EEL 851: Biometrics

An Overview of

Statistical Pattern Recognition

Outline

- Introduction
  - Pattern
  - Feature
  - Noise
- Example
- Problem Analysis
  - Segmentation
  - Feature Extraction
  - Classification
- Design Cycle
- Bayesian Decision Theory
  - Discriminant Functions
  - Decision Regions
  - Minimum Distance Classification
- Nearest Neighbor Classification
- Parameter Estimation
- Decision Boundaries

Introduction

- What is a Pattern?
  - Object process of event consisting of both deterministic/stochastic components
  - Record of dynamic occurrences influenced by both deterministic and stochastic factors
  - Examples – voice, image, characters

- Kind of Patterns
  - **Visual patterns**
  - **Temporal patterns**
  - **Logical patterns**

- What is Pattern Class?
  - **Set of patterns sharing set of common attributes (or features)**
  - **Usually originating from the same source**

Introduction

- Feature
  - Relevant characteristics that make patterns apart from each other
  - Data extractable through measurements or processing

- Examples
  - **Patterns**
    - **Speech waveforms, crystals, textures, weather patterns**
  - **Features**
    - **Age, color, height, width**

- Classifications
  - **Assigning patterns into classes based on features**

- Noise
  - **Distortions associated with pattern processing and/or training samples that effect the classification performance of system**

Example

- Automation in fish packing plant
  - *Sort incoming fish on a conveyor according to species using optical sensing*

- Sorting species
  - Sea bass
  - Salmon

Problem Analysis

- Setup camera
  - Take some sample images
  - Features?

- Suggested features
  - Length
  - Lightness
  - Width
  - Number and shape of fins
  - Position of mouth, etc..

- Explore above suggested features!

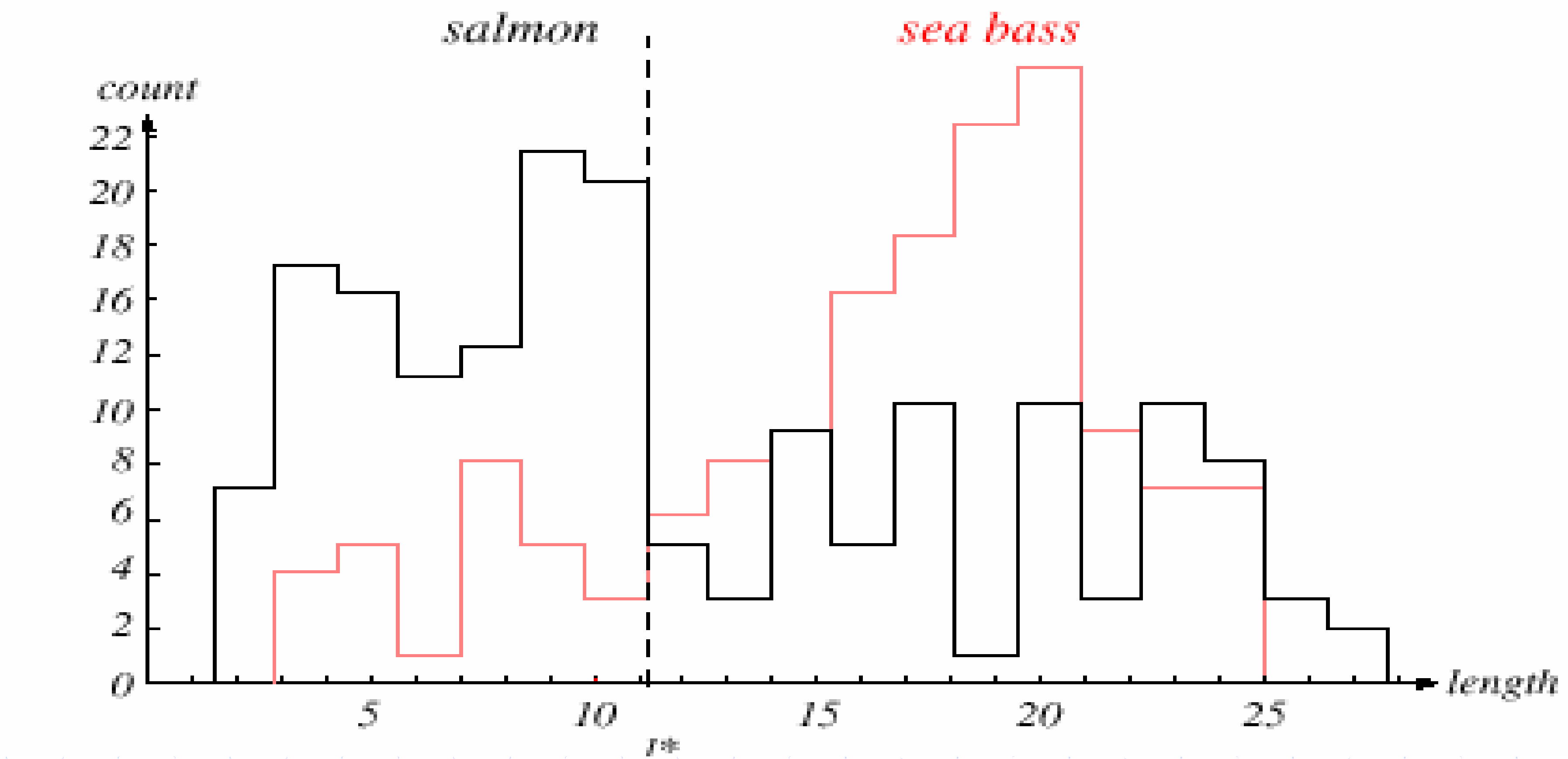
Preprocessing

- Segmentation
  - Isolate fishes from background
  - Isolate fishes from one another
- Feature extraction
  - Reduce the data by measuring certain features
- Classifier
  - Evaluates the evidence presented $\rightarrow$ Makes final decision

Preprocessing

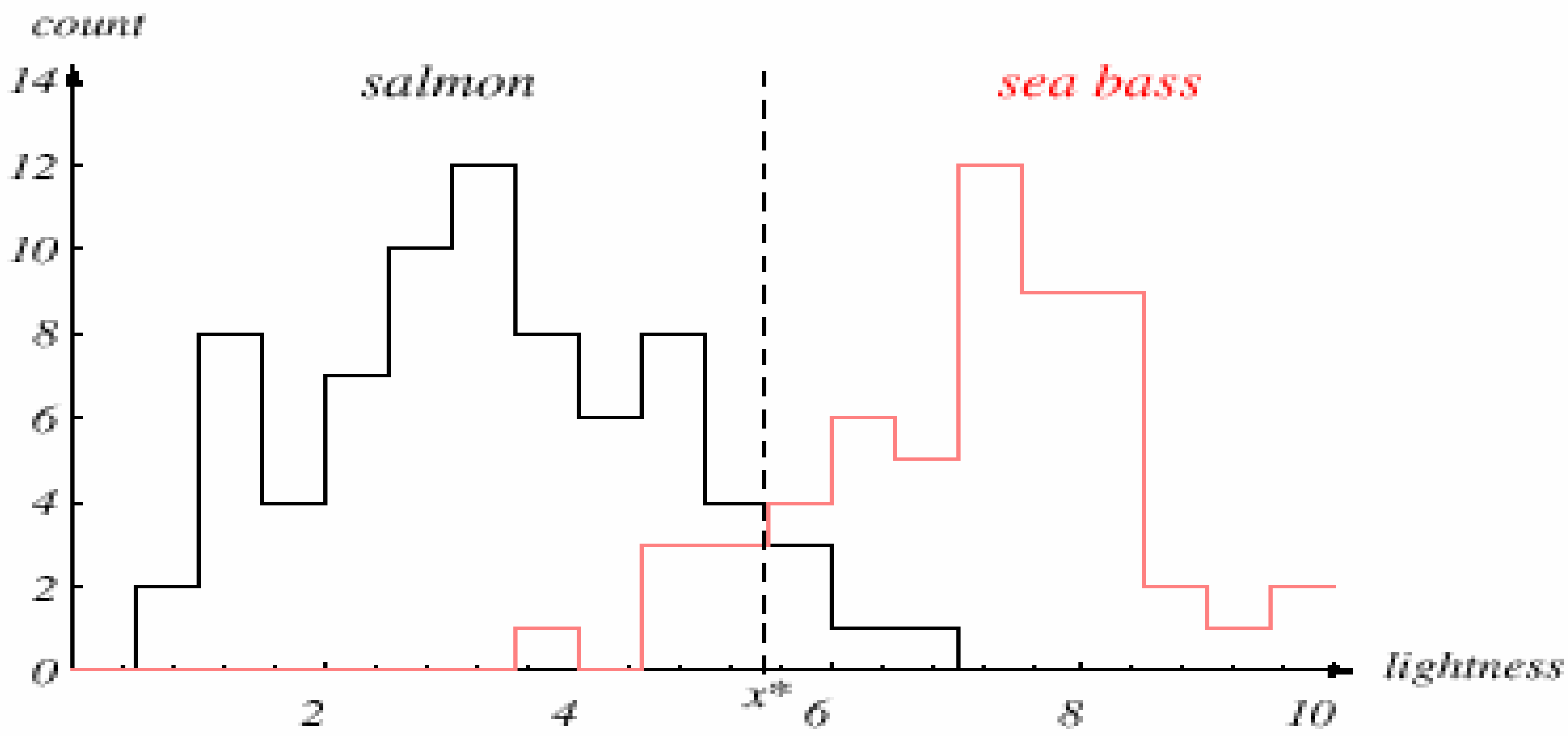
Classification

- Selecting length of fish → Possible feature for classification



Classification

- Length is a poor feature!
- Select brightness as a possible feature



Threshold Decision Boundary

- Cost
  - Consequences of our decision
  - Equal?

