Bayesian Decision Theory

Mingmin Chi

SCS Fudan University, Shanghai, China

1. Introduction
2. Minimum-Error-Rate Classification
   - Prior
   - Decision by Likelihood
   - Decision by Posterior
3. Minimum-Risk Classification
4. Gaussian (Normal) Density
5. Discriminant Functions
6. Discriminant Function for the Normal Density
   - Case $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$
   - Common covariance matrix
   - Different $\mathbf{\Sigma}$ for each class
7. Error Probabilities
8. Receiver Operating Characteristic (ROC)

1. Introduction
2. Minimum-Error-Rate Classification
   - Prior
   - Decision by Likelihood
   - Decision by Posterior
3. Minimum-Risk Classification
4. Gaussian (Normal) Density
5. Discriminant Functions
6. Discriminant Function for the Normal Density
   - Case $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$
   - Common covariance matrix
   - Different $\mathbf{\Sigma}$ for each class
7. Error Probabilities
8. Receiver Operating Characteristic (ROC)

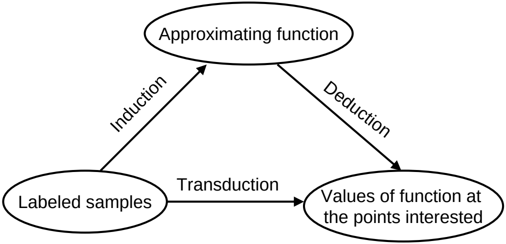
Learning Types

Imagine an organism or machine which experiences a series of sensory inputs: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \cdots$

- Supervised learning: The machine is also given desired outputs $y_1, y_2, y_3, \cdots$, and its goal is to learn to produce the correct output given a new input
- Unsupervised learning: The goal of the machine is to build a model of $\mathbf{x}$ that can be used for reasoning, decision making, predicting things, communicating etc.
- Reinforcement learning: The machine can also produce actions $a_1, a_2, , \cdots$ which affect the state of the world, and receives rewards (or punishments) $r_1, r_2, , \cdots$. Its goal is to learn to act in a way that maximizes rewards in the long term

Inference Types

- Inductive Learning (specific-to-general): Learning is a problem of function estimation on the basis of empirical data
- Transductive Learning (specific-to-specific): To estimate the values of the function for a given finite number of samples of interest



General Decision Theory

Foundation of pattern recognition is probability theory

- Minimize the expected number of misclassifications by assigning each input $\mathbf{x}$ to the class $\mathcal{C}_k$ which maximizes the posterior

$$P(\mathcal{C}_k|\mathbf{x}).$$

General Decision Theory

Foundation of pattern recognition is probability theory

- Minimize the expected number of misclassifications by assigning each input $\mathbf{x}$ to the class $\mathcal{C}_k$ which maximizes the posterior

$$P(\mathcal{C}_k|\mathbf{x}).$$

- Two phases
  1. Inference: model the posterior probabilities
  2. Decision: choose the optimal output

Generative Vs. Discriminative Models

- Generative approaches: separately model the class-conditional densities and the priors

$$p(\mathbf{x}|\mathcal{C}_k), \quad P(\mathcal{C}_k)$$

then evaluate the posterior with the Bayes' Theorem

$$P(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)P(\mathcal{C}_j)}$$

Generative Vs. Discriminative Models

- Generative approaches: separately model the class-conditional densities and the priors

$$p(\mathbf{x}|\mathcal{C}_k), \quad P(\mathcal{C}_k)$$

then evaluate the posterior with the Bayes' Theorem

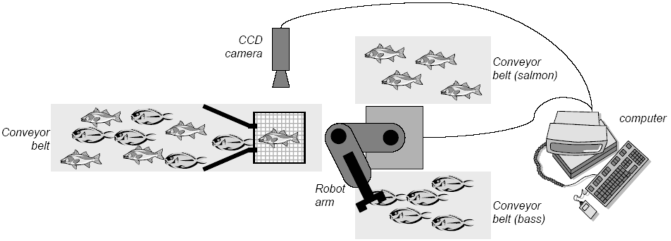
$$P(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)P(\mathcal{C}_j)}$$

- Discriminative approaches: directly model the posterior

$$P(\mathcal{C}_k|\mathbf{x})$$

Scenario

- Design of a classifier to separate two kinds of fish: sea bass and salmon
- What's the next emerging along the conveyor belt (prediction)?
- Does the sequence of types of fish appear to be random?



1. Introduction
2. Minimum-Error-Rate Classification
   - Prior
   - Decision by Likelihood
   - Decision by Posterior
3. Minimum-Risk Classification
4. Gaussian (Normal) Density
5. Discriminant Functions
6. Discriminant Function for the Normal Density
   - Case $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$
   - Common covariance matrix
   - Different $\mathbf{\Sigma}$ for each class
7. Error Probabilities
8. Receiver Operating Characteristic (ROC)

1. Introduction
2. Minimum-Error-Rate Classification
   - Prior
   - Decision by Likelihood
   - Decision by Posterior
3. Minimum-Risk Classification
4. Gaussian (Normal) Density
5. Discriminant Functions
6. Discriminant Function for the Normal Density
   - Case $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$
   - Common covariance matrix
   - Different $\mathbf{\Sigma}$ for each class
7. Error Probabilities
8. Receiver Operating Characteristic (ROC)

Decision by Prior

- The type of fish, or state of nature, or class, ω, is a random variable
- As each fish emerges nature is in one or the other of the two possible states

$$\omega = \begin{cases} \omega_1 & \text{if fish is sea bass} \\ \omega_2 & \text{if fish is salmon} \end{cases}$$

- As the ω is unpredictable, it must be described probabilistically
- Assuming there is some *a priori* probability (prior), which reflects our knowledge of how likely each type of fish will appear before we actually see it.

Prior

- Assuming that the catch of salmon and sea bass is equiprobable (uniform priors),

Prior

- Assuming that the catch of salmon and sea bass is equiprobable (uniform priors),
  - $P(\omega_1) = P(\omega_2)$

Prior

- Assuming that the catch of salmon and sea bass is equiprobable (uniform priors),
  - $P(\omega_1) = P(\omega_2)$
- Assume there are no other types of fish

Prior

- Assuming that the catch of salmon and sea bass is equiprobable (uniform priors),
  - $P(\omega_1) = P(\omega_2)$
- Assume there are no other types of fish
  - $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustivity)
- May use different values depending on the fishing area, time of the year, etc.

