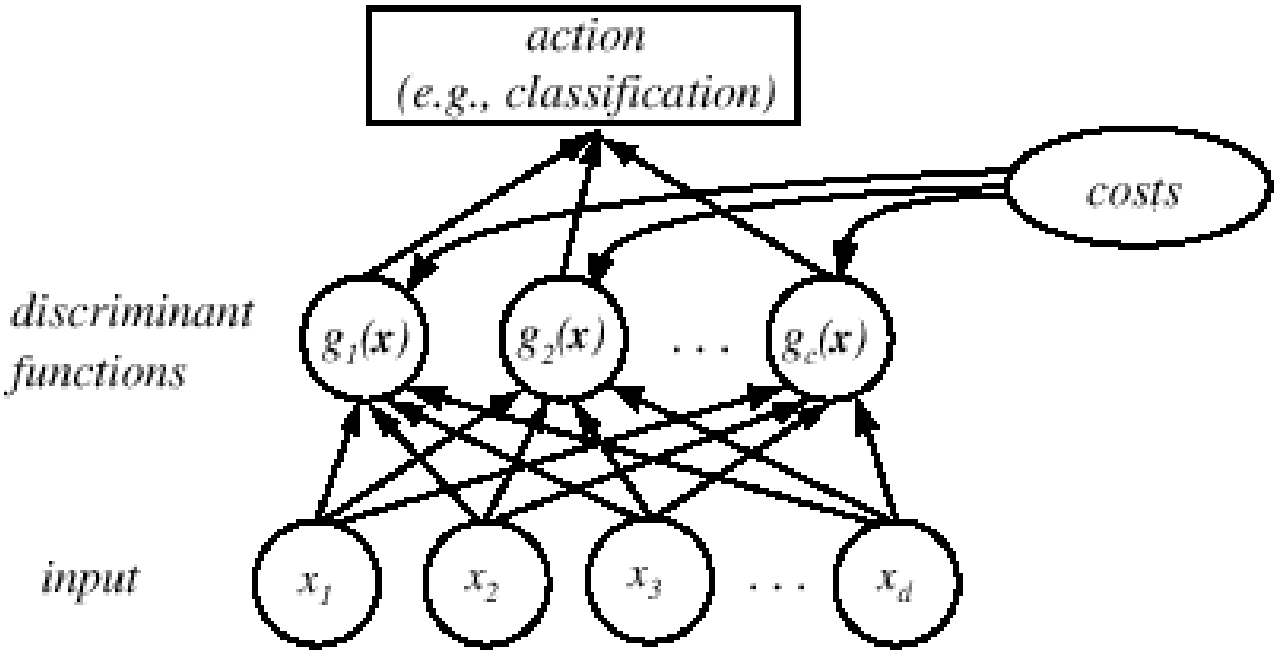


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Discriminant Functions

- *Let* $g_i(\boldsymbol{x}) = -R(\alpha_i \mid \boldsymbol{x})$
  (max. discriminant corresponds to min. risk!)
- For the minimum error rate, we take

$$g_i(\boldsymbol{x}) = P(\omega_i \mid \boldsymbol{x})$$

(max. discrimination corresponds to max. posterior!)

$$g_i(\boldsymbol{x}) \equiv p(\boldsymbol{x} \mid \omega_i)\, P(\omega_i)$$

$$g_i(\boldsymbol{x}) = \ln p(\boldsymbol{x} \mid \omega_i) + \ln P(\omega_i)$$

(ln: natural logarithm!)

Discriminant functions

- Discriminant functions do not change the decision, when scaled by some positive constant '*k*'.

- The decision is not affected when a constant is added to all discriminant functions.

Discriminant Functions

- Feature space divided into c decision regions

$$\text{if } g_i(\mathbf{x}) > g_j(\mathbf{x}) \ \forall \ j \neq i \text{ then } x \text{ is in } \mathcal{R}_i$$

($\mathcal{R}_i$ means assign x to ω_i)

- The two-category case
  - A classifier is a "dichotomizer" that has two discriminant functions g_1 and g_2

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

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

Dichotomizer

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

(ln: natural logarithm!)