- Task of decision theory
  - Move the decision boundary towards smaller values of lightness, Cost ↓
  - Reduce number of sea brass classified as salmon

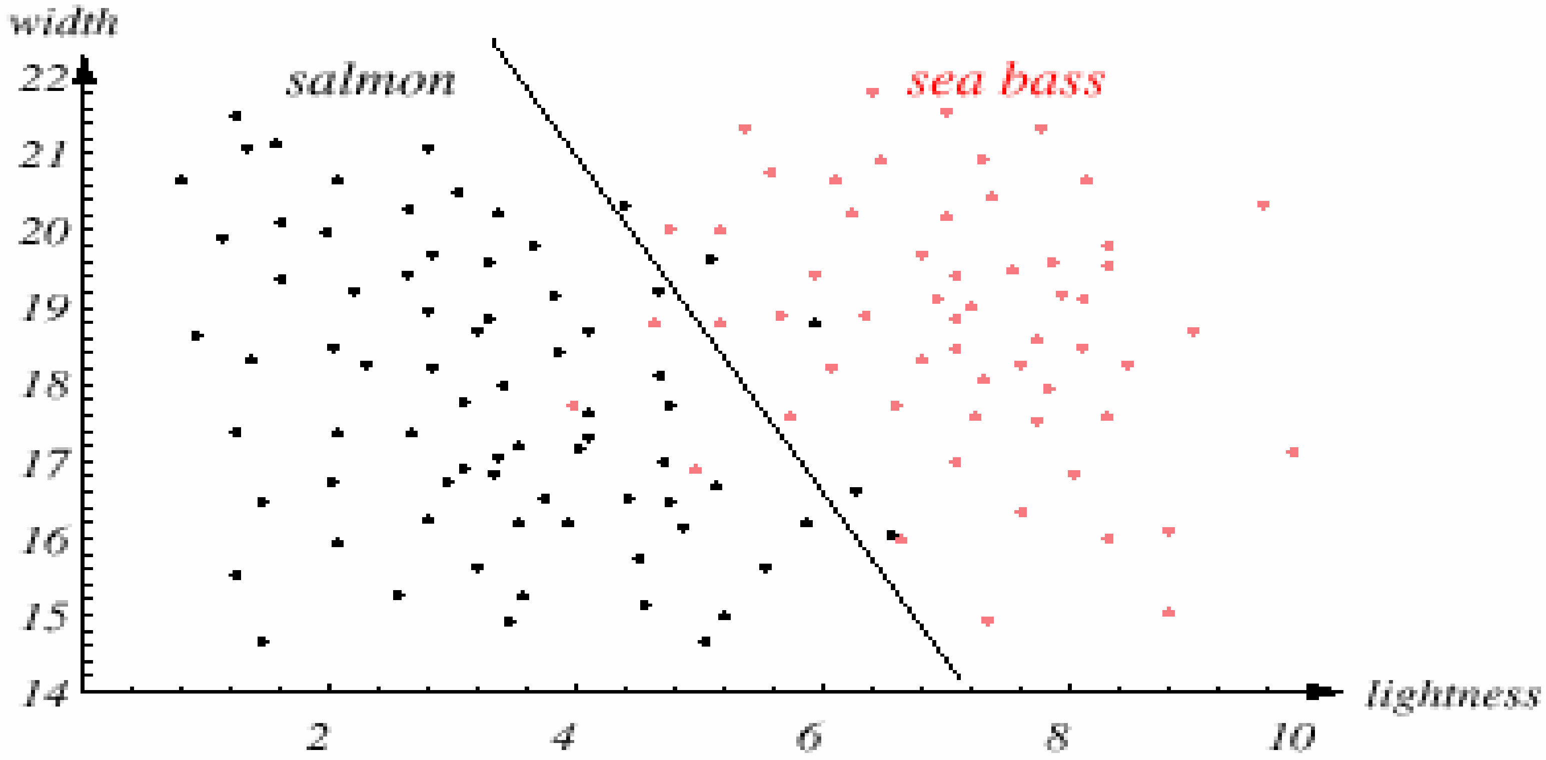
Classification

- Adopt *lightness* and the *width* of fish

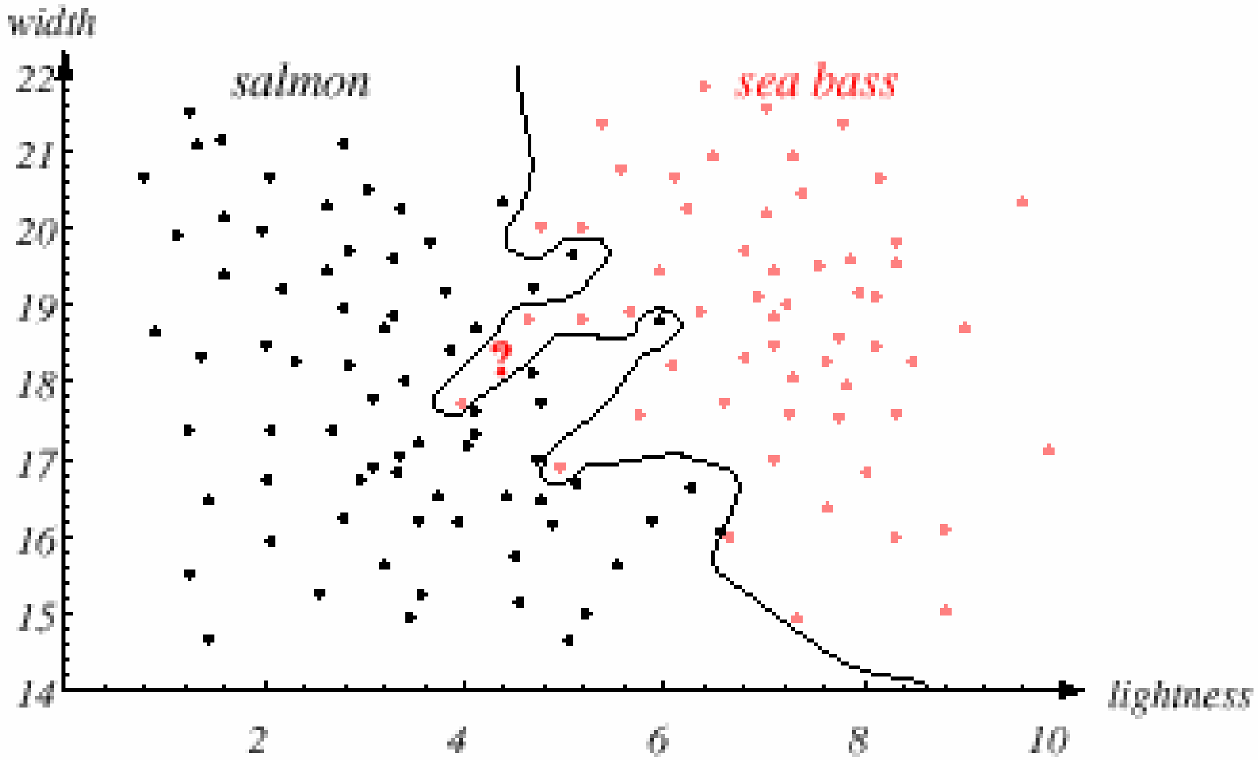
Fish $\Longrightarrow$ $x = [x_1, x_2]$

Lightness **Width**

Classification

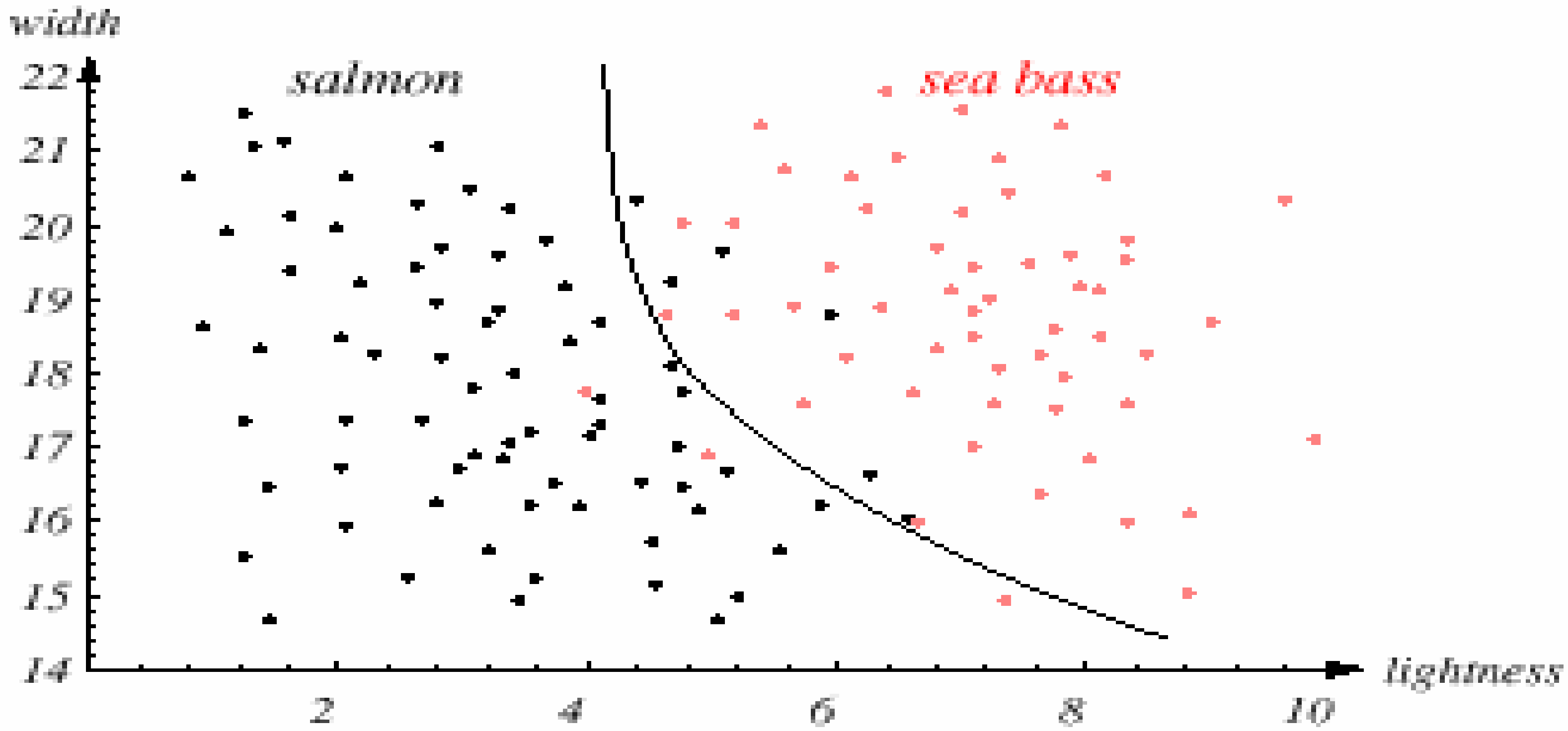


Classification



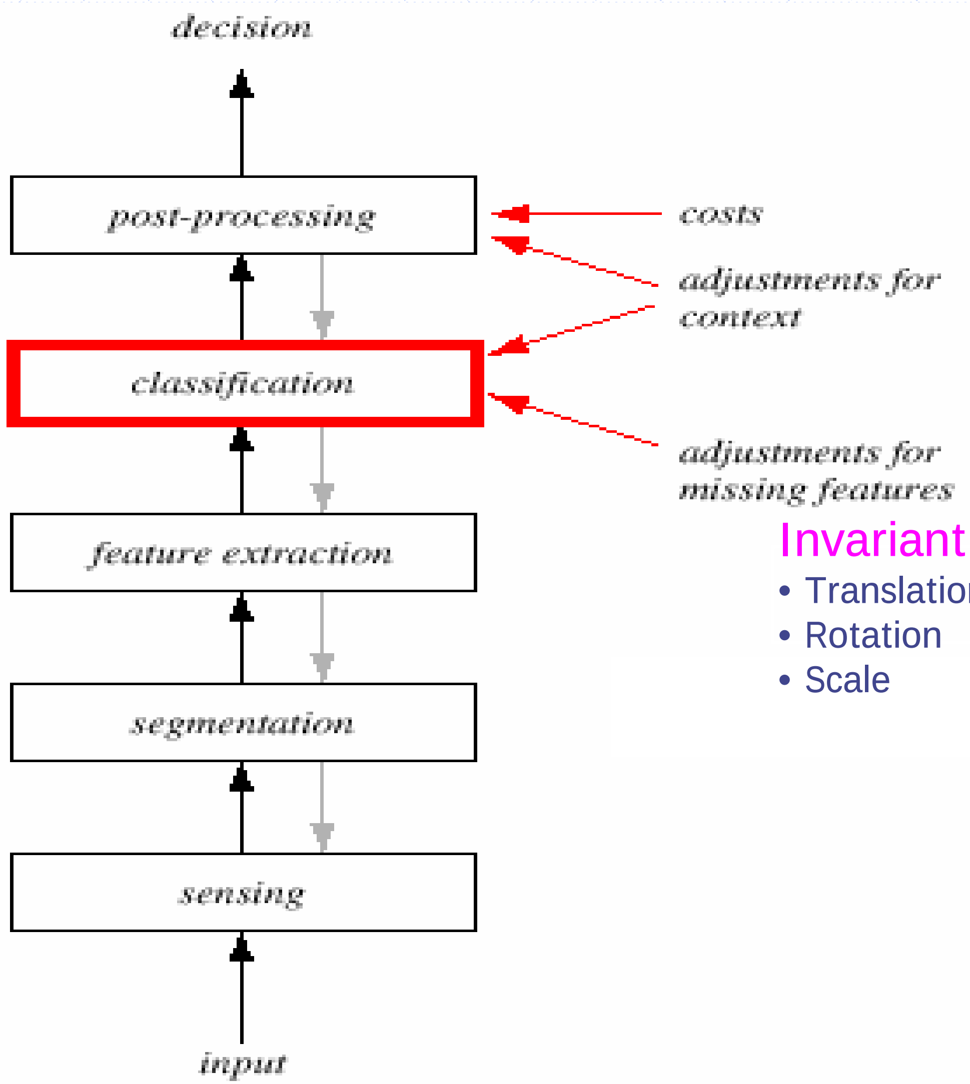
- Our satisfaction → Premature