Decision with Prior

No more information available, if we are forced to make a decision

Decision with Prior

No more information available, if we are forced to make a decision

$$\omega = \begin{cases} \omega_1 & \text{if } P(\omega_1) > P(\omega_2) \\ \omega_2 & \text{if } P(\omega_1) < P(\omega_2) \\ \omega_1/\omega_2 & \text{if } P(\omega_1) = P(\omega_2) \end{cases}$$

1. Introduction
2. Minimum-Error-Rate Classification
   - Prior
   - Decision by Likelihood
   - Decision by Posterior
3. Minimum-Risk Classification
4. Gaussian (Normal) Density
5. Discriminant Functions
6. Discriminant Function for the Normal Density
   - Case $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$
   - Common covariance matrix
   - Different $\mathbf{\Sigma}$ for each class
7. Error Probabilities
8. Receiver Operating Characteristic (ROC)

Probability Density

- Let's try to improve the decision using the lightness measurement x
- different fish with different lightness values $x_1, x_2, \cdots$

$$\Rightarrow x \text{ is a random variable in probabilistic terms}$$

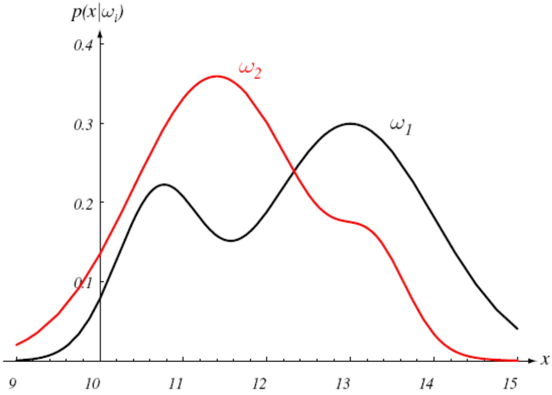
- Assume x to be a continuous random variable whose distribution depends on the state of nature:

$$p(x|\omega)$$

  which is class-conditional probability density of measuring a particular feature value x given the pattern is in category (class) ω
- likelihood: if all other things are equal, larger $p(x|\omega_i)$ is of more "likely" that the true category is ω_i

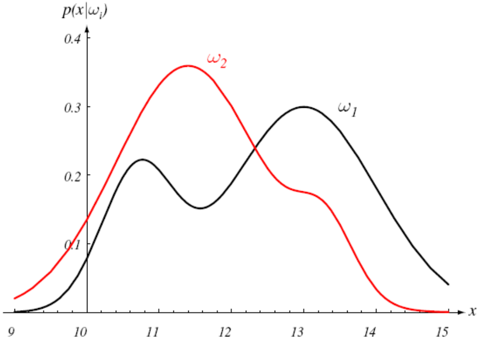
Probability Density

$p(x|\omega_1)$ and $p(x|\omega_2)$ describe the difference in lightness between population of sea bass and salmon



Decision by Likelihood

$$\omega = \begin{cases} \omega_1 & \text{if } p(x|\omega_1) > p(x|\omega_2) \\ \omega_2 & \text{if } p(x|\omega_1) < p(x|\omega_2) \\ \omega_1/\omega_2/\text{reject} & \text{if } p(x|\omega_1) = p(x|\omega_2) \end{cases}$$



1. Introduction
2. Minimum-Error-Rate Classification
   - Prior
   - Decision by Likelihood
   - Decision by Posterior
3. Minimum-Risk Classification
4. Gaussian (Normal) Density
5. Discriminant Functions
6. Discriminant Function for the Normal Density
   - Case $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$
   - Common covariance matrix
   - Different $\mathbf{\Sigma}$ for each class
7. Error Probabilities
8. Receiver Operating Characteristic (ROC)

Decision-theoretic terminology

- state of nature: ω
- *a priori* probability (prior) $P(\omega)$
- class-conditional probability density function $p(x|\omega_i)$

  the likelihood of ω_i wrt x
- *a posteriori* probability $P(\omega_i|x)$: the probability of the state of nature being ω_i given that feature value x has been measured

Posterior

Posterior

- Product of the likelihood and the prior probability
- Bayes formula:

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

$\Rightarrow$

Posterior

Posterior

- Product of the likelihood and the prior probability
- Bayes formula:

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

$\Rightarrow$

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

Making Decision

- Suppose we know the likelihood and the prior probability
- How can we make a decision after observing the value of x?

Making Decision

- Suppose we know the likelihood and the prior probability
- How can we make a decision after observing the value of x?

$$\omega = \begin{cases} \omega_1 & \text{if } P(\omega_1|x) > P(\omega_2|x) \\ \omega_2 & \text{otherwise} \end{cases}$$

- We can rewrite the decision rule by

$$\omega = \begin{cases} \omega_1 & \text{if } \frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)} \\ \omega_2 & \text{otherwise} \end{cases}$$

Probability of the Error

What is the probability of error for this decision:

Probability of the Error

What is the probability of error for this decision:

$$P(\text{error}|x) =$$

Probability of the Error

What is the probability of error for this decision:

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

What is the average probability of error?

Probability of the Error

What is the probability of error for this decision:

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

What is the average probability of error?

$$P(\text{error})$$

Probability of the Error

What is the probability of error for this decision:

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

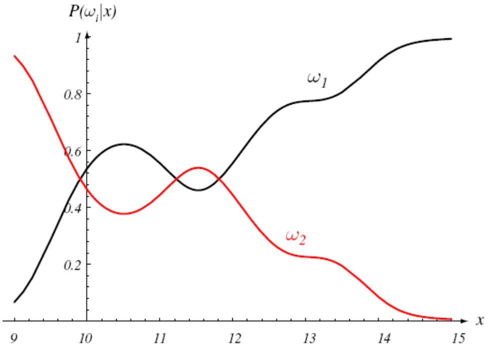
What is the average probability of error?

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

Bayes decision rule minimizes this error since

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

Bayes Decision Rule



MAP and MLE

Maximum a posteriori (MAP)

$$\text{Decide } \omega_1 \text{ if } P(\omega_1|x) \ > \ P(\omega_2|x)$$

MAP and MLE

Maximum a posteriori (MAP)

$$\text{Decide } \omega_1 \text{ if } P(\omega_1|x) \; > \; P(\omega_2|x)$$

$$P(\omega_k|x) = \frac{p(x|\omega_k)P(\omega_k)}{p(x)} \quad \Downarrow \quad p(x) = \sum_{i=1}^{2} p(x|\omega_i)P(\omega_i)$$

MAP and MLE

Maximum a posteriori (MAP)

$$\text{Decide } \omega_1 \text{ if } P(\omega_1|x) \;>\; P(\omega_2|x)$$

$$P(\omega_k|x) = \frac{p(x|\omega_k)P(\omega_k)}{p(x)} \;\Downarrow\; p(x) = \sum_{i=1}^{2} p(x|\omega_i)P(\omega_i)$$

Maximum Likelihood Estimation (MLE)

$$\text{Decide } \omega_1 \text{ if } p(x|\omega_1)P(\omega_1) \;>\; p(x|\omega_2)P(\omega_2)$$

$$\Downarrow \; P(\omega_1) = P(\omega_2)$$

$$\text{Decide } \omega_1 \text{ if } p(x|\omega_1) \;>\; p(x|\omega_2)$$

1. Introduction
2. Minimum-Error-Rate Classification
   - Prior
   - Decision by Likelihood
   - Decision by Posterior
3. **Minimum-Risk Classification**
4. Gaussian (Normal) Density
5. Discriminant Functions
6. Discriminant Function for the Normal Density
   - Case $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$
   - Common covariance matrix
   - Different $\mathbf{\Sigma}$ for each class
7. Error Probabilities
8. Receiver Operating Characteristic (ROC)

How can we generalize to

- more than one feature:

How can we generalize to

- more than one feature:
  - replacing the scalar x by the *feature vector* $\mathbf{x} \in \mathcal{R}^d$

How can we generalize to

- more than one feature:
  - replacing the scalar x by the *feature vector* $\mathbf{x} \in \mathcal{R}^d$
- more than two states of nature
  - multiple classes

How can we generalize to

- more than one feature:
  - replacing the scalar x by the *feature vector* $\mathbf{x} \in \mathcal{R}^d$
- more than two states of nature
  - multiple classes
- allowing actions other than just decision
  - allowing the possibility of rejection

