Statistical Methods in Artificial Intelligence
CSE471 - Monsoon 2016 : Lecture 09

Avinash Sharma
CVIT, IIIT Hyderabad

Lecture Plan

- Introduction
  - Probability Theory Revision
  - Toy Example Walkthrough
- Bayesian Decision Theory
  - Bayes Formula
    - Prior, Likelihood, Posteriori and Evidence
  - Bayes Decision Rule
  - Bayes Risk
- Minimum-Error-Rate Classification

Introduction

- Random Variables
  - Boolean (True/False values)
  - Discrete (Categorical/Exact values like weather, Birth Year)
  - Continuous (Continuous values like Temperature, Time, Weight)
- Fundamental Rules
  - Probability of Union: $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
  - Joint Probabilities: $P(A, B) = P(A|B)P(B)$
  - Conditional Probability: $P(A|B)$=$P(A, B)/P(B)$ if $P(B) > 0$
  - Marginal Distribution: $P(A) = \sum_b P(A, B) = \sum_b P(A|B = b)P(B = b)$
- Probability Density Functions (PDFs)
  - Discrete - histograms
  - Continuous $p(x)$ where $P(a \leq x \leq b) = \int_a^b p(x)dx$ & $\int_{-\infty}^{+\infty} p(x)dx = 1$

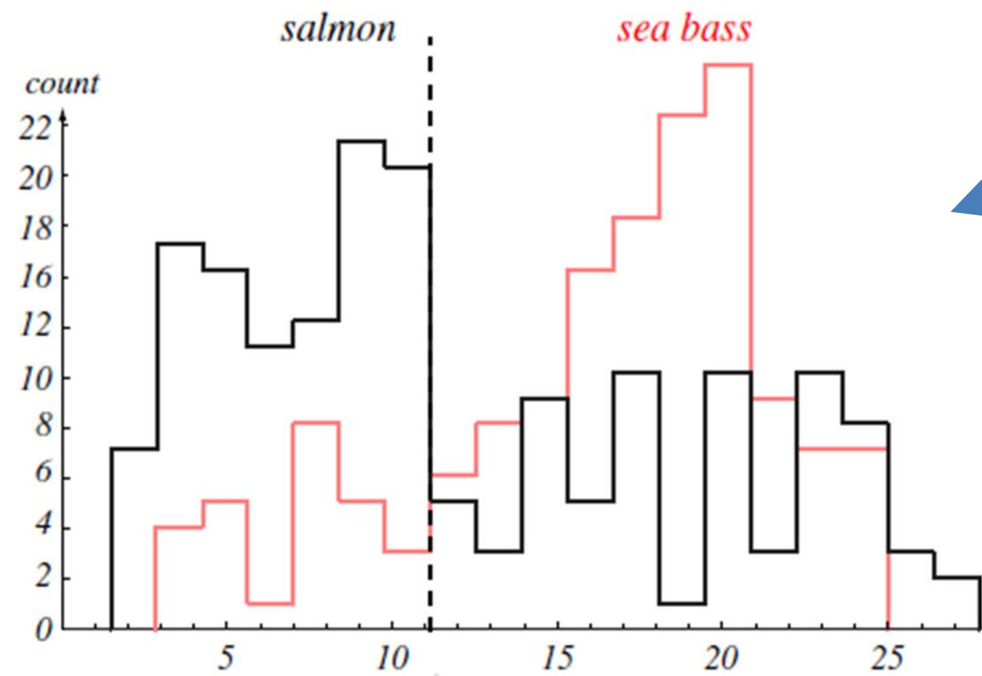
Toy Examples Walkthrough

- Image Based Fish Classification (Salmon v/s Sea Bass)