– The computation of $g(\mathbf{x})$ for dichotomizer

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

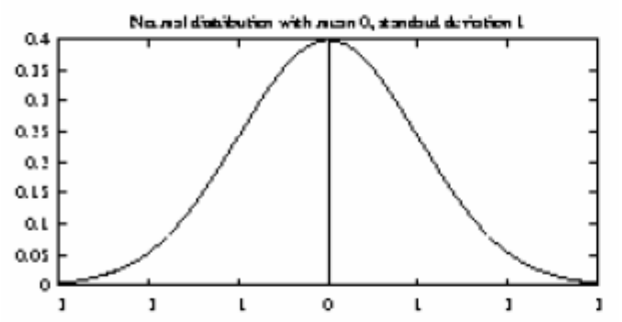
or

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

Normal /Gaussian Distribution

The Normal Distribution

A bell-shaped distribution defined by the probability density function



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

If the random variable X follows a normal distribution, then

- The probability that X will fall into the interval (a,b) is given by $\int_a^b p(x)dx$
- Expected, or mean, value of X is $E[X] = \int_{-\infty}^{\infty} xp(x)dx = \mu$
- Variance of X is $Var(x) = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \sigma^2$

The Normal Density in Pattern Recognition

- Univariate density
  - Analytically tractable, continuous
  - A lot of processes are asymptotically Gaussian
  - Central Limit Theorem: aggregate effect of a sum of a large number of small, independent random disturbances will lead to a Gaussian distribution
  - Handwritten characters, speech sounds are ideal or prototype corrupted by random process

Multivariate Gaussian Distribution

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}.$$

Where $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}$ and $\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] = \begin{bmatrix} \mathcal{E}[x_1] \\ \mathcal{E}[x_2] \\ \vdots \\ \mathcal{E}[x_d] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix}$:

$\boldsymbol{\Sigma}$ = *d* x *d* Covariance matrix

Similarly, the *covariance matrix* $\mathbf{\Sigma}$ is defined as the (square) matrix whose ij^{th} element σ_{ij} is the covariance of x_i and x_j:

$$\sigma_{ij} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)] \qquad i, j = 1 \ldots d,$$

$$\begin{aligned}
\mathbf{\Sigma} &= \begin{bmatrix} \mathcal{E}[(x_1 - \mu_1)(x_1 - \mu_1)] & \mathcal{E}[(x_1 - \mu_1)(x_2 - \mu_2)] & \ldots & \mathcal{E}[(x_1 - \mu_1)(x_d - \mu_d)] \\ \mathcal{E}[(x_2 - \mu_2)(x_1 - \mu_1)] & \mathcal{E}[(x_2 - \mu_2)(x_2 - \mu_2)] & \ldots & \mathcal{E}[(x_2 - \mu_2)(x_d - \mu_d)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{E}[(x_d - \mu_d)(x_1 - \mu_1)] & \mathcal{E}[(x_d - \mu_d)(x_2 - \mu_2)] & \ldots & \mathcal{E}[(x_d - \mu_d)(x_d - \mu_d)] \end{bmatrix} \\
&= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \ldots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \ldots & \sigma_d^2 \end{bmatrix}.
\end{aligned} \tag{73}$$

Multivariate Gaussian Distribution

Mean vector has its components which are means of variables

Covariance :
Diagonal elements are variances of variables
Cross-diagonal elements are covariances of pairs of variables
Statistical independence means off-diagonal elements are zero

Covariance matrix property

- If **w** is any *d* dimensional vector, the variance of $\mathbf{w}^T\mathbf{x}$ can not be negative.

- This leads to the quadratic form $\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$ to be non-negative → positive semi-definite nature of $\boldsymbol{\Sigma}$

- Eigen values of $\boldsymbol{\Sigma}$ are non-negative.

Multivariate Gaussian Distribution

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

$$\boldsymbol{\mu} = \mathbf{E}[\mathbf{x}] = \int \mathbf{x}\, p(\mathbf{x})d\mathbf{x}, \qquad \boldsymbol{\Sigma} = \mathbf{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t]$$

Reminder: the covariance matrix is symmetric and positive semidefinite.