How can we generalize to

- more than one feature:
  - replacing the scalar x by the *feature vector* $\mathbf{x} \in \mathcal{R}^d$
- more than two states of nature
  - multiple classes
- allowing actions other than just decision
  - allowing the possibility of rejection
- different risks for the decision
  - define how costly each action is

- Let $\mathbf{x}$ be the d-dimensional random variable, called the feature vector

- Let $\mathbf{x}$ be the d-dimensional random variable, called the feature vector
- Let $\{\omega_1, \omega_2, \cdots, \omega_c\}$ be the finite set of the c states of nature ("categories" or "classes")

- Let $\mathbf{x}$ be the d-dimensional random variable, called the feature vector
- Let $\{\omega_1, \omega_2, \cdots, \omega_c\}$ be the finite set of the c states of nature ("categories" or "classes")
- Let $\{\alpha_1, \cdots, \alpha_a\}$ be the finite set of a possible actions

- Let **x** be the d-dimensional random variable, called the feature vector
- Let $\{\omega_1, \omega_2, \cdots, \omega_c\}$ be the finite set of the *c* states of nature ("categories" or "classes")
- Let $\{\alpha_1, \cdots, \alpha_a\}$ be the finite set of *a* possible actions
- Let $\lambda(\alpha_i|\omega_j)$ be the loss incurred for taking action α_i when the state of nature is ω_j

Bayesian Decision Theory

Posterior

- $P(\omega_i)$ is the prior probability when the state of nature is ω_i
- $p(x|\omega_i)$ is the class-conditional probability density function
- $P(\omega_i|x)$ is the posterior probability which can be computed by

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

Conditional Risk

- Suppose we observe $\mathbf{x}$ and take action α_i
- If the true state of nature is ω_j, we incur the loss $\lambda(\alpha_i|\omega_j)$
- the expected loss with taking action α_i

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

which is also called the conditional risk

Minimum Risk Classification

- The general decision rule $\boldsymbol{\alpha}(\mathbf{x})$ tells us which action $(\alpha_i, i = 1, \cdots, a)$ to take for every possible observation
- We want to find the decision rule that minimizes the overall risk

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

- Bayes decision rule minimizes the overall risk by selecting the action α_i when $R(\alpha_i|\mathbf{x})$ is the minimum
- The resulting minimum overall risk is called the Bayes risk and is the best performance that can be achieved.

Binary-Class Case: Conditional risk

- α_1: deciding ω_1
- α_2: deciding ω_2
- $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$:

Binary-Class Case: Conditional risk

- α_1: deciding ω_1
- α_2: deciding ω_2
- $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$: loss incurred for deciding ω_i when the true state of nature is ω_j

Binary-Class Case: Conditional risk

- α_1: deciding ω_1
- α_2: deciding ω_2
- $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$: loss incurred for deciding ω_i when the true state of nature is ω_j

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 $i \neq j$

Conditional risk:

- $R(\alpha_1|\mathbf{x}) = \lambda_{11}P(\omega_1|\mathbf{x}) +$

Binary-Class Case: Conditional risk

- α_1: deciding ω_1
- α_2: deciding ω_2
- $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$: loss incurred for deciding ω_i when the true state of nature is ω_j

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 $i \neq j$

Conditional risk:

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

- α_1: deciding ω_1
- α_2: deciding ω_2
- $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$

Minimum-risk decision rule:

$$\text{Decide} \begin{cases} \omega_1 & \text{if } R(\alpha_1|\mathbf{x}) \; < \; R(\alpha_2|\mathbf{x}) \\ \omega_2 & \text{otherwise} \end{cases}$$

This corresponds to

$$\begin{aligned} & R(\alpha_1|\mathbf{x}) \; < \; R(\alpha_2|\mathbf{x}) \\ & \Rightarrow \lambda_{21} - \lambda_{11})P(\omega_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(\omega_2|\mathbf{x}) \\ & \Rightarrow \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)} \end{aligned}$$

Likelihood Ratio

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

- the form of decision rule focuses on the **x**-dependence of probability densities
- likelihood ratio exceeds a threshold value that is independent of the observation **x**

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

Recall: $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$

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 $i \neq j$

- Define the zero-one loss

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad i,j = 1, \cdots, c$$

  (all the errors are equally costly)

- Conditional risk becomes

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

Minimum Error Rate

$$R(\alpha_i|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$$

- Minimizing the risk requires maximizing $P(\omega_i|\mathbf{x})$ and results in the minimum-error decision rule

$$\text{Decide } \omega_i \text{ if } P(\omega_i|\mathbf{x}) \; > \; P(\omega_j|\mathbf{x}), \; \forall \, j \neq i$$

- The resulting error is called the Bayes error and is the best performance that can be achieved

Example

Recall likelihood ratio

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda \text{ then}$$

Example

Recall likelihood ratio

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda \text{ then decide } \omega_1 \text{ if } \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \theta_\lambda$$

- If Λ is the zero-one loss function,

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} =$$

Example

Recall likelihood ratio

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda \text{ then decide } \omega_1 \text{ if } \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \theta_\lambda$$

- If Λ is the zero-one loss function,

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

Example

Recall likelihood ratio

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda \text{ then decide } \omega_1 \text{ if } \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \theta_\lambda$$

- If Λ is the zero-one loss function,

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \theta_\lambda = \frac{P(\omega_2)}{P(\omega_1)} = \theta_a$$

- If loss function penalizes misclassifying ω_2 as ω_1 more than the converse, say, 1.2 folds,

$$\Lambda =$$

Example

Recall likelihood ratio

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}\frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda \text{ then decide } \omega_1 \text{ if } \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \theta_\lambda$$

- If Λ is the zero-one loss function,

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \theta_\lambda = \frac{P(\omega_2)}{P(\omega_1)} = \theta_a$$

- If loss function penalizes misclassifying ω_2 as ω_1 more than the converse, say, 1.2 folds,

$$\Lambda = \begin{pmatrix} 0 & 1.2 \\ 1 & 0 \end{pmatrix},$$

Example

Recall likelihood ratio

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda \text{ then decide } \omega_1 \text{ if } \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \theta_\lambda$$

- If Λ is the zero-one loss function,

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \theta_\lambda = \frac{P(\omega_2)}{P(\omega_1)} = \theta_a$$

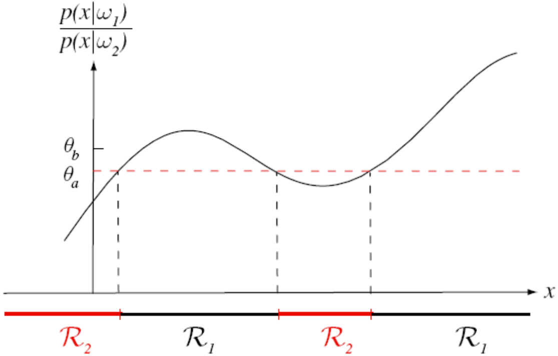
- If loss function penalizes misclassifying ω_2 as ω_1 more than the converse, say, 1.2 folds,

$$\Lambda = \begin{pmatrix} 0 & 1.2 \\ 1 & 0 \end{pmatrix}, \quad \theta_\lambda = \frac{1.2P(\omega_2)}{P(\omega_1)} = \theta_b$$

Example

Table: A cost matrix.