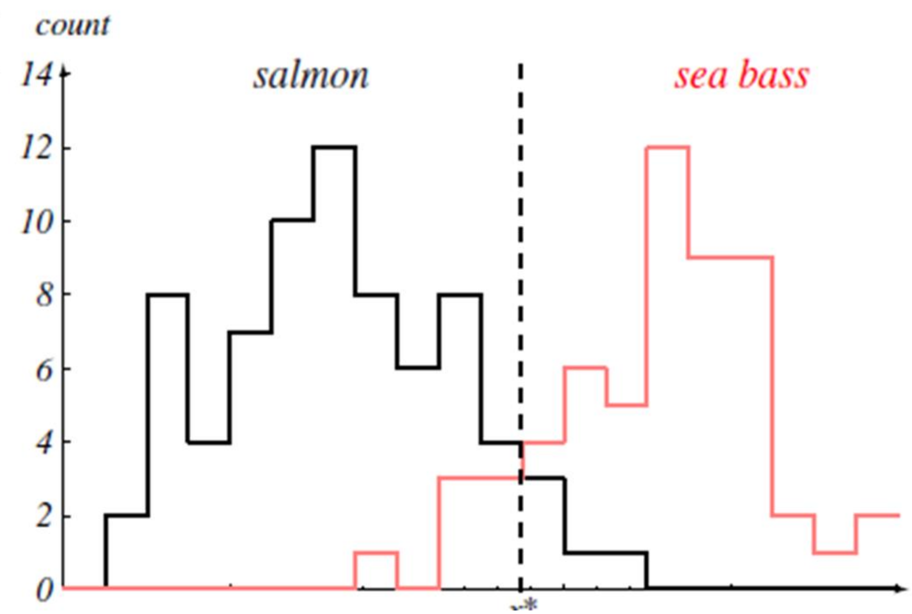
Toy Examples Walkthrough

- State of nature (ω)
  - Let variable ω be the discrete random variable which can assume only two categorical values, namely, ω_1(i.e., sea bass) or ω_2(i.e., solmon)
- Prior knowledge based classification
  - $P(\omega = \omega_2)$ represents the prior probability of any new sample belonging to class ω_2.
  - Let $P(\omega_1)$ and $P(\omega_2)$ be the class prior probabilities of the next fish on conveyer belt being sea bass and solmon, respectively. .
  - $P(\omega_1)$+ $P(\omega_2) = 1$ i.e., only if two types of fishes are caught
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ else ω_2. (**Decision Rule**)
- But we can use more information !!