  - Central aim → Correctly classify a novel input
  - Issue of Generalization!

Classification



- Syntactic Pattern Recognition
  - Recursive description using method of formal languages
- Structural Pattern Recognition
  - Derivation of descriptions using formal rules

Pattern Recognition Systems



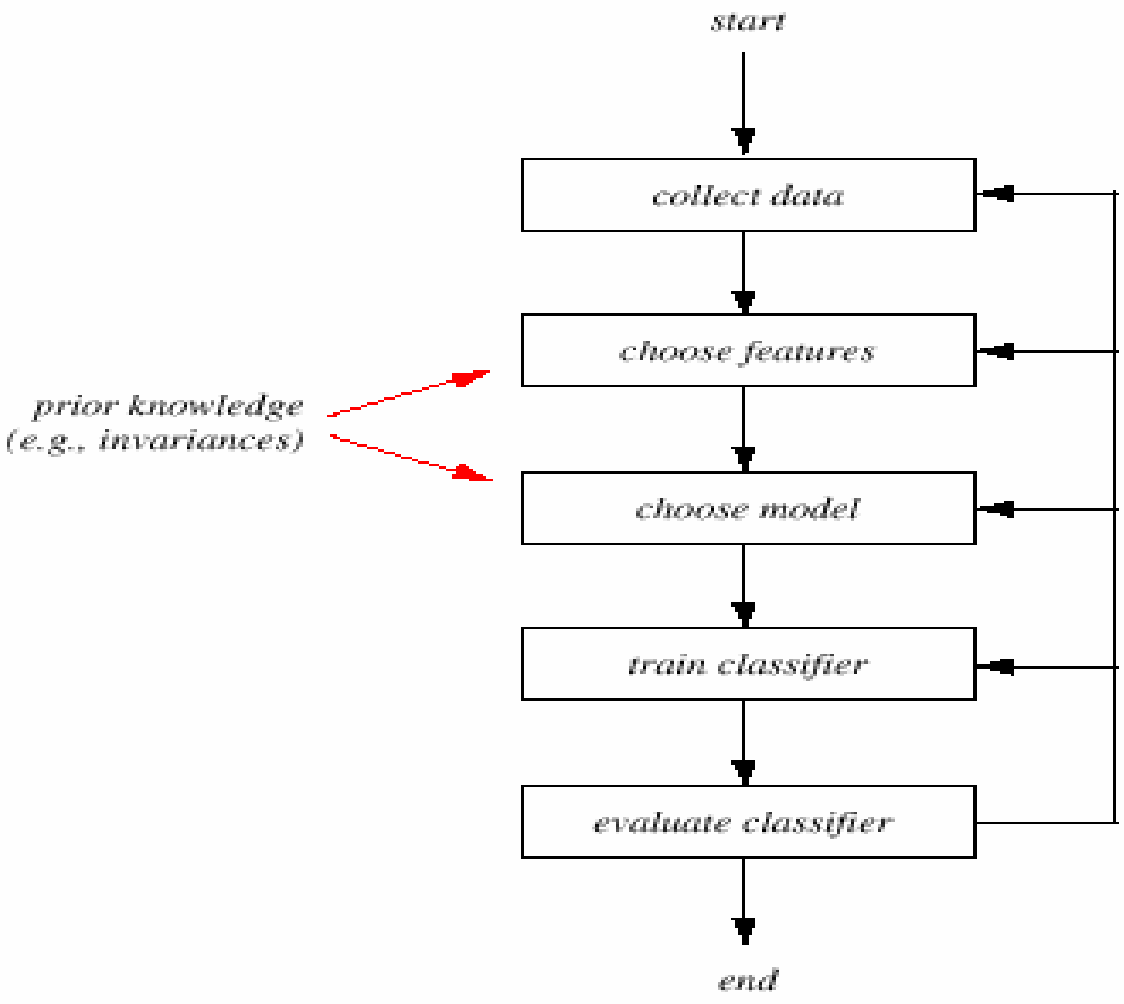
Invariant Features

- Translation
- Rotation
- Scale

Design Cycle

- Data Collection
- Feature Choice
- Model Choice
- Training
- Evaluation
- Computational Complexity