	actual normal	actual cancer
predicted normal	$\lambda_{11} = 0$	$\lambda_{12} = 1.2$
predicted cancer	$\lambda_{21} = 1$	$\lambda_{22} = 0$



1. Introduction
2. Minimum-Error-Rate Classification
   - Prior
   - Decision by Likelihood
   - Decision by Posterior
3. Minimum-Risk Classification
4. **Gaussian (Normal) Density**
5. Discriminant Functions
6. Discriminant Function for the Normal Density
   - Case $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}$
   - Common covariance matrix
   - Different $\mathbf{\Sigma}$ for each class
7. Error Probabilities
8. Receiver Operating Characteristic (ROC)

Gaussian can be considered as a model where the feature vectors for a given class are continuous-valued, randomly corrupted versions of a single typical or prototype vector

Some properties of the Gaussian

- analytically tractable
- Completely specified by the 1st and 2nd moments
- A lot of processes are asymptotically Gaussian (Central Limit Theorem)
- Linear transformations of a Gaussian are also Gaussian
- Uncorrelatedness implies independence

Univariate Density

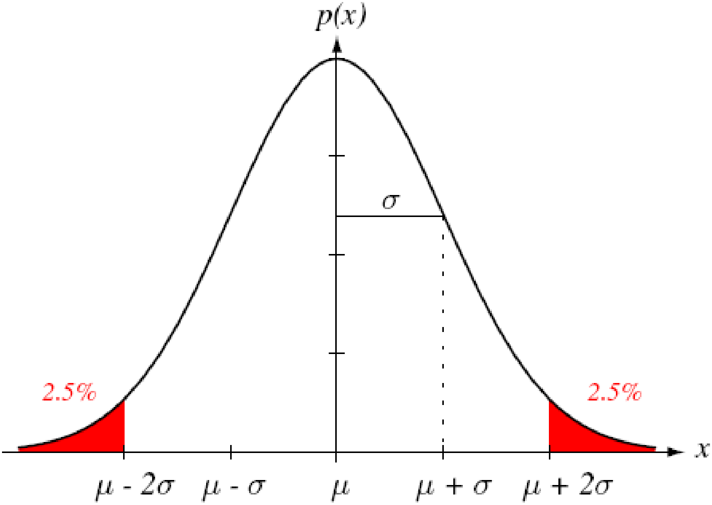
For $x \in \mathcal{R}$

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

where

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

Univariate Density



Multivariate Density

For $\mathbf{x} \in \mathcal{R}^d$

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

where

$$\boldsymbol{\mu} = E[\mathbf{x}] = \int_{-\infty}^{\infty} \mathbf{x}p(\mathbf{x})d\mathbf{x}$$
$$\boldsymbol{\Sigma} = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top] = \int_{-\infty}^{\infty} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top p(\mathbf{x})d\mathbf{x}$$

statistically independent,

Multivariate Density

For $\mathbf{x} \in \mathcal{R}^d$

$$
\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \frac{1}{2\pi^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]
\end{aligned}
$$

where

$$
\boldsymbol{\mu} = E[\mathbf{x}] = \int_{-\infty}^{\infty} \mathbf{x} p(\mathbf{x}) d\mathbf{x}
$$

$$
\boldsymbol{\Sigma} = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top] = \int_{-\infty}^{\infty} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top p(\mathbf{x}) d\mathbf{x}
$$

statistically independent, $\sigma_{ij} = 0$

Linear Transformation

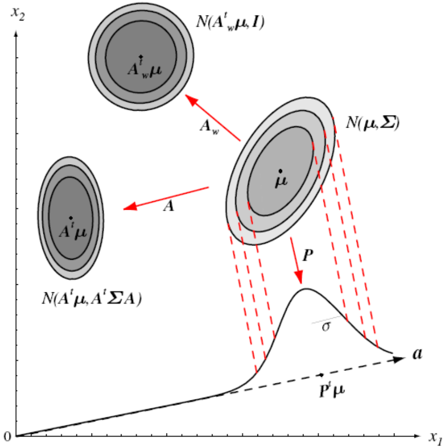
The linear transformation of a Gaussian is also Gaussian

- $p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- $\mathbf{z} = \mathbf{A}^\top \mathbf{x}, \quad \mathbf{A} \in \mathcal{R}^{d \times k}$
- $p(\mathbf{z}) =$

Linear Transformation

The linear transformation of a Gaussian is also Gaussian

- $p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- $\mathbf{z} = \mathbf{A}^\top \mathbf{x}, \quad \mathbf{A} \in \mathcal{R}^{d \times k}$
- $p(\mathbf{z}) = \mathcal{N}(\mathbf{A}^\top \boldsymbol{\mu}, \mathbf{A}^\top \boldsymbol{\Sigma} \mathbf{A})$



x_2
$N(A_w^t\mu, I)$
$A_w^t\mu$
A_w
$N(\mu, \Sigma)$
μ
A
$A^t\mu$
$N(A^t\mu, A^t\Sigma A)$
P
σ
a
$P^t\mu$
0
x_1

Projection onto a line

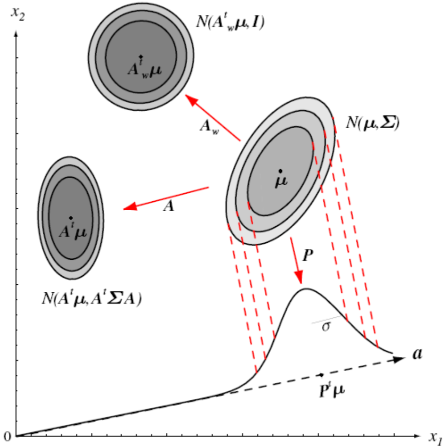
Remember $\mathbf{z} = \mathbf{P}^\top \mathbf{x}, \quad \mathbf{P} \in \mathcal{R}^{d \times k}$

- if $k = 1$ and $\mathbf{P}$ is a unit-length vector, $\mathbf{a}$
- then,

Projection onto a line

Remember $\mathbf{z} = \mathbf{P}^\top \mathbf{x}, \quad \mathbf{P} \in \mathcal{R}^{d \times k}$

- if $k = 1$ and $\mathbf{P}$ is a unit-length vector, $\mathbf{a}$
- then, $z = \mathbf{a}^\top \mathbf{x}$ is a scalar, representing a projection of $\mathbf{x}$ onto a line in the direction $\mathbf{a}$



x_2
$N(A^t_w\mu, I)$
$A^t_w\mu$
A_w
$N(\mu,\Sigma)$
μ
A
$A^t\mu$
$N(A^t\mu, A^t\Sigma A)$
P
σ
a
$P^t\mu$
0
x_1

Whitening Transformation

Coordinate transformation

- arbitrarily Gaussian distribution $\rightarrow$ a spherical one
- $p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \rightarrow p(\mathbf{z}) =$

Whitening Transformation

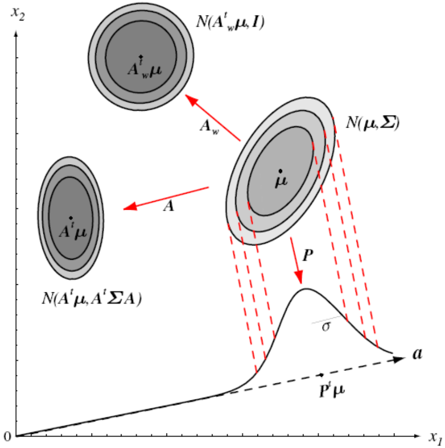
Coordinate transformation

- arbitrarily Gaussian distribution $\rightarrow$ a spherical one
- $p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \rightarrow p(\mathbf{z}) = \mathcal{N}(\mathbf{A}_w^\top \boldsymbol{\mu}, \mathbf{I}_d)$
- finding $\mathbf{A}_w$