Toy Examples Walkthrough

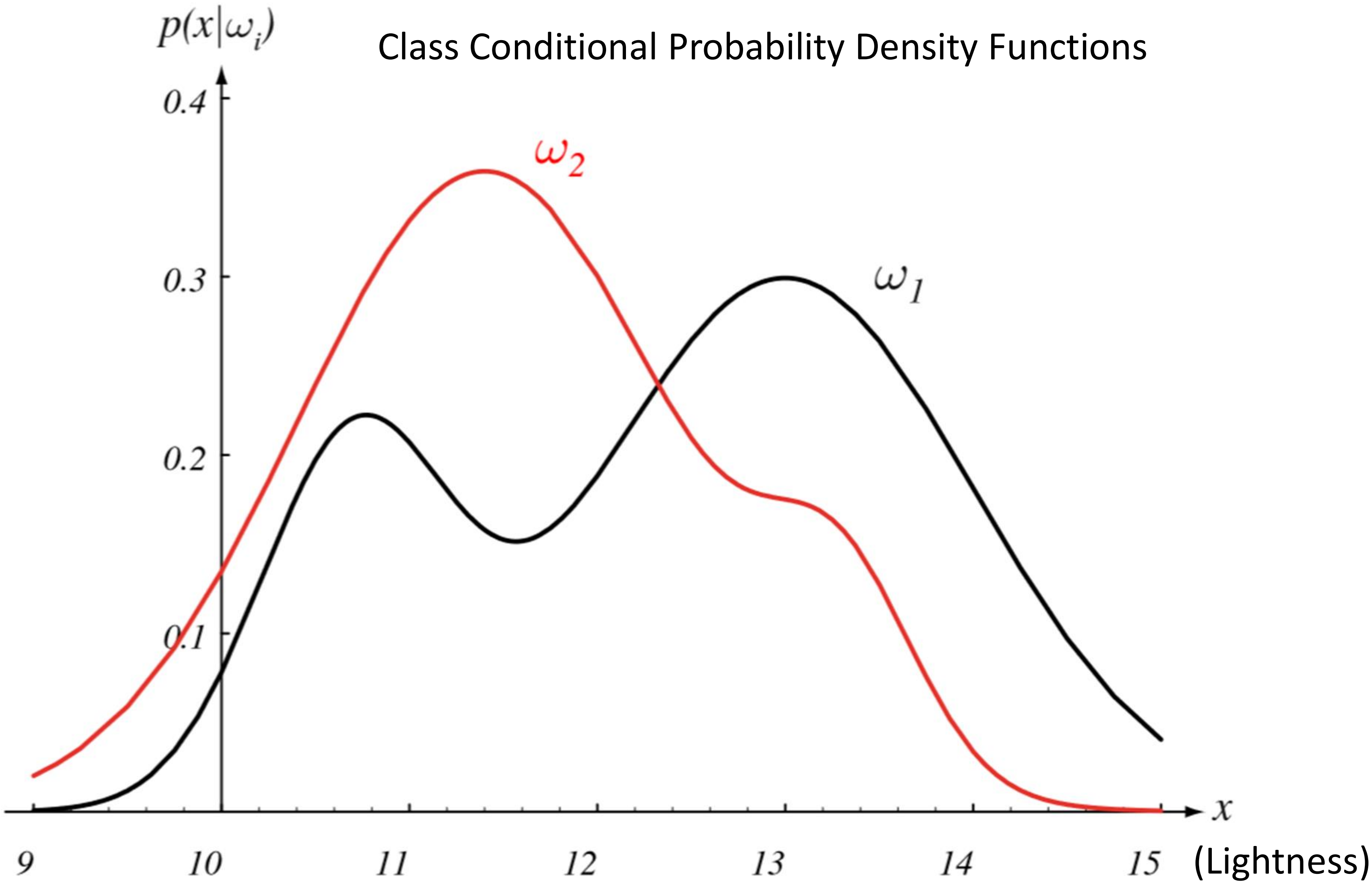


Length Histogram Feature



Lightness Histogram Feature

Toy Example Walkthrough



Bayesian Decision Theory

- The joint probability density of finding a sample which is in category ω_j and has feature value x is given by:

$$p(\omega_j, x) = P(\omega_j|x)p(x)$$
$$= p(x|\omega_j)P(\omega_j)$$

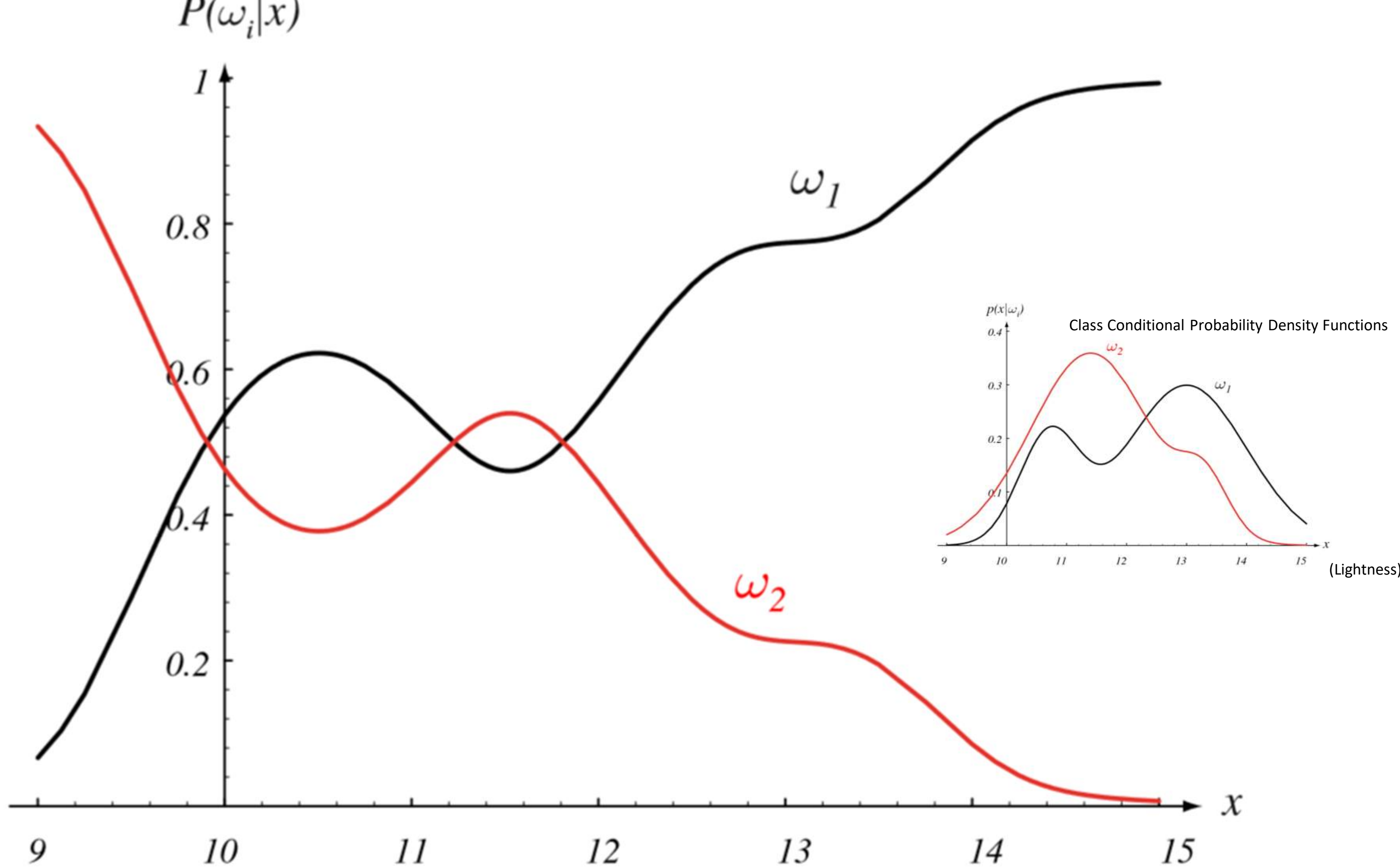
- **Bayes Formula**

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

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

Bayesian Decision Theory

Posteriori probabilities for fixed prior probabilities
$P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$



Bayesian Decision Theory

- Evidence act as normalizing term as:

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

- Bayes Formula (Two Category)

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

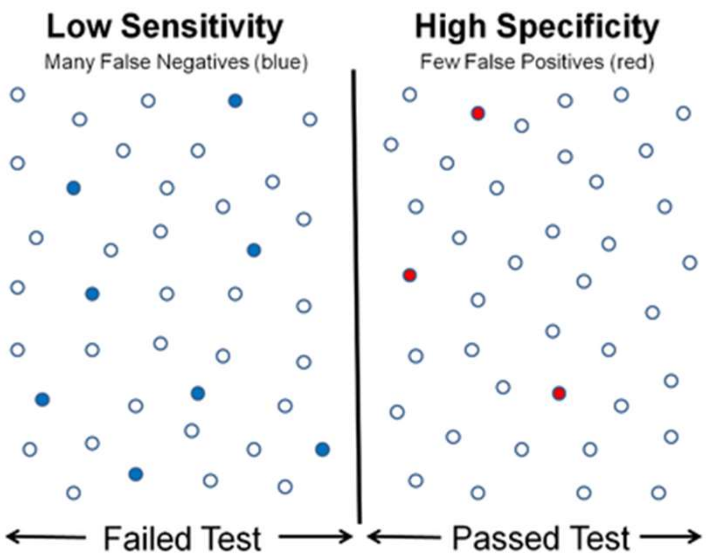
- **Bayes Decision Rule**

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

Bayesian Decision Theory

- Practical Example
  - Cancer Test

- **Sensitivity** (TPR), *Recall*, Hit-rate = $\frac{TP}{TP+F}$
- Precision (PPV) = $\frac{TP}{TP+}$
- **Specificity** (TNR) = $\frac{TN}{TN+FP}$
- NPV= $\frac{TN}{FN+TN}$
- Fall-out (FPR)= $\frac{FP}{FP+TN}$
- False Discovery Rate (FDR)= $\frac{FP}{FP+TP}$



Bayesian Decision Theory

- Bayes formula can be generalized to:
  - Multi-dimensional feature space i.e., $\mathbf{x} = [x_1, \quad \cdots, \quad x_d]^T$
  - Multiple classes i.e., $\{\omega_1, \dots, \omega_c\}$
  - Allowing more generic actions like rejection apart from assigning class label $\{\alpha_1, \dots, \alpha_a\}$

- $P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)\, P(\omega_j)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\omega_j)\, P(\omega_j)}{\sum_{i=1}^{c} p(\mathbf{x}|\omega_i) P(\omega_i)}$

- Let's define loss function as $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$ where α_i is the action taken while being in stat of nature ω_j.