Design Cycle



Types of Learning

- Supervised Learning
  - **Teacher → Category label or cost for each label in training set**
  - **Desired input → Desired output**
    - Goals → Produce a correct output given a new input

- Goals of Supervised Learning
  - **Classification**
    - Desired output → Discrete class labels
  - **Regression**
    - Desired output → Continuous valued

- Unsupervised Learning
  - The system forms ***clusters*** or natural ***groupings*** of input pattern
  - Goals → Build a model or find useful representations of data
  - Usage → Reasoning, finding clusters, dimensionality reduction, decision making, prediction, etc.
    - Data compression, Outlier detection, Help classification

Bayesian Decision Theory

- Assumptions
  - **Problem $\rightarrow$ Probabilistic terms**
  - **Knowledge of relevant probabilities**

- Sea Bass/Salmon example
  - **State of nature**
    - Random Variable
    - $\omega \rightarrow \omega_1$ (sea bass)
    - $\omega \rightarrow \omega_1$ (salmon)

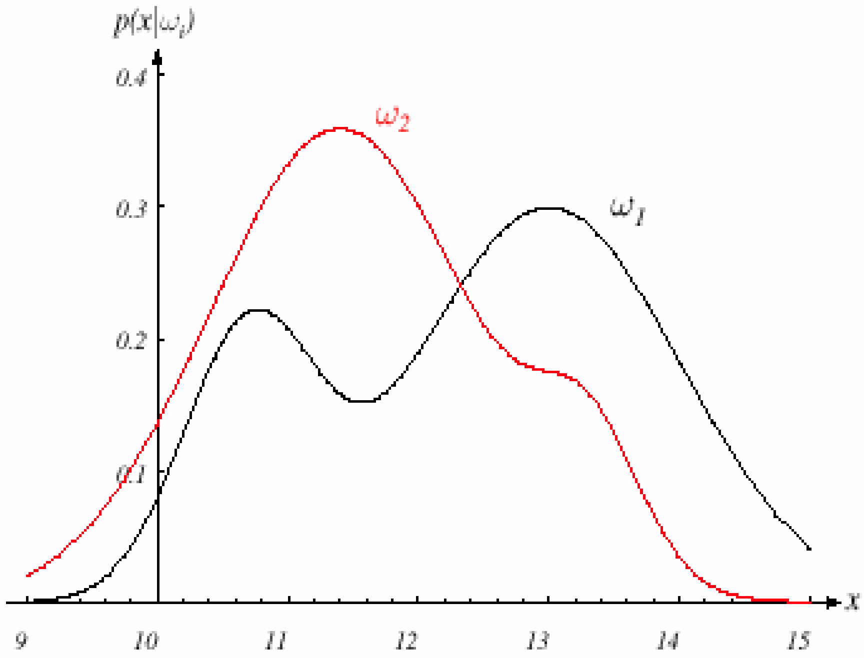
  - **Priori probability**
    - $P(\omega_1) \rightarrow$ sea bass; $P(\omega_2) \rightarrow$ salmon
    - $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustivity)

Bayesian Decision Theory

- Decision Rule
  - **Only prior information, Cannot see!**
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ otherwise decide ω_2
  - **More info in most cases**

- Class Conditional Info
  - **Class-conditional pdf** $\rightarrow$ $\boldsymbol{p}(x|\omega)$
  - $p(x|\omega_1)$ and $p(x|\omega_2)$
    - Difference in lightness between populations of sea and salmon

Bayesian Decision Theory



- Class-conditional pdf
  - $x \rightarrow$ Lightness of fish
  - Normalized

Bayes Formula

- Lightness of fish $\rightarrow$ x, Class?

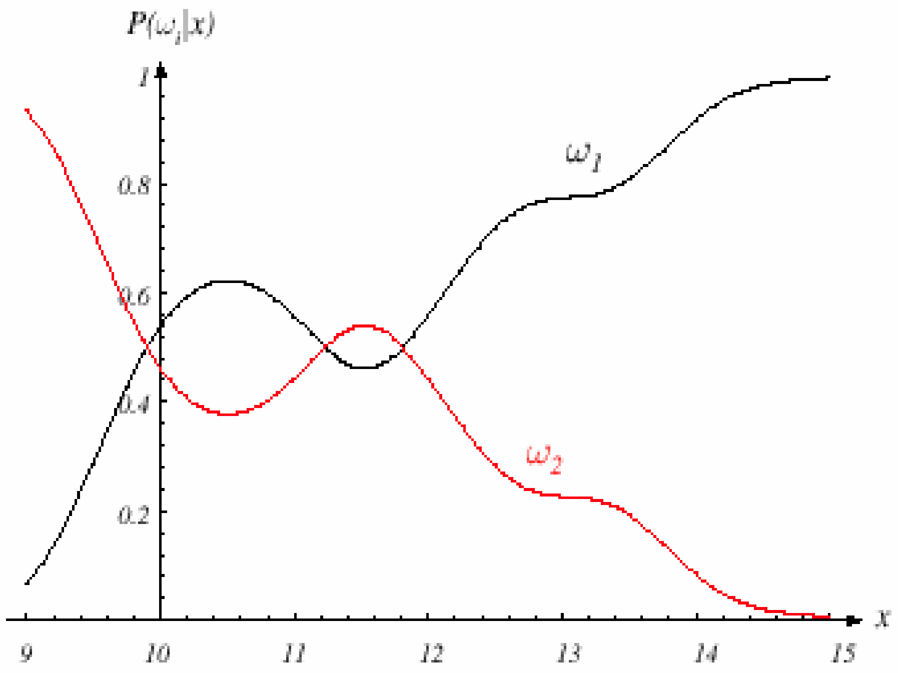
$$P(\omega_j \mid x) = p(x \mid \omega_j) \,.\, P(\omega_j) \,/\, p(x)$$

where in case of two categories

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

- Posterior = (Likelihood $\times$ Prior) / Evidence
  - $p(\omega_j \mid x) \rightarrow$ Likelihood
  - $p(x) \rightarrow$ Evidence (scale factor)

Posterior Probabilities



- Prior probabilities
  - $P(\omega_1) = 2/3$
  - $P(\omega_2) = 1/3$

Bayes Decision Rule

- Decision given posterior probabilities
  - **Observation $\rightarrow$ x**
  - if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ $\rightarrow$ True state of nature = ω_1
  - if $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ $\rightarrow$ True state of nature = ω_2

- Probability of error
  - $P(\text{error} \mid x) = P(\omega_1 \mid x)$ if we decide ω_2
  - $P(\text{error} \mid x) = P(\omega_2 \mid x)$ if we decide ω_1

Bayes Decision Rule

- Minimize probability of error
  - Decide ω_1 if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise decide ω_2
- Bayes decision
  - $P(\text{error} \mid x) = \min\,[P(\omega_1 \mid x), P(\omega_2 \mid x)]$

Bayesian Classifier

- Generalization of Preceding Ideas
  - **More than one feature**
  - **More than two state of nature**
  - **Actions not only deciding state of nature**
  - **Loss function → More general than probability of error**

- Bayes Decision Rule
  - *Input pattern* C *is classified into class* ω_k *, for a given feature vector* **x***, maximizes the* *<u>posterior</u>* *probability;*

    $P(\omega_k | \mathbf{x}) \geq P(\omega_j | \mathbf{x}) \quad \text{for } \forall\, j \neq k$

    - Likelihood conditions $p(\mathbf{x} | \omega_k)$ → Training data measurements
    - Prior probabilities $P(\omega_k)$ → Supposed to known within given population of sample patterns
    - $p(\mathbf{x})$ is same for all class alternatives → Ignored, normalization factor

Loss Function

- Certain classification errors may be costly than others
  - False Negative may be much more costly than False Positive
  - Examples $\rightarrow$ Fire alarm, Medical diagnosis
  - $\Omega = \{\omega_1, \omega_1, \ldots \omega_n\}$ $\rightarrow$ Possible set of nature/classes
  - $\Delta = \{\alpha_1, \alpha_2, \ldots \alpha_n\}$ $\rightarrow$ Possible decisions/actions

- Loss Function
  - $\lambda(\alpha_i|\omega_j)$
  - Loss incurred by taking action α_i when true state of nature is ω_j

- Conditional Risk
  - $R(\alpha_i|\mathbf{x})$
  - Loss expected by taking action α_i when observed evidence is $\mathbf{x}$

Minimum Error Rate Classification

- Action α_i is associated with class ω_j
- All errors are equally likely
- Zero-One Loss
  - Classification decision α_i is correct <u>only</u> if state of nature is ω_j
  - Symmetrical function