Whitening Transformation

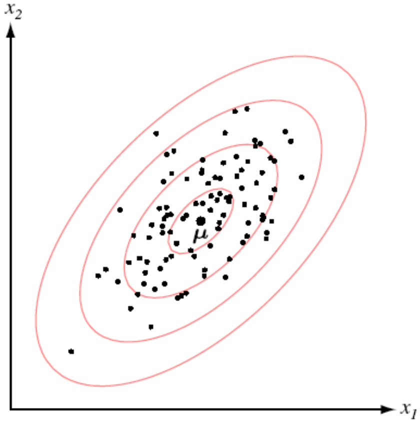
Coordinate transformation

- arbitrarily Gaussian distribution $\rightarrow$ a spherical one
- $p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \rightarrow p(\mathbf{z}) = \mathcal{N}(\mathbf{A}_w^\top \boldsymbol{\mu}, \mathbf{I}_d)$
- finding $\mathbf{A}_w$
    - $\mathbf{A}_w^\top \boldsymbol{\Sigma} \mathbf{A}_w = \mathbf{I}_d$
    - $\boldsymbol{\Sigma} = \mathbf{U} \Lambda \mathbf{U}^\top$
    - $\mathbf{A}_w = \mathbf{U} \Lambda^{-1/2}$


x_2
$N(A^t_w\mu, I)$
$A^t_w\mu$
A_w
$N(\mu,\Sigma)$
μ
A
$A^t\mu$
$N(A^t\mu, A^t\Sigma A)$
P
σ
a
$P^t\mu$
0
x_1

Cloud

- cluster
- position: mean
- shape: covariance matrix
- $\#$ parameters: $d + d(d-1)/2$



Mahalanobis distance

- the shape of data points with equal density is a hyperellipsoid (locus of points of constant density)
- principle axes of hyperellipsoids:

Mahalanobis distance

- the shape of data points with equal density is a hyperellipsoid (locus of points of constant density)
- principle axes of hyperellipsoids: eigenvector of $\mathbf{\Sigma}$
- distance for $\mathbf{x}$ to $\boldsymbol{\mu}$: $r^2 = (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$, called squared Mahalanobis distance
  - determining similarity of an unknown sample set to a known one
  - scale-invariant, i.e. not dependent on the scale of measurements
  - dissimilarity measure between two random vectors with covariance matrix $\mathbf{\Sigma}$: $(\mathbf{x}_1 - \mathbf{x}_2)^\top \mathbf{\Sigma}^{-1} (\mathbf{x}_1 - \mathbf{x}_2)$
    - $\mathbf{\Sigma} = \mathbf{I}$, Euclidean distance
    - $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \cdots, \sigma_d)$, normalized Euclidean distance

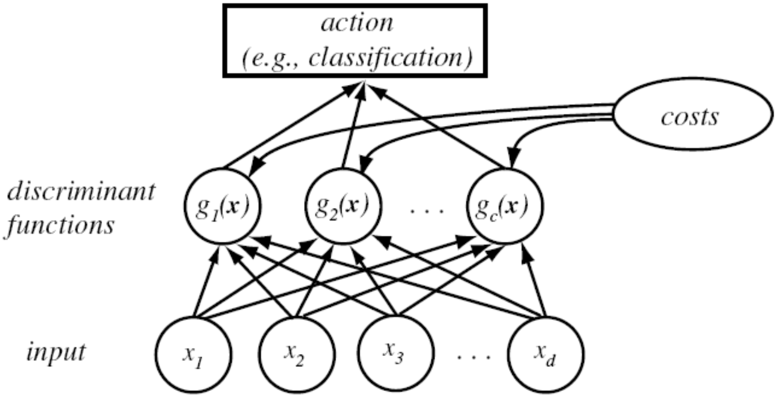
1. Introduction
2. Minimum-Error-Rate Classification
   - Prior
   - Decision by Likelihood
   - Decision by Posterior
3. Minimum-Risk Classification
4. Gaussian (Normal) Density
5. **Discriminant Functions**
6. Discriminant Function for the Normal Density
   - Case $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$
   - Common covariance matrix
   - Different $\mathbf{\Sigma}$ for each class
7. Error Probabilities
8. Receiver Operating Characteristic (ROC)

- A useful way of representing classifiers is through discriminant functions $g_i(\mathbf{x}),\ i = 1, \cdots, c$, where the classifier assigns a feature vector $\mathbf{x}$ to class ω_i if

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

or

$$\mathbf{x} \in \omega_m \text{ if and only if } g_m(\mathbf{x}) = \arg \max_{i=1,\cdots,c} \{g_i(\mathbf{x})\}$$


action
(e.g., classification)
costs
discriminant
functions
$g_1(\mathbf{x})$
$g_2(\mathbf{x})$
. . .
$g_c(\mathbf{x})$
input
x_1
x_2
x_3
. . .
x_d

- For the classifier that minimizes conditional risk

- For the classifier that minimizes conditional risk

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

- For the classifier that minimizes conditional risk

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

- For the classifier that minimizes error

- For the classifier that minimizes conditional risk

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

- For the classifier that minimizes error

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

- These functions divide the feature space into c decision regions, $\mathcal{R}_1, \mathcal{R}_2, \cdots, \mathcal{R}_c$, separated by decision boundaries
- the choice of discriminant function is not unique
  - multiplicative or additive operations without influencing the decision
  - with a monotonically increasing function $f(\cdot)$, replacing $g_i(\mathbf{x})$ by $f(g_i(\mathbf{x}))$, i.e.,

$$\begin{aligned} g_i(\mathbf{x}) &= p(\mathbf{x}|\omega_i)P(\omega_i) \Rightarrow \\ \ln g_i(\mathbf{x}) &= \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i) \end{aligned}$$

  - This may lead to significant analytical and computational simplifications

Example for the Two-category case

- special case of multi-category case, dichotomizer
- discriminant function:

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

$$\omega = \begin{cases} \omega_1, & \text{if } g(x) > 0 \\ \omega_2, & \text{otherwise} \end{cases}$$

- for minimum-error-rate discriminant function,

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

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

1. Introduction
2. Minimum-Error-Rate Classification
   - Prior
   - Decision by Likelihood
   - Decision by Posterior
3. Minimum-Risk Classification
4. Gaussian (Normal) Density
5. Discriminant Functions
6. Discriminant Function for the Normal Density
   - Case $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$
   - Common covariance matrix
   - Different $\mathbf{\Sigma}$ for each class
7. Error Probabilities
8. Receiver Operating Characteristic (ROC)

- Recall that the minimum error-rate classification can be achieved by the discriminant function

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

- For multivariate normal density, $p(\mathbf{x}|\omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$
- so we can have

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

1. Introduction
2. Minimum-Error-Rate Classification
   - Prior
   - Decision by Likelihood
   - Decision by Posterior
3. Minimum-Risk Classification
4. Gaussian (Normal) Density
5. Discriminant Functions
6. Discriminant Function for the Normal Density
   - Case $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$
   - Common covariance matrix
   - Different $\mathbf{\Sigma}$ for each class
7. Error Probabilities
8. Receiver Operating Characteristic (ROC)