Bayesian Decision Theory

- Bayes Risk (*conditional risk*)

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

- Two-Category Classification

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

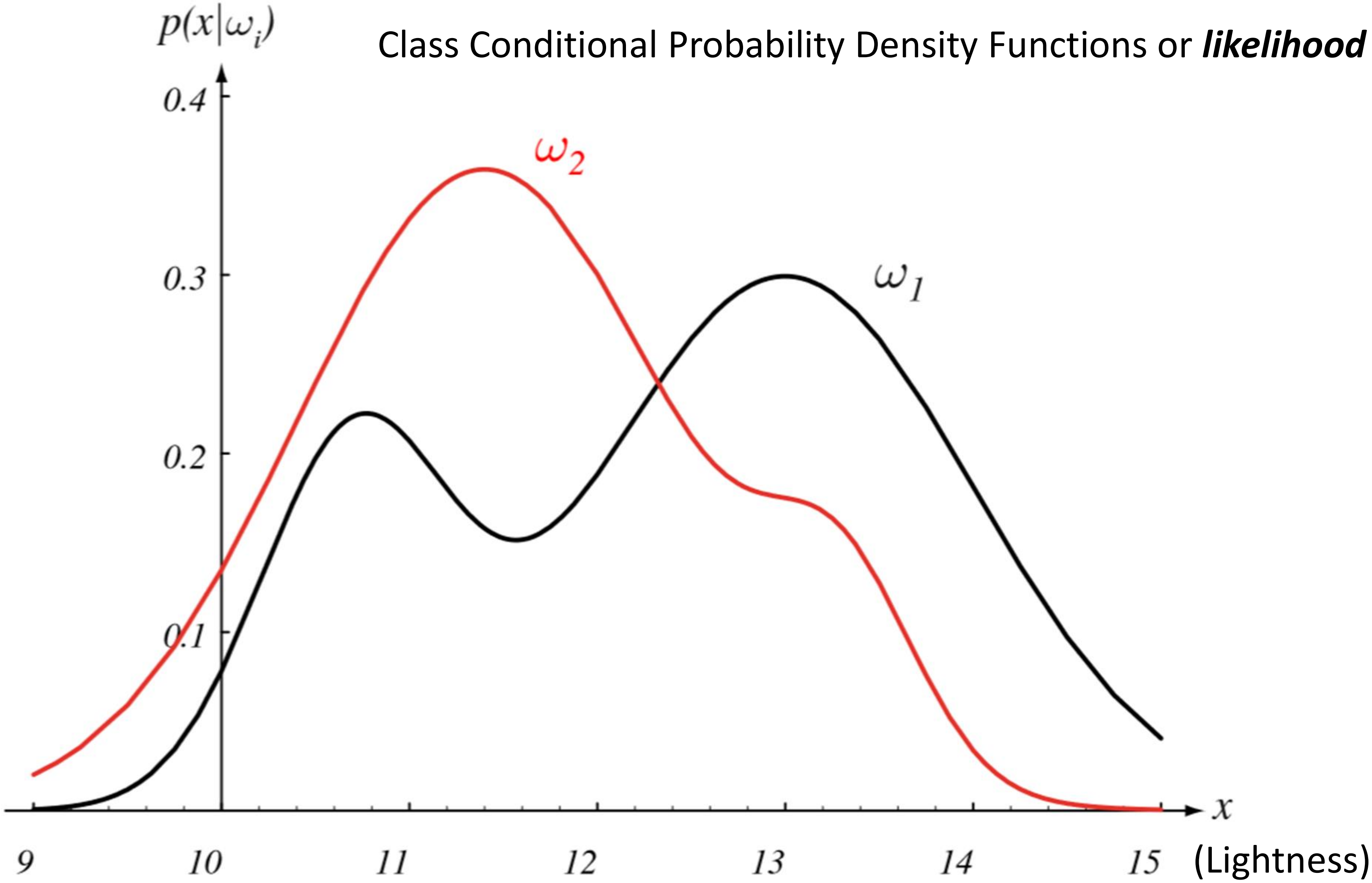
  – Choose ω_1 if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$ or

$$(\lambda_{21} - \lambda_{11})\, P(\omega_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})\, P(\omega_2|\mathbf{x}) \text{ or}$$