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

- Conditional Risk
  - $R(\alpha_i|\mathbf{x}) = \sum \lambda(\alpha_i|\omega_j)\, P(\omega_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$

- Minimize Risk
  - Select the class <u>maximizing</u> the posterior probability
  - Suggested by Bayes Decision Rule, Minimum error rate classification

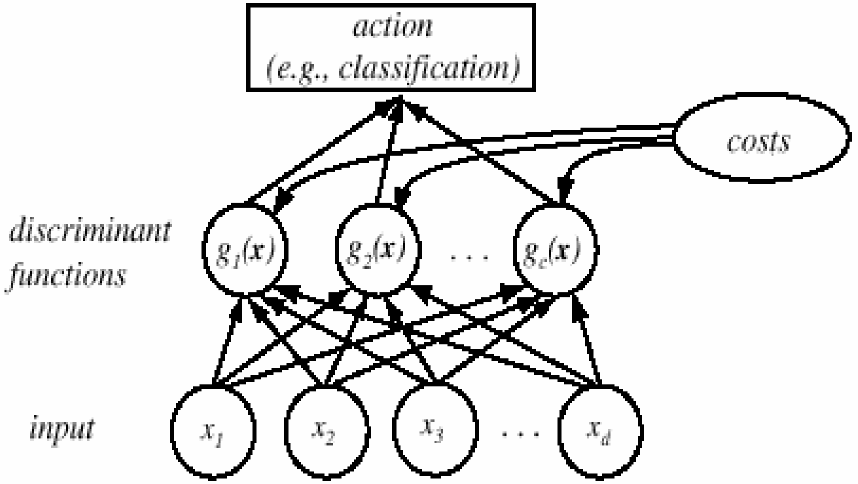
Neyman-Pearson Criterion

- Minimize overall risk subject to constraints
  - $\int R(\alpha_i|\mathbf{x})\, d\mathbf{x} < \textit{constant}$
  - Misclassification limited to a frequency
  - Fish example $\rightarrow$ Do not misclassify more than 1% of salmon as bass
  - Minimize the chance of classifying sea bass as salmon

Discriminant Functions

- Classifier Representation
  - Function employed for discriminating among classes
  - $g_i(\mathbf{x})$, $i = 1, \ldots, k$
  - Classifier $\rightarrow$ Assign $\mathbf{x}$ to class ω_i *if*

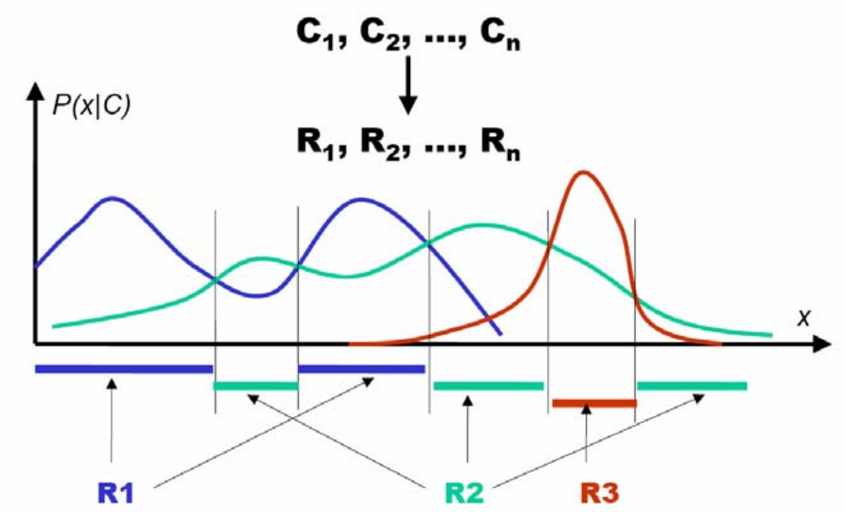
    $g_i(\mathbf{x}) \geq g_j(\mathbf{x})$ for $\forall\, j \neq i$

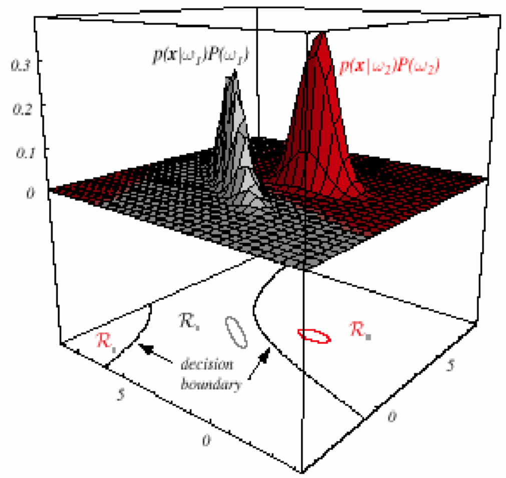


- Possible Choices
  - $g_i(\mathbf{x}) = P(\omega_j|\mathbf{x})$
  - $g_i(\mathbf{x}) = p(\mathbf{x}|\omega_j)P(\omega_j)$
  - $g_i(\mathbf{x}) = \ln\{p(\mathbf{x}|\omega_j)\} + \ln\{P(\omega_j)\}$

Decision Region

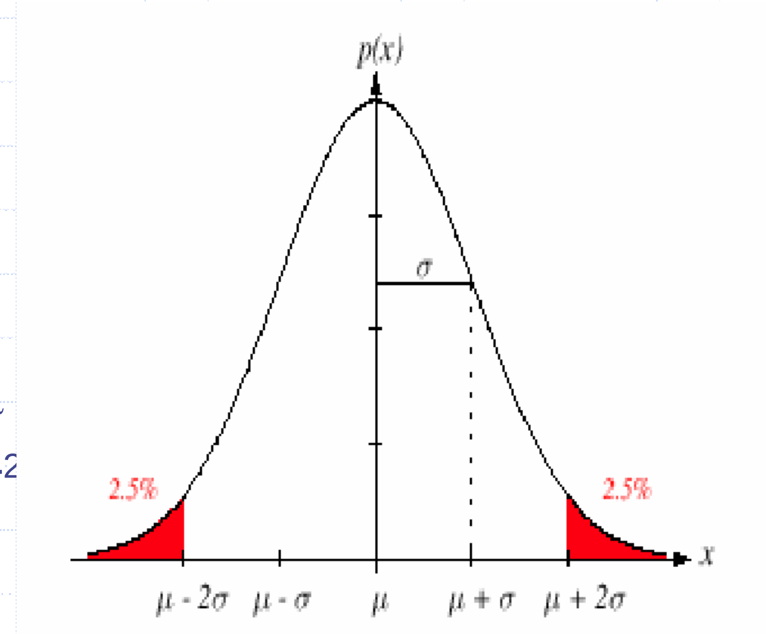
- Effect of Decision Rule
  - Feature space → Decision regions
  - Decision Boundaries
    - Surface in feature space →Ties occur among largest discriminant functions





Normal Density

- Univariate Density
  - Analytically tractable, Maximum entropy
  - Continuous-valued density
  - A lot of processes are asymptotically Gaussian
  - Central Limit Theorem
    - Aggregate effect of independent random disturbances $\rightarrow$ Gaussian
    - Many patterns $\rightarrow$ Prototype corrupted by large number of RP



μ

σ^2 r variance

Normal Density

- Multivariate density

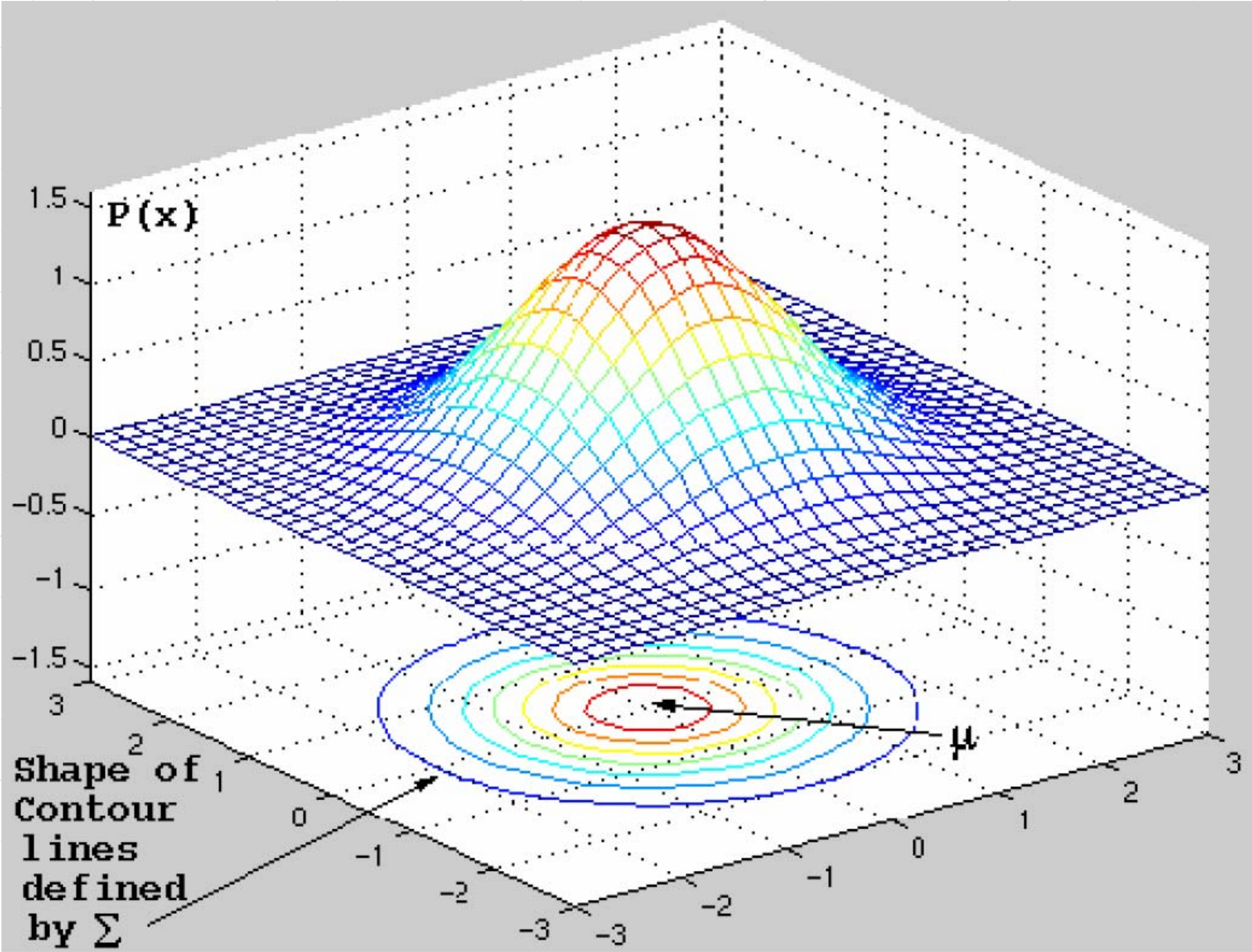
  - Multivariate normal density in d dimensions is:

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

where:
$x = (x_1, x_2, \ldots, x_d)^t$ feature vector
$\mu = (\mu_1, \mu_2, \ldots, \mu_d)^t$ mean vector
Σ = d*d covariance matrix
$|\Sigma|$ and Σ^{-1} are determinant and inverse respectively

  - $\Sigma \rightarrow$ Shape of Gaussian curve
  - $(x - \mu)^t \Sigma^{-1}(x - \mu) \rightarrow$ Mahalanobis distance

Normal Density



Discriminant Functions

- Minimum Error Rate Classification
  - $g_i(x) = \ln p(x \mid \omega_i) + \ln P(\omega_i)$

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

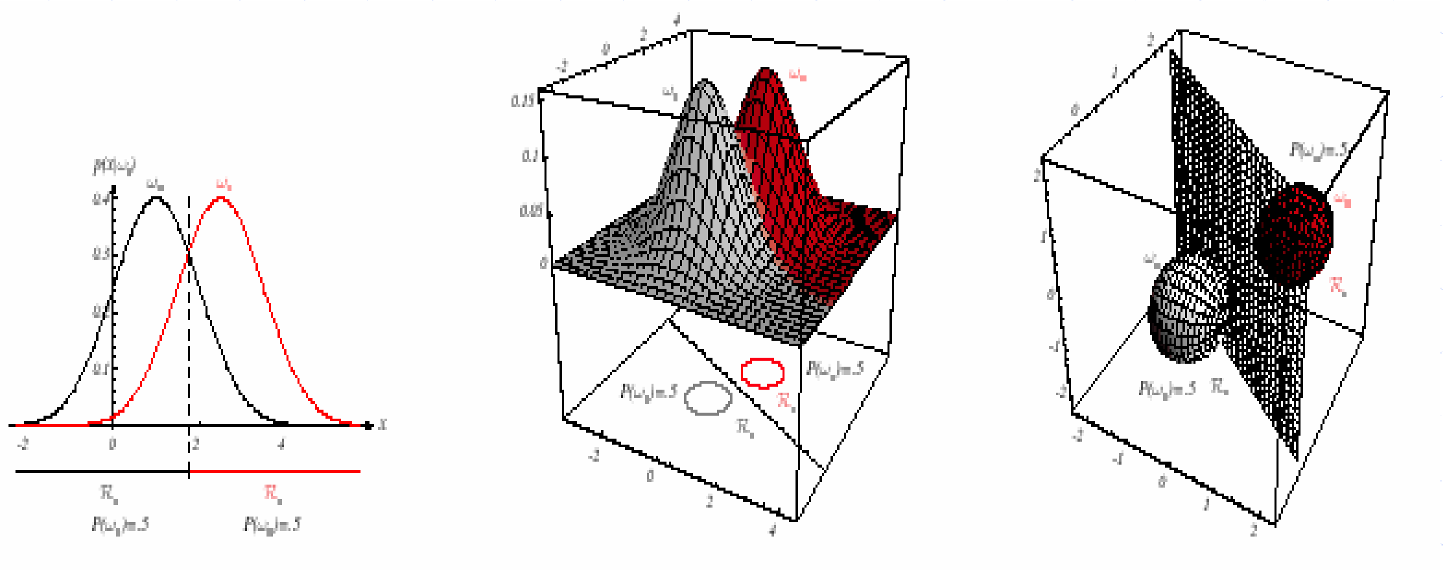
- Case $\rightarrow \Sigma_i = \sigma^2.I$
  - Features are statistically independent, same variance σ^2
  - Diagonal covariance matrix
  - $$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0}$$

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

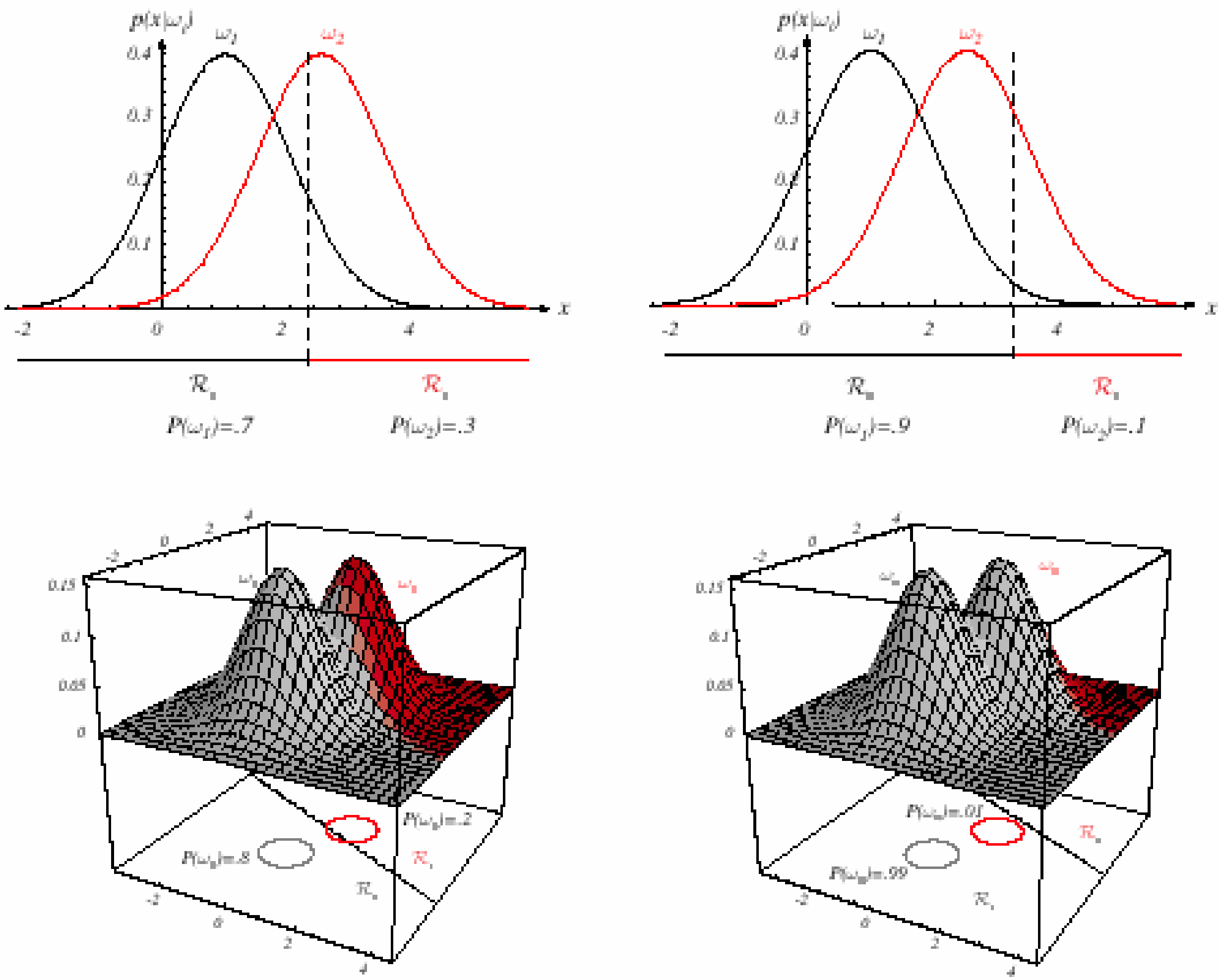
  - Linear machines
  - Decision surface $\rightarrow$ Pieces of hyperplanes