Linear Discriminant Function (1)

- features are statistically independent
- each feature has the same variance, σ^2
- Recall that the discriminant functions

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma}_i| + \ln P(\omega_i)$$

Linear Discriminant Function (1)

- features are statistically independent
- each feature has the same variance, σ^2
- Recall that the discriminant functions

$$
\begin{aligned}
g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma}_i| + \ln P(\omega_i) \\
\Rightarrow &\quad -\frac{||\mathbf{x}-\boldsymbol{\mu}_i||^2}{2\sigma^2} + \ln P(\omega_i)
\end{aligned}
$$

where $||\cdot||$ denotes the Euclidean norm, i.e.,

$$
||\mathbf{x}-\boldsymbol{\mu}_i||^2 = (\mathbf{x}-\boldsymbol{\mu}_i)^\top(\mathbf{x}-\boldsymbol{\mu}_i)
$$

Linear Discriminant Function (2)

- $\mathbf{x}$ equal distance to mean, optimal decision by a priori
- otherwise, not necessary to compute distance

$$
\begin{aligned}
g_i(\mathbf{x}) &= -\frac{||\mathbf{x}-\boldsymbol{\mu}_i||^2}{2\sigma^2} + \ln P(\omega_i) \\
&= -\frac{1}{2\sigma^2}[\mathbf{x}^\top\mathbf{x} - 2\boldsymbol{\mu}_i^\top\mathbf{x} + \boldsymbol{\mu}_i^\top\boldsymbol{\mu}_i] + \ln P(\omega_i) \\
&\propto -\frac{1}{2\sigma^2}[-2\boldsymbol{\mu}_i^\top\mathbf{x} + \boldsymbol{\mu}_i^\top\boldsymbol{\mu}_i] + \ln P(\omega_i) \\
&= \mathbf{w}_i^\top\mathbf{x} + w_{i0}
\end{aligned}
$$

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

Linear Machine (1)

- A classifier that uses linear discriminant functions is called a linear machine
- decision surfaces $\mathcal{R}_i, \mathcal{R}_j$ are pieces of hyperplanes defined by

$$g_i(\mathbf{x}) \equiv g_j(\mathbf{x}) \Rightarrow \quad \mathbf{w}^\top(\mathbf{x} - \mathbf{x}_0) = 0$$

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

Decision hyperplane

- separating $\mathcal{R}_i$ and $\mathcal{R}_j$
- through the point $\mathbf{x}_0$
- orthogonal to $\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ and so

Linear Machine (1)

- A classifier that uses linear discriminant functions is called a linear machine
- decision surfaces $\mathcal{R}_i, \mathcal{R}_j$ are pieces of hyperplanes defined by

$$g_i(\mathbf{x}) \equiv g_j(\mathbf{x}) \Rightarrow \quad \mathbf{w}^\top(\mathbf{x} - \mathbf{x}_0) = 0$$

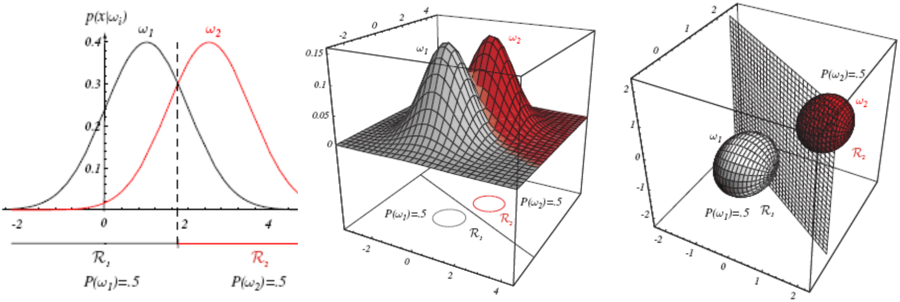
$$\text{where} \begin{cases} \mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \\ \mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||^2} \ln \frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \end{cases}$$

Decision hyperplane

- separating $\mathcal{R}_i$ and $\mathcal{R}_j$
- through the point $\mathbf{x}_0$
- orthogonal to $\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ and so orthogonal to the line linking the means

Equal Prior

$P(\omega_i) = P(\omega_j)$, the point $\mathbf{x}_0$ is halfway between the means, and the hyperplane is the perpendicular bisector of the line between the means



Equal Priors for multiple classes

If the priors are the same for all c classes, the ln $P(\omega_i)$ term becomes unimportant constant

minimum-distance classifier

- measure the Euclidean distance $d_i = ||\mathbf{x} - \boldsymbol{\mu}_i||$
- $\mathbf{x} \in \omega_m$, $d_m = \arg\min_{i=1,\cdots,c} d_i$

If mean as ideal prototype or template, template-matching

Unequal Prior

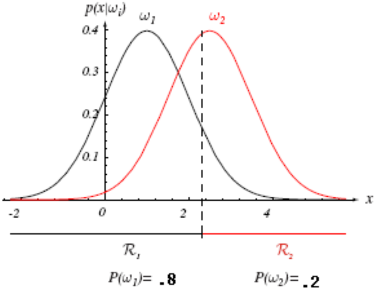
$$g_i(\mathbf{x}) \equiv g_j(\mathbf{x}) \Rightarrow \quad \mathbf{w}^\top(\mathbf{x} - \mathbf{x}_0) = 0$$

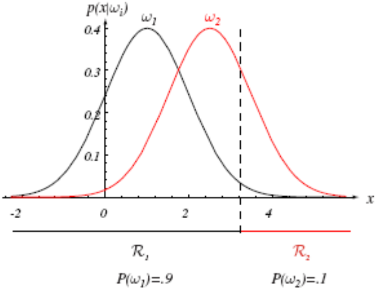
$$\text{where} \begin{cases} \mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \\ \mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||^2} \ln \frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \end{cases}$$

$P(\omega_i) \neq P(\omega_j)$

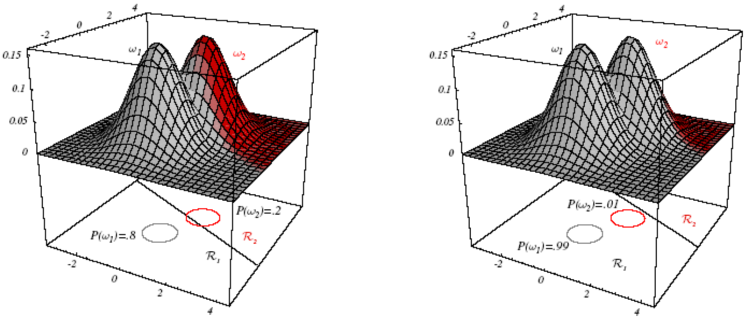
- the point $\mathbf{x}_0$ shifts away from the more likely mean
- if $\frac{\sigma^2}{||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||^2}$ is small, the position of the decision boundary is relatively insensitive to the exact values of the prior probability

Unequal Prior (1-d)