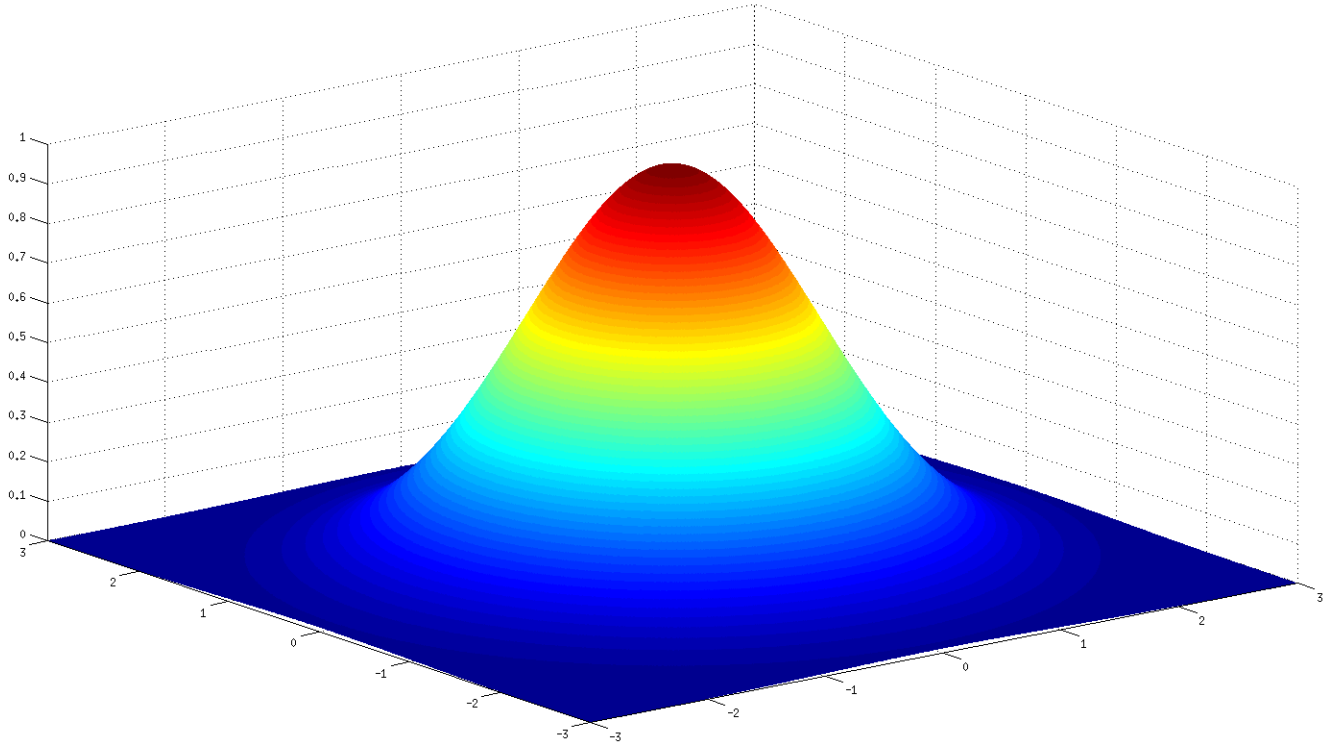
$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \mathbf{y} = \mathbf{A}^t\mathbf{x} \qquad p(\mathbf{y}) \sim N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\boldsymbol{\Sigma}\mathbf{A})$$

Entropy - the measure of uncertainty

Normal distribution has the maximum entropy over all distributions with a given mean and variance.

