

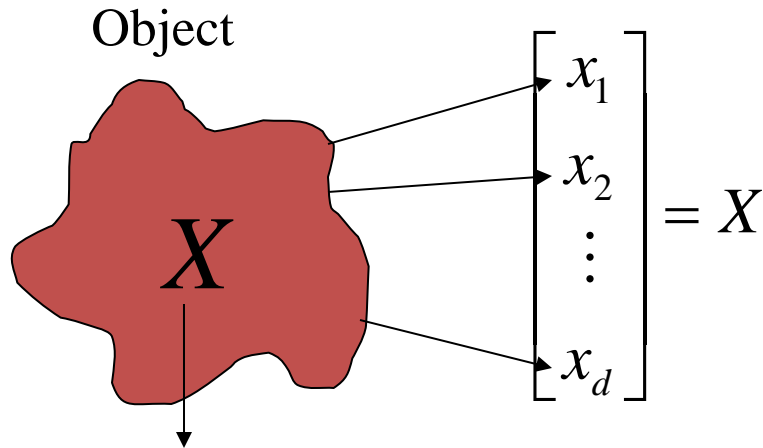
Bayesian Decision Theory

Primary source of reference: *Pattern Classification* – Duda
and Hart

Introduction

Bayesian Decision Theory–Continuous Features

Basic concepts



Feature vector $X \in \mathcal{X}$

- A vector of observations (measurements).
- X is a point in feature space \mathcal{X} .

Class to which X belongs is $y \in Y$

-Needs to be estimated, based on training set.

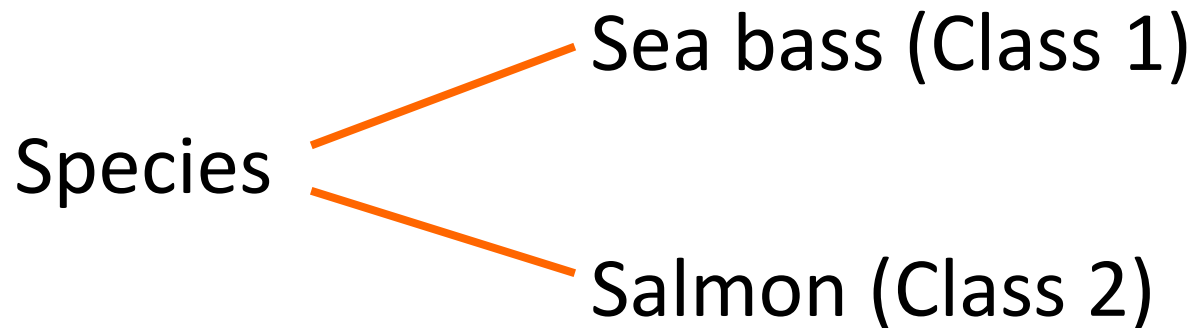
Task

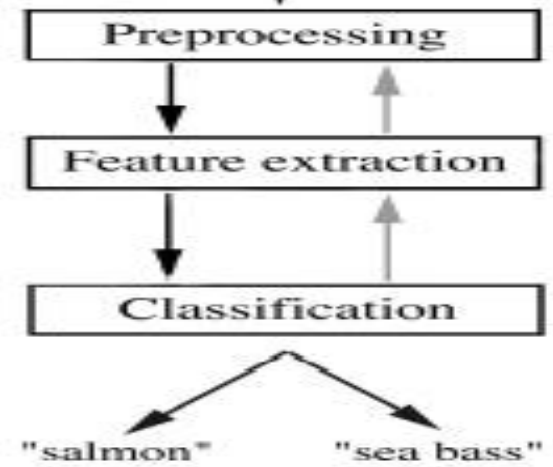
- To design a classifier (decision rule) $f : \mathcal{X} \rightarrow Y$ which decides about the class label based on X .

- Given X , we want to find its class label.
- For this we use the function f
- We give X as input to the function f and we output $y = f(X)$.
- f is normally learnt from the given training set which is $\{(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)\}$.

An Example

- “Sorting incoming Fish on a conveyor according to species using optical sensing”