    $g_i(x) = g_j(x)$

Discriminant Functions



Discriminant Functions



Discriminant Functions

- Case $\rightarrow \Sigma_i = \Sigma$
  - Covariance matrices of all classes $\rightarrow$ Identical but arbitrary

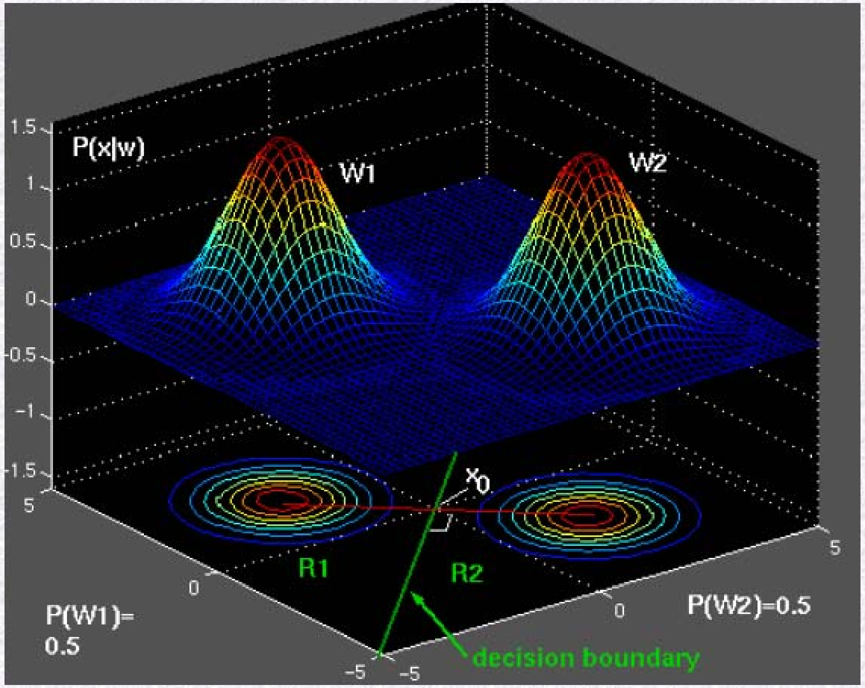
$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^\mathbf{t}\Sigma^{-1}(\mathbf{x}-\mu_\mathbf{i}) + \ln P(\omega_i)$$

  - Hyperplane separating R_i and R_j

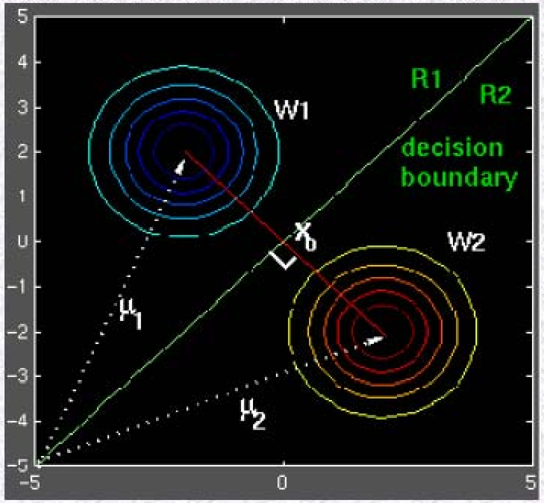
$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\mu_i - \mu_j)^t \Sigma^{-1}(\mu_i - \mu_j)} . (\mu_i - \mu_j)$$

  - Hyperplane separating R_i and $R_j \rightarrow$ Not orthogonal to the line between the means

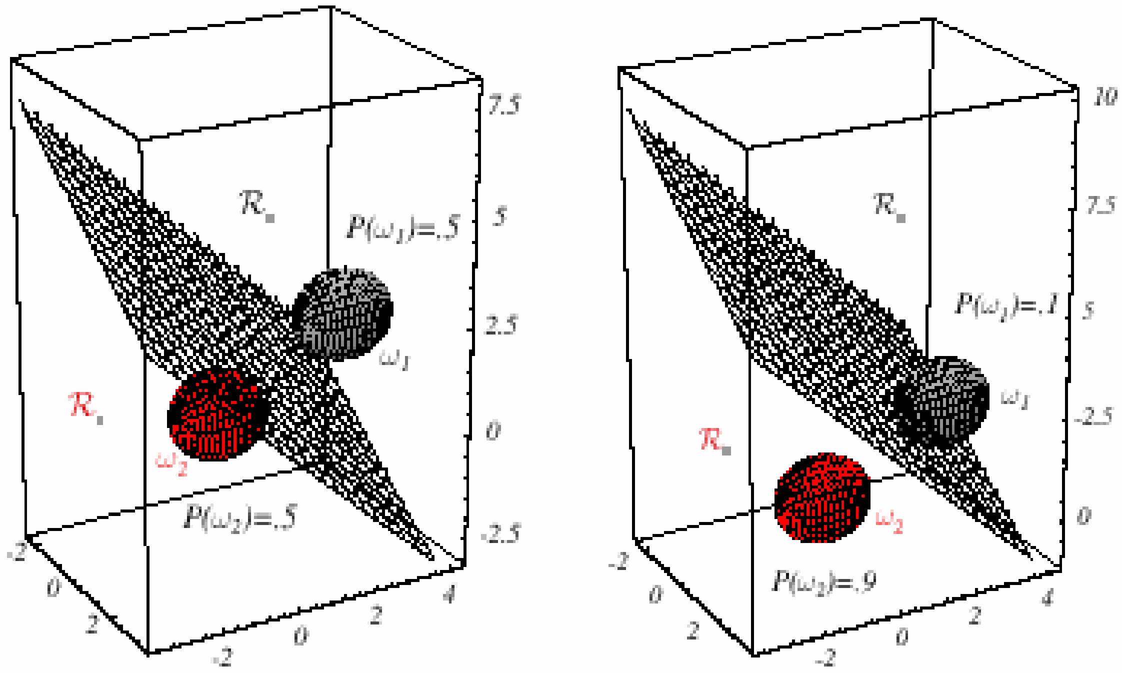
Discriminant Functions



Projected contour lines from the above 3D graph.



Discriminant Functions



Discriminant Functions

- Case $\rightarrow$ Σ_i = arbitrary

  - The covariance matrices are different for each category

$$g_i(x) = x^t W_i x + w_i^t x = w_{i0}$$

$$where:$$

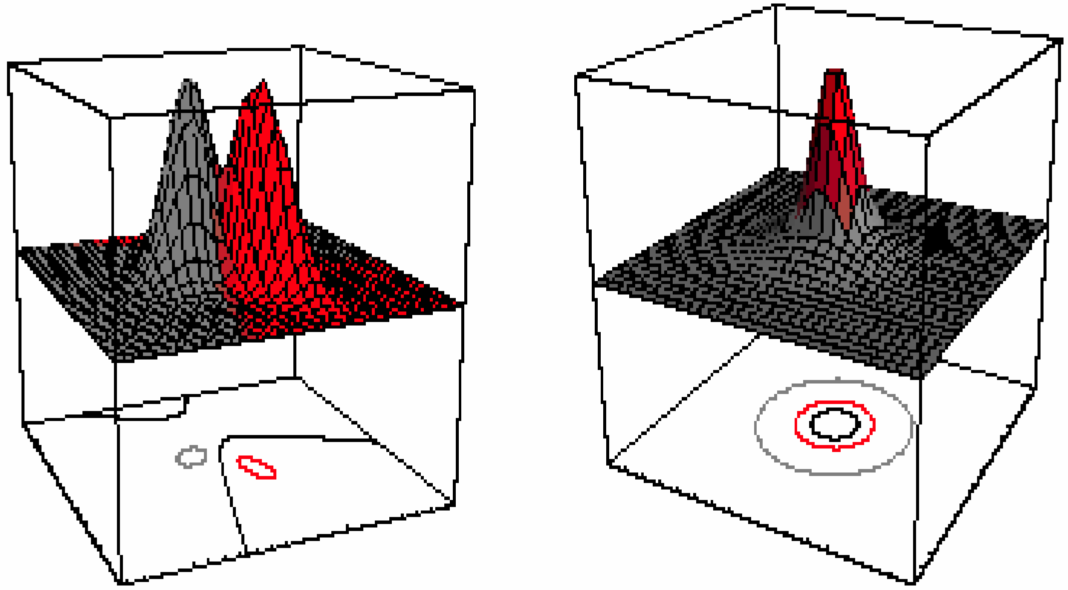
$$\mathrm{W_i} = -\frac{1}{2}\Sigma_i^{-1}$$

$$\mathrm{w_i} = \Sigma_i^{-1}\mu_i$$

$$\mathrm{w}_{i0} = -\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

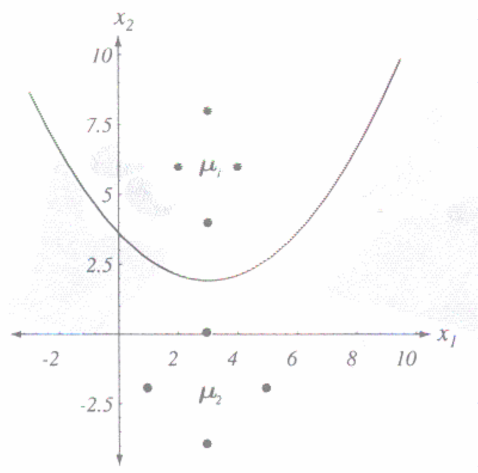
  - Hyperquadrics
    - Hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, hyperhyperboloids