$$H(p(x)) = -\int p(x) \ln p(x) dx$$

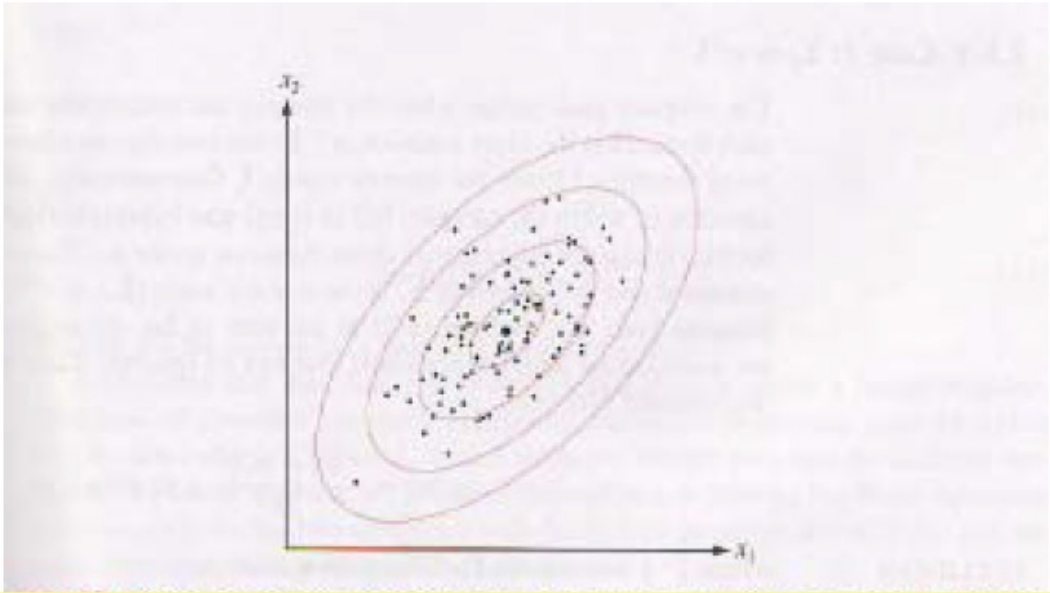
2D Gaussian



Multivariate Gaussian Distribution

Multivariate Normal Density

- Specified by d+d(d+1)/2 parameters: mean and independent elements of covariance matrix



Locii of points of constant density are hyperellipsoids

Samples drawn from a 2-D Gaussian lie in a cloud centered at the mean μ. Ellipses show lines of equal probability density of the Gaussian

Mahalanobis Distance

$$r^2 = (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}),$$

Contours of constant Density are hyperellipsoids of constant Mahanalobis Distance