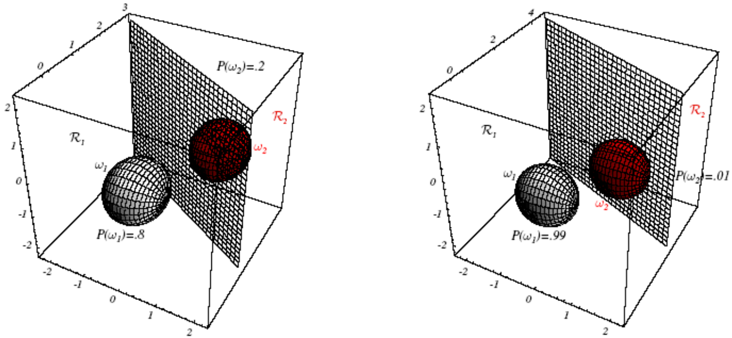




Unequal Prior (2-d)



Unequal Prior (3-d)



1. Introduction
2. Minimum-Error-Rate Classification
   - Prior
   - Decision by Likelihood
   - Decision by Posterior
3. Minimum-Risk Classification
4. Gaussian (Normal) Density
5. Discriminant Functions
6. Discriminant Function for the Normal Density
   - Case $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}$
   - Common covariance matrix
   - Different $\mathbf{\Sigma}$ for each class
7. Error Probabilities
8. Receiver Operating Characteristic (ROC)

Linear discriminant

Common covariance matrix, i.e., $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$

Discriminant function is

$$
\begin{aligned}
g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) \ + \ \ln P(\omega_i) \\
&= -\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} \ + \ \boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \ + \ \ln P(\omega_i) \\
&\propto \boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \ + \ \ln P(\omega_i) \\
&= \mathbf{w}_i^\top \mathbf{x} + w_{i0}
\end{aligned}
$$

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

Decision boundary (1)

If $\mathcal{R}_i$ and $\mathcal{R}_j$ are continuous, we have

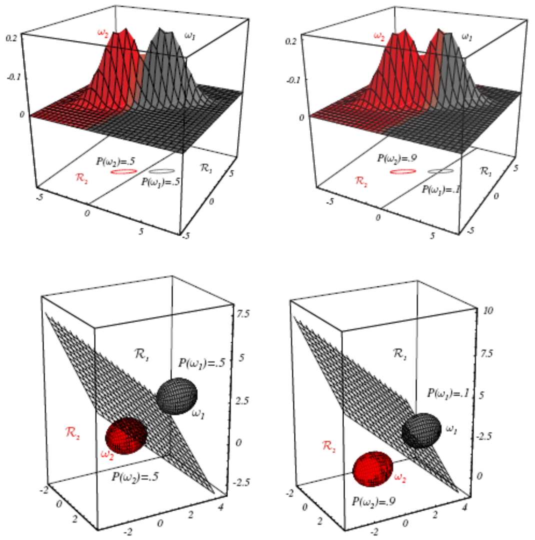
$$\mathbf{w}^\top(\mathbf{x} - \mathbf{x}_0) = 0$$

where

$$\begin{aligned}
\mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \\
\mathbf{x}_0 &= \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln(P(\omega_i)/P(\omega_j))}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)
\end{aligned}$$

the hyperplane passes through $\mathbf{x}_0$ but is necessarily not orthogonal to the line between the means

Decision boundary (2)



1. Introduction
2. Minimum-Error-Rate Classification
   - Prior
   - Decision by Likelihood
   - Decision by Posterior
3. Minimum-Risk Classification
4. Gaussian (Normal) Density
5. Discriminant Functions
6. Discriminant Function for the Normal Density
   - Case $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$
   - Common covariance matrix
   - Different $\mathbf{\Sigma}$ for each class
7. Error Probabilities
8. Receiver Operating Characteristic (ROC)

Discriminant function

$$
\begin{aligned}
g_i(\mathbf{x}) &= -(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) \;+\; \ln P(\omega_i) \\
&= -\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}_i^{-1}\mathbf{x} \;+\; \boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}_i^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i \;+\; \ln P(\omega_i) \\
&= \mathbf{x}^\top \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^\top \mathbf{x} + w_{i0}
\end{aligned}
$$

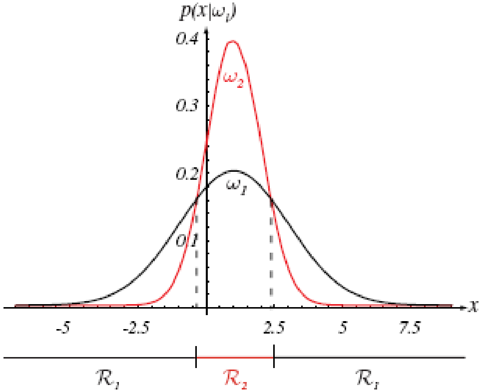
where

$$
\begin{aligned}
\mathbf{W}_i &= -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}, \; \mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i \\
w_{i0} &= -\frac{1}{2}\boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \;-\; \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| \;+\; \ln P(\omega_i)
\end{aligned}
$$

g_i is quadratic discriminant

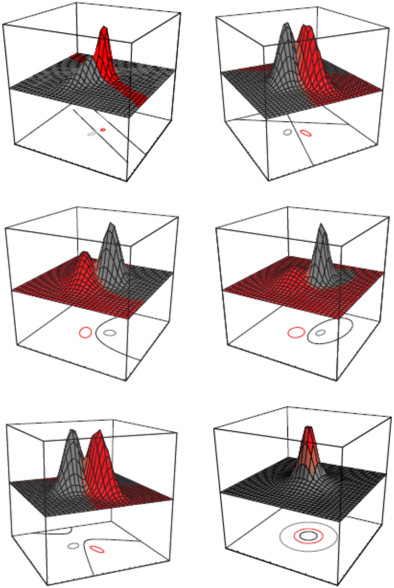
Decision boundary are hyperquadrics,

Decision boundary (1-d)

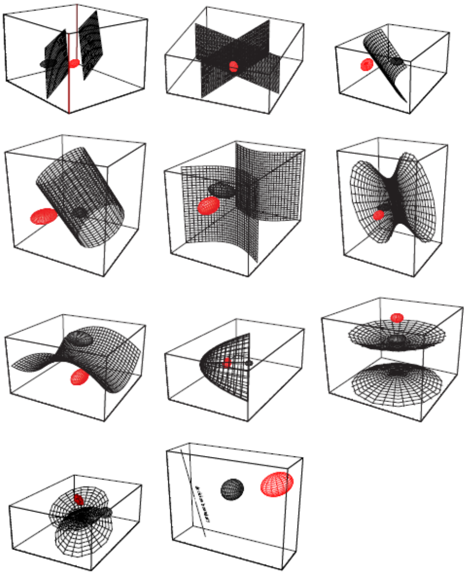


Equal prior

Decision boundary (2-d)



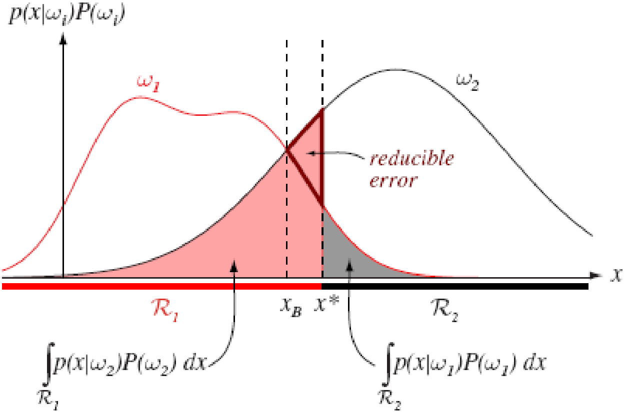
Decision boundary (3-d)