Discriminant Functions



Decision Boundary - Example

- Compute Bayes decision boundary
  - Gaussian Distributions



- Means and Covariance's $\rightarrow$ Discrete versions
  - $\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix} \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix}$
  - $\mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix} \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$
  - $\mathbf{x}_2 = 3.514 - 1.125\mathbf{x}_1 + 0.1875\mathbf{x}_1^2$

Supervised Learning

- Parametric Approaches
  - Bayesian parameter estimation
  - Maximum likelihood estimation
- Estimation Problem
  - Estimate $P(\omega_j)$
  - Estimate $p(\omega_j \mid x) \rightarrow$ Tough
    - High dimensional feature spaces, small number of training samples
- Simplifying Assumptions
  - Feature Independence
  - Independently drawn samples $\rightarrow$ **I. I. D.** model
  - Assume that $p(\omega_j \mid x)$ is Gaussian
    - Estimation problem $\rightarrow$ Parameters of normality

Estimation Methods

- Bayesian Estimation (MAP)
  - Distribution parameters are random values that follow a known (*i.e.* Gaussian) distribution
  - Behavior of training data helps in revising parameter values
  - Large training samples $\rightarrow$ Better chances of refining posterior probabilities (parameter peaking)

- Maximum-Likelihood (ML) Estimation
  - Parameters of probabilistic distributions are fixed but unknown values
  - Parameters $\rightarrow$ Unknown constants, Identify using training data
  - Best estimates of parameter values
    - Class-conditional probabilities are maximized over the available samples

Parametric Approaches

- Curse of dimensionality

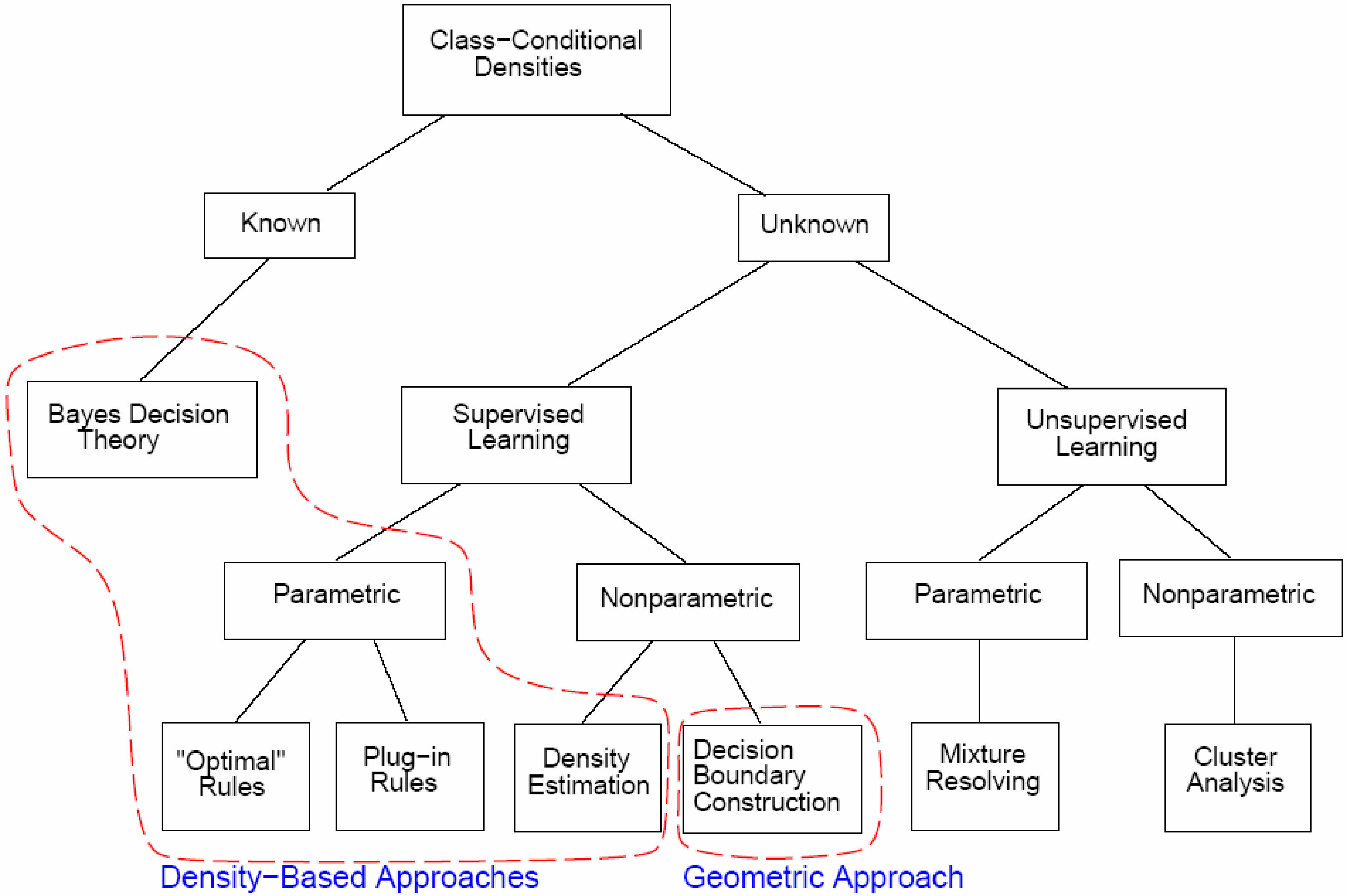
- Estimate the parameters known distribution
  - Smaller number of samples
  - Some priori information

Non-parametric Approaches

- Parzen window pdf estimation (KDE)
  - Estimate $p(\omega_j \mid x)$ directly from sample patterns

- K_n nearest-neighbor
  - Directly construct the decision boundary based on training data

Statistical Approaches



Nearest-Neighbor Methods

- Statistical, nearest-neighbor or memory-based methods
- *k*-nearest-neighbor
  - New pattern category $\rightarrow$ Plurality of its *k* closest neighbor
  - Large *k* $\rightarrow$ Decreases the chance of undue influence by noisy training pattern
  - Small *k* $\rightarrow$ Reduces acuity of method
  - Distance metric usually Euclidean

$$\sqrt{\sum_{j=1}^{n}(x_{1j}-x_{2j})^2}$$

  - In practice features are scaled $\rightarrow a_j$

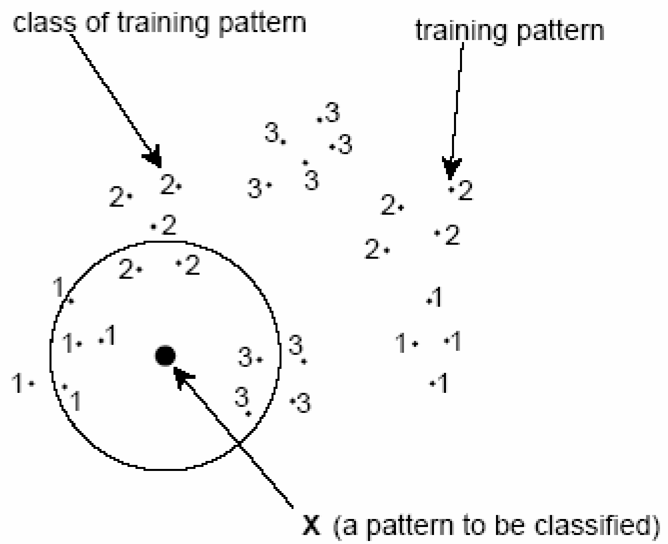
$$\sqrt{\sum_{j=1}^{n}a_j^2(x_{1j}-x_{2j})^2}$$

  - Advantage $\rightarrow$ Does not require training

Nearest-Neighbor Decision

➢ Example

- Class of training patterns → 4



k = 8

four patterns of category 1
two patterns of category 2
two patterns of category 3

plurality are in category 1, so
decide **X** is in category 1

References

- Richard O. Duda, Peter E. Hart, and David G.Stork, *Pattern Classification*, 2nd edition,John Wiley & Sons, 2001 (most contents)
- Nils J. Nilsson, *Introduction to Machine Learning*, http://ai.stanford.edu/people/nilsson/mlbook.html
- http://www.rii.ricoh.com/%7Estork/DHS.html
- A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Trans PAMI*, pp. 4-37, Jan. 2000

Other Useful References

- Fukunaga, *Introduction to Statistical Pattern Recognition,* Academic Press, New York, NY, 1972.
- Therrien, *Decision Estimation and Classification,* John-Wiley & sons, New York, NY, 1989.
- Hertz, Krogh, and Palmer, *Introduction to the Theory of Neural Computation,* Addison-Wesley, Reading, MA, 1991.
- Bose and Liang, *Neural Network Fundamentals with Graphs, Algorithms, and Applications,* McGrawll-Hill, New York, NY, 1996