Bayesian Decision Making

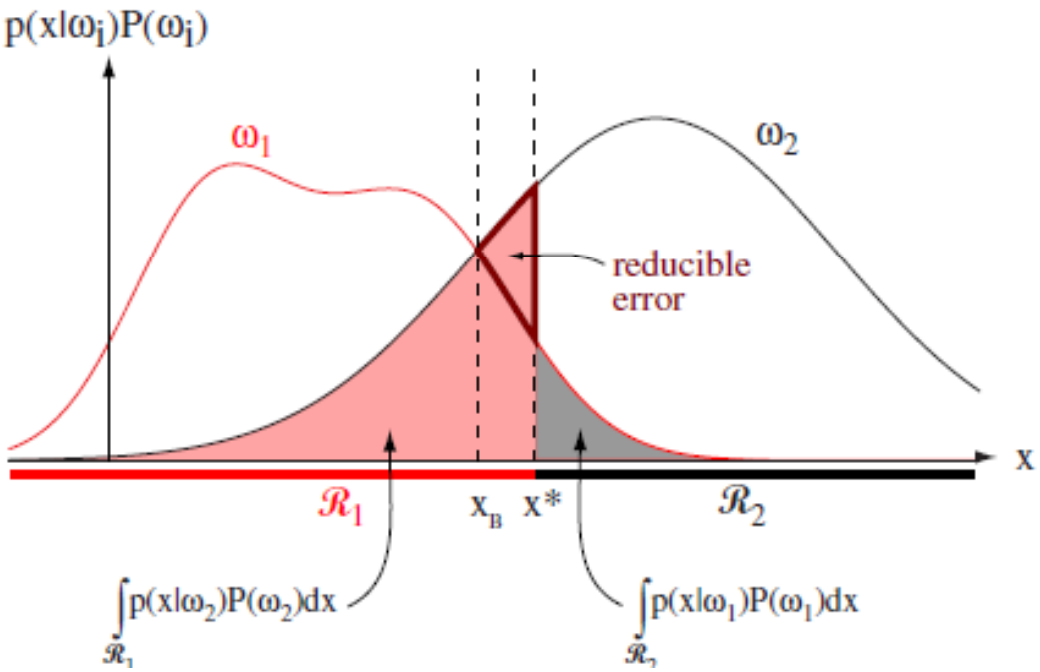
Error Computation

$$P(e) = P(\mathbf{x} \in R_1, \omega_2) + P(\mathbf{x} \in R_2, \omega_1)$$

$$= P(\mathbf{x} \in R_1 \mid \omega_2) P(\omega_2) + P(\mathbf{x} \in R_2 \mid \omega_1) P(\omega_1)$$

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

Error computation



Error computation

$$P(correct) = \sum_{i=1}^{C} P(x \in R_i, \omega_i)$$

$$= \sum_{i=1}^{C} P(x \in R_i \mid \omega_i) P(\omega_i)$$

$$= \sum_{i=1}^{C} \int_{R_i} p(\mathbf{x} \mid \omega_i) P(\omega_i)$$

Bayes Decision Theory – Discrete Features

- Components of **x** are binary or integer valued.

- **x** can take only one of m discrete values

$$\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m$$

$$\boxed{P(\omega_j|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_j)P(\omega_j)}{P(\mathbf{x})},}$$

where

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

Bayes Decision Theory – Discrete Features

The definition of the conditional risk $R(\alpha|\mathbf{x})$ is unchanged, and the fundamental Bayes decision rule remains the same: To minimize the overall risk, select the action α_i for which $R(\alpha_i|\mathbf{x})$ is minimum, or stated formally,

$$\alpha^* = \arg\max_i R(\alpha_i|\mathbf{x}).$$

Bayes Decision Theory – Discrete Features

- Case of independent binary features in 2 category problem

Let $\boldsymbol{x} = [x_1, x_2, \ldots, x_d]^T$ where each x_i is either 0 or 1, with probabilities:

$$p_i = P(x_i = 1 \mid \omega_1)$$
$$q_i = P(x_i = 1 \mid \omega_2)$$

Bayes Decision Theory – Discrete Features

Assuming, conditional independence of components of a feature vector, likelihoods can be computed as

$$P(\mathbf{x}|\omega_1) = \prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i}$$

and

$$P(\mathbf{x}|\omega_2) = \prod_{i=1}^{d} q_i^{x_i}(1-q_i)^{1-x_i}.$$

Then the likelihood ratio is given by

$$\frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} = \prod_{i=1}^{d} \left(\frac{p_i}{q_i}\right)^{x_i} \left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$$

Bayes Decision Theory – Discrete Features

$$g(\mathbf{x}) = \sum_{i=1}^{d} \left[x_i \ln \frac{p_i}{q_i} + (1 - x_i) \ln \frac{1 - p_i}{1 - q_i} \right] + \ln \frac{P(\omega_1)}{P(\omega_2)}.$$

this discriminant function is linear in the x_i

$$g(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i + w_0,$$

where

$$w_i = \ln \frac{p_i(1 - q_i)}{q_i(1 - p_i)} \qquad i = 1, ..., d$$

and

$$w_0 = \sum_{i=1}^{d} \ln \frac{1 - p_i}{1 - q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}.$$

$p_i = q_i$ ⟹ dependence on prior only

Bayes Decision Theory – Discrete Features

We have:

$$g(x) = \sum_{i=1}^{d} w_i x_i + w_0$$

where :