1. Introduction
2. Minimum-Error-Rate Classification
   - Prior
   - Decision by Likelihood
   - Decision by Posterior
3. Minimum-Risk Classification
4. Gaussian (Normal) Density
5. Discriminant Functions
6. Discriminant Function for the Normal Density
   - Case $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$
   - Common covariance matrix
   - Different $\mathbf{\Sigma}$ for each class
7. **Error Probabilities**
8. Receiver Operating Characteristic (ROC)

- For binary classification

$$
\begin{aligned}
P(\text{error}) &= P(\mathbf{x} \in \mathcal{R}_2, \omega_1) + P(\mathbf{x} \in \mathcal{R}_1, \omega_2) \\
&= P(\mathbf{x} \in \mathcal{R}_2 | \omega_1) P(\omega_1) + P(\mathbf{x} \in \mathcal{R}_1, \omega_2) P(\omega_2) \\
&= \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1) P(\omega_1) d\mathbf{x} + \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2) P(\omega_2) d\mathbf{x}
\end{aligned}
$$


$p(x|\omega_i)P(\omega_i)$
ω_1
ω_2
reducible error
x
$\mathcal{R}_1$
x_B
x^*
$\mathcal{R}_2$
$\int_{\mathcal{R}_1} p(x|\omega_2)P(\omega_2)\,dx$
$\int_{\mathcal{R}_2} p(x|\omega_1)P(\omega_1)\,dx$

- For multicategory classification

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

1. Introduction
2. Minimum-Error-Rate Classification
   - Prior
   - Decision by Likelihood
   - Decision by Posterior
3. Minimum-Risk Classification
4. Gaussian (Normal) Density
5. Discriminant Functions
6. Discriminant Function for the Normal Density
   - Case $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$
   - Common covariance matrix
   - Different $\mathbf{\Sigma}$ for each class
7. Error Probabilities
8. Receiver Operating Characteristic (ROC)

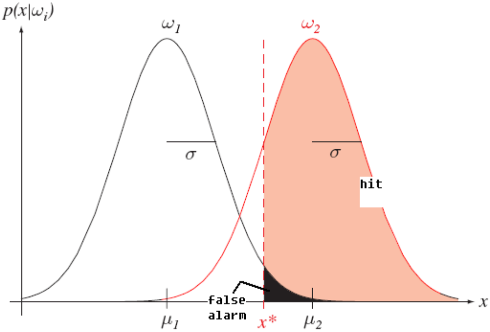
History

The ROC curve

- was first developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battle fields, also known as the signal detection theory
  - ω_1: object is not present (negative), e.g., detecting a weak pulse
  - ω_2: object is present (positive)
- was soon introduced in psychology to account for perceptual detection of signals
- has been widely used in medicine, radiology, and other areas for many decades

An Example

- A detector to detect whether there is an external signal (pulse) denoted by a signal x
- x is a random variable due to the random noise within and outside the detector itself
  - ω_1: when the signal is not present (negative), $p(x|\omega_1) = \mathcal{N}(\mu_1, \sigma^2)$
  - ω_2: when the signal is present (positive), $p(x|\omega_2) = \mathcal{N}(\mu_2, \sigma^2)$



discriminability:

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

Confusion matrix

		Predicted		Total
		ω_1	ω_2	
State of Nature	ω_1	correct rejection (true negative) $P(\mathbf{x} < \mathbf{x}^* \mid \mathbf{x} \in \omega_1)$	false alarm (false positive) $P(\mathbf{x} > \mathbf{x}^* \mid \mathbf{x} \in \omega_1)$	N
	ω_2	miss (false negative) $P(\mathbf{x} < \mathbf{x}^* \mid \mathbf{x} \in \omega_2)$	hit (true positive) $P(\mathbf{x} > \mathbf{x}^* \mid \mathbf{x} \in \omega_2)$	P

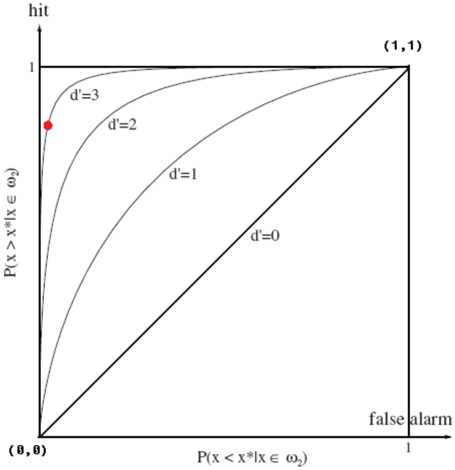
- true positive rate, TPR (hit rate, recall, sensitivity): $\frac{TP}{FN+TP}$, which determines a classifier or a diagnostic test performance on classifying positive instances correctly among all positive samples available during the test
- false positive rate, FPR (false alarm, $1 -$ sensitivity): $\frac{FP}{TN+FP}$, which defines how many incorrect positive results occur among all negative samples available during the test

ROC

- only TPR and FPR are needed to draw an ROC curve
- fixed density, i.e.,

ROC

- only TPR and FPR are needed to draw an ROC curve
- fixed density, i.e., d'
- changeable $\mathbf{x}^*$



Example

- four prediction results from 100 positive and 100 negative instances
- C' is the mirror of C across the center point $(0.5, 0.5)$

A

TP=63	FP=28	91
FN=37	TN=72	109
100	100	200

TPR = 0.63
FPR = 0.28
ACC = 0.68

B

TP=77	FP=77	154
FN=23	TN=23	46
100	100	200

TPR = 0.77
FPR = 0.77
ACC = 0.50

C

TP=24	FP=88	112
FN=76	TN=12	88
100	100	200

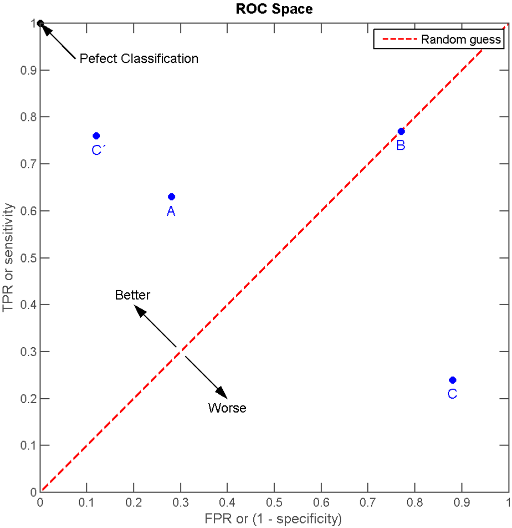
TPR = 0.24
FPR = 0.88
ACC = 0.18

C'

TP=76	FP=12	88
FN=24	TN=88	112
100	100	200

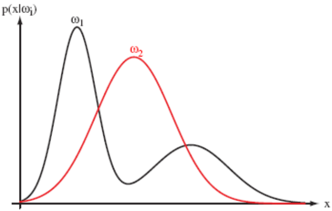
TPR = 0.76
FPR = 0.12
ACC = 0.82

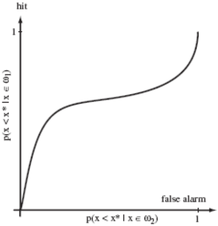
Example


ROC Space
Random guess
Pefect Classification
C´
A
B
C
Better
Worse
TPR or sensitivity
FPR or (1 - specificity)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Extension

to non-Gaussian assumption





- ROC analysis provides tools to select possibly optimal models