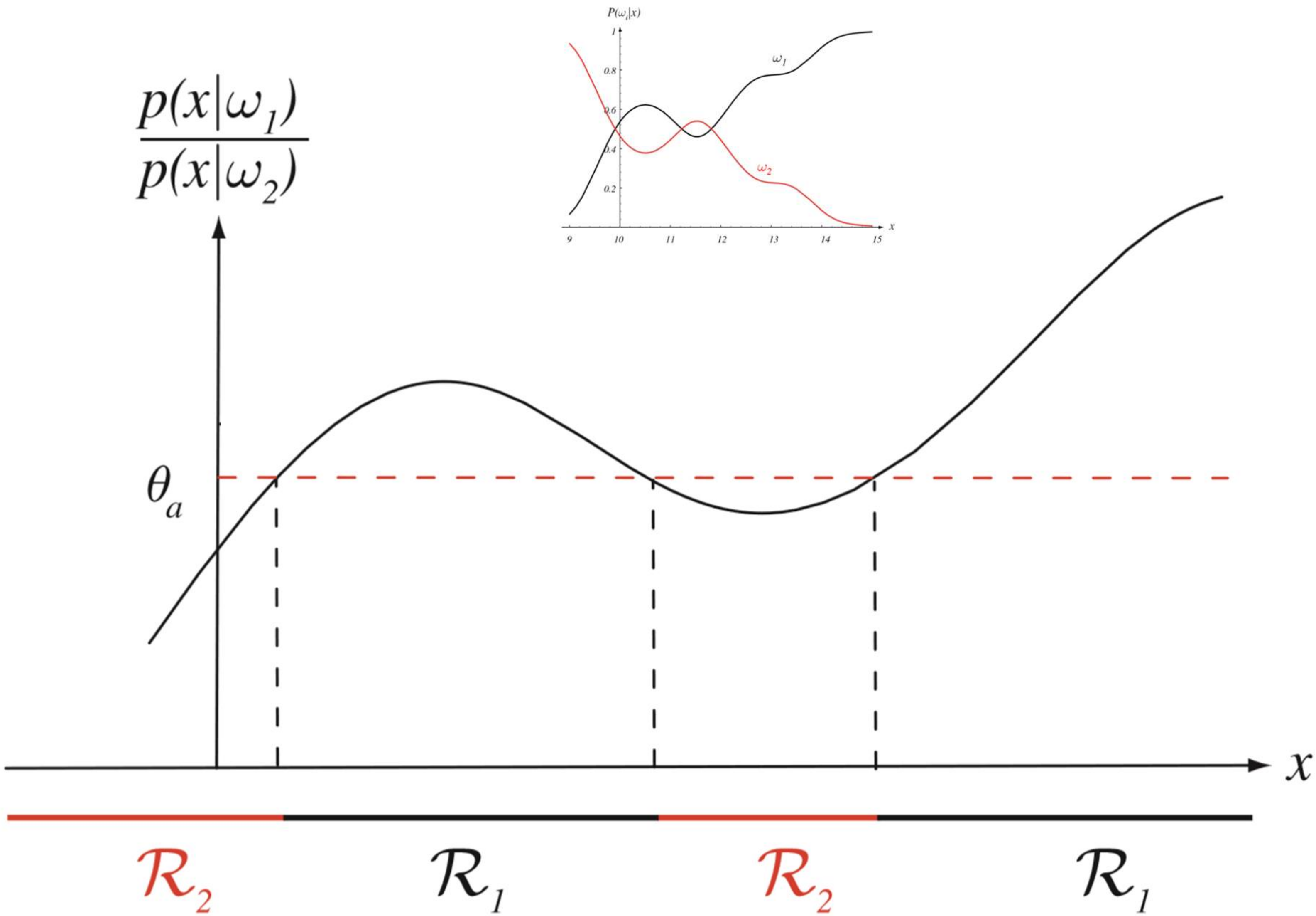
$$(\lambda_{21} - \lambda_{11})p(\mathbf{x}|\omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22})p(\mathbf{x}|\omega_2)P(\omega_2) \text{ or}$$

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

Bayesian Decision Theory



Bayesian Decision Theory



Minimum-Error-Rate Classification

- Let $\lambda_{ij} = \lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$

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

- If we choose ω_i corresponding to largest $P(\omega_i|\mathbf{x})$ then we minimize $R(\alpha_i|\mathbf{x})$

- Decision rule:
  Decide ω_i if $P(\omega_i|x) > P(\omega_j|x) \quad \forall j \neq i$

Minimum-Error-Rate Classification

- Let $\lambda_{12} > \lambda_{21}$(penalize miss-categorizing ω_2 as ω_1)

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