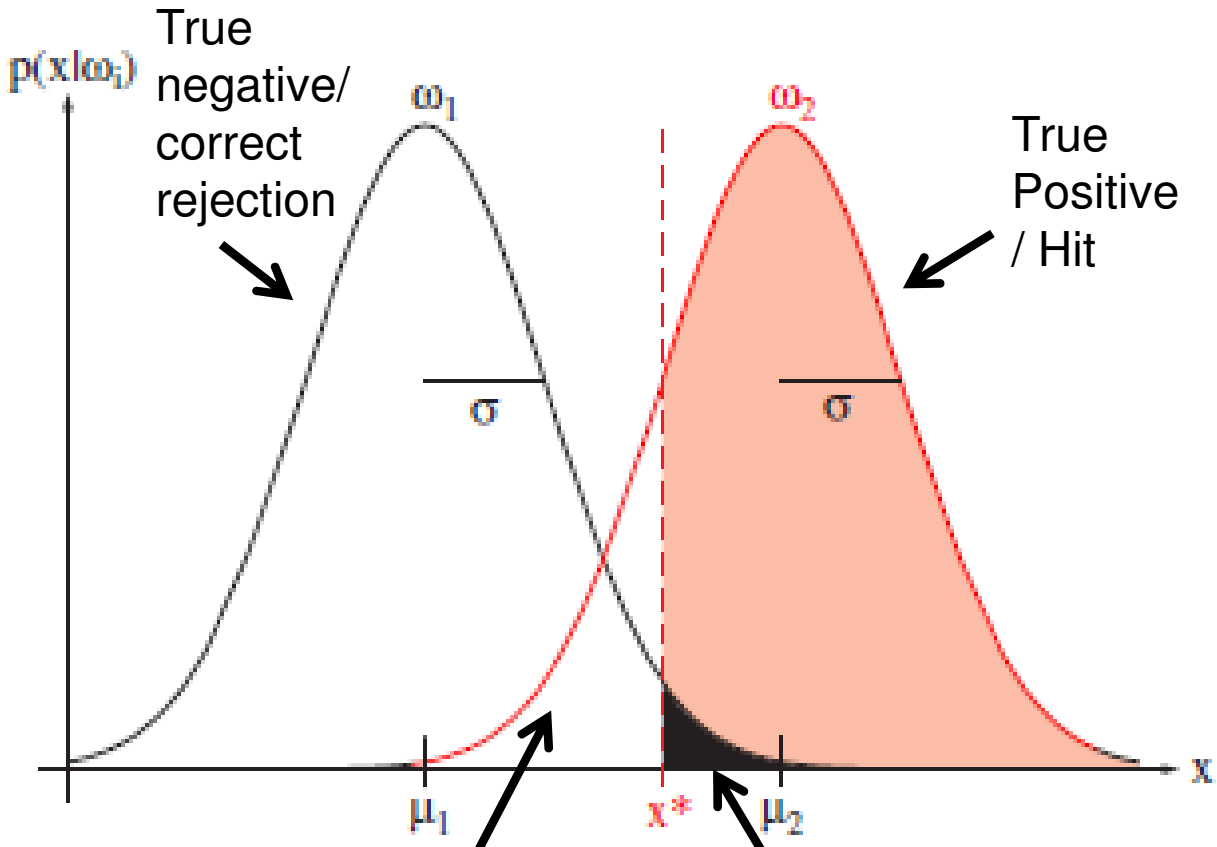
$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad i = 1,\ldots,d$$

and :

$$w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

decide ω_1 if $g(x) > 0$ and ω_2 if $g(x) \leq 0$

Evaluation of classifier



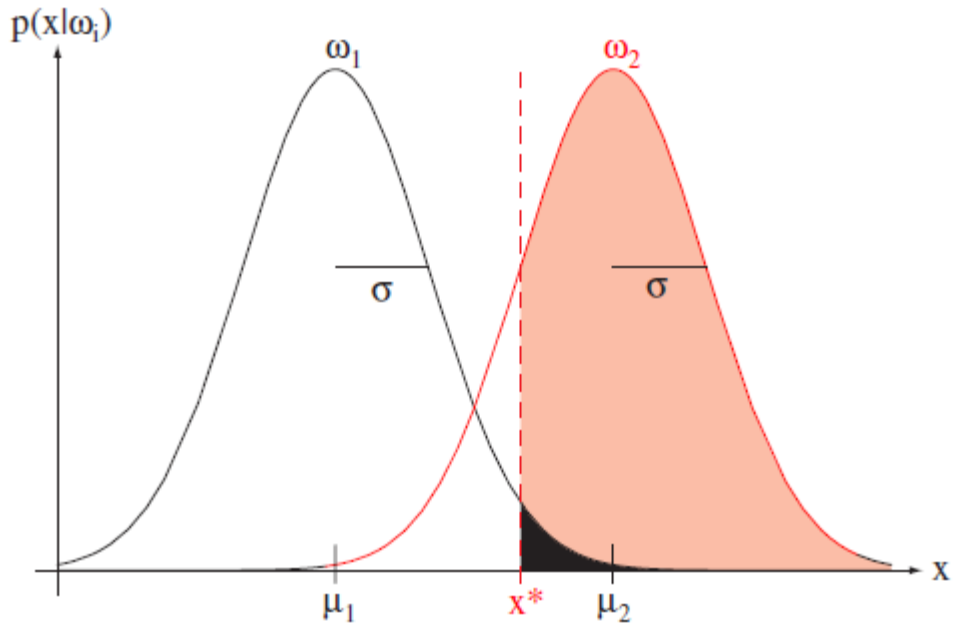
$P(x > x^*|x \in \omega_2)$: a *hit*
$P(x > x^*|x \in \omega_1)$: a *false alarm*
$P(x < x^*|x \in \omega_2)$: a *miss*
$P(x < x^*|x \in \omega_1)$: a *correct rejection*

Discriminability ratio

$$d' = \frac{|\mu_2 - \mu_1|}{\sigma}$$

Discriminability ratio between two distributions

A high d' is of course desirable.



Receiver operating characteristics

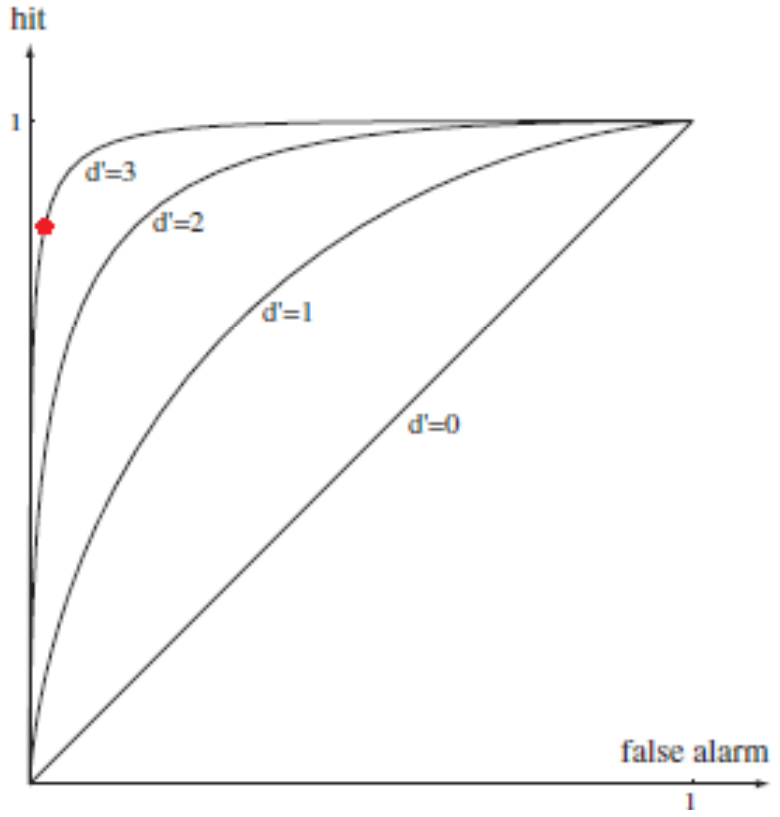
Consider a pair of distributions with discriminability ratio d'.

Vary threshold x^*.

We note that the true positive / hit probabilities and false positive / alarm probabilities will vary with a threshold x^*.

The variation of these probabilities for a given discriminability ratio present a smooth curve.

The curve is called 'Receiving Operating Characteristics'



$$d' = \frac{|\mu_2 - \mu_1|}{\sigma}$$