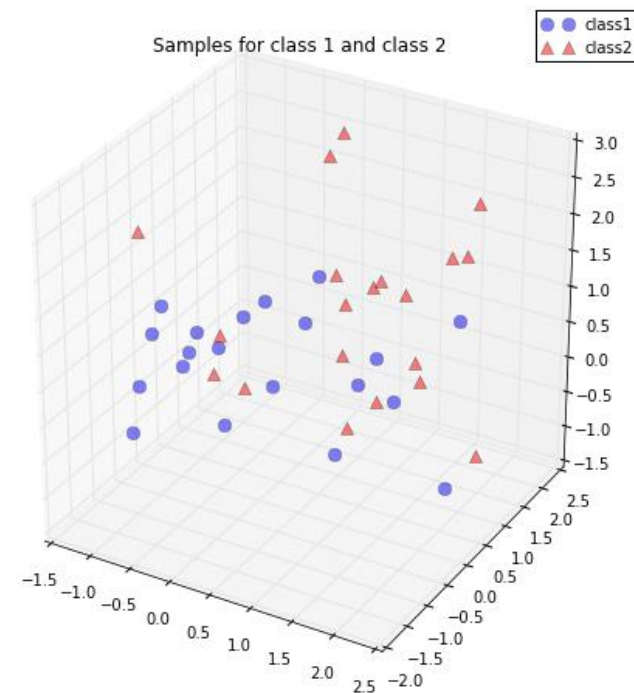


- Problem Analysis

- Set up a camera and take some sample images to extract features like

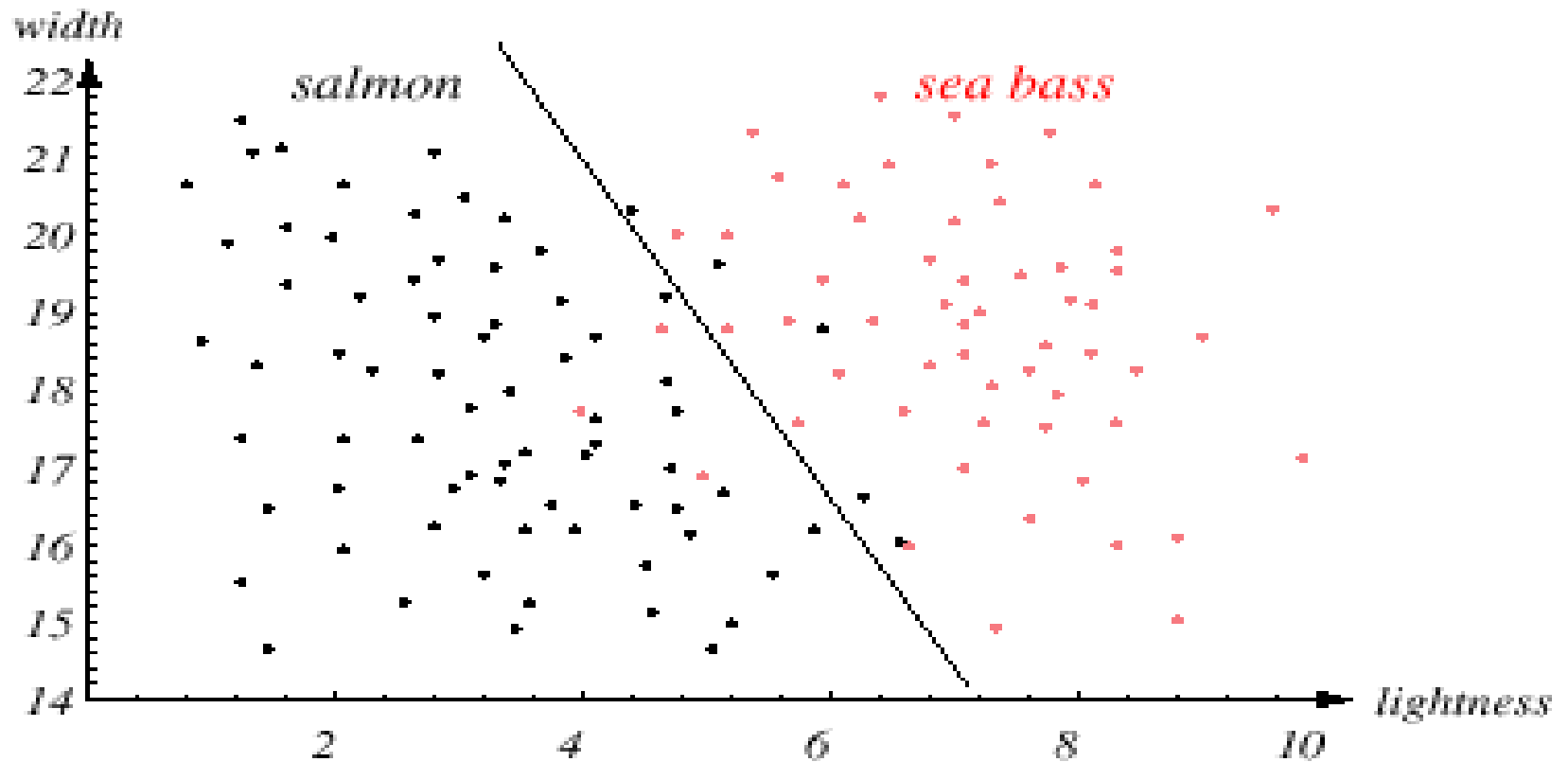
- Length of the fish
 - Lightness (based on the gray level)
 - Width of the fish

- So, $X = \begin{bmatrix} length \\ Lightness \\ Width \end{bmatrix}$



Considering only two features –

A linear classifier also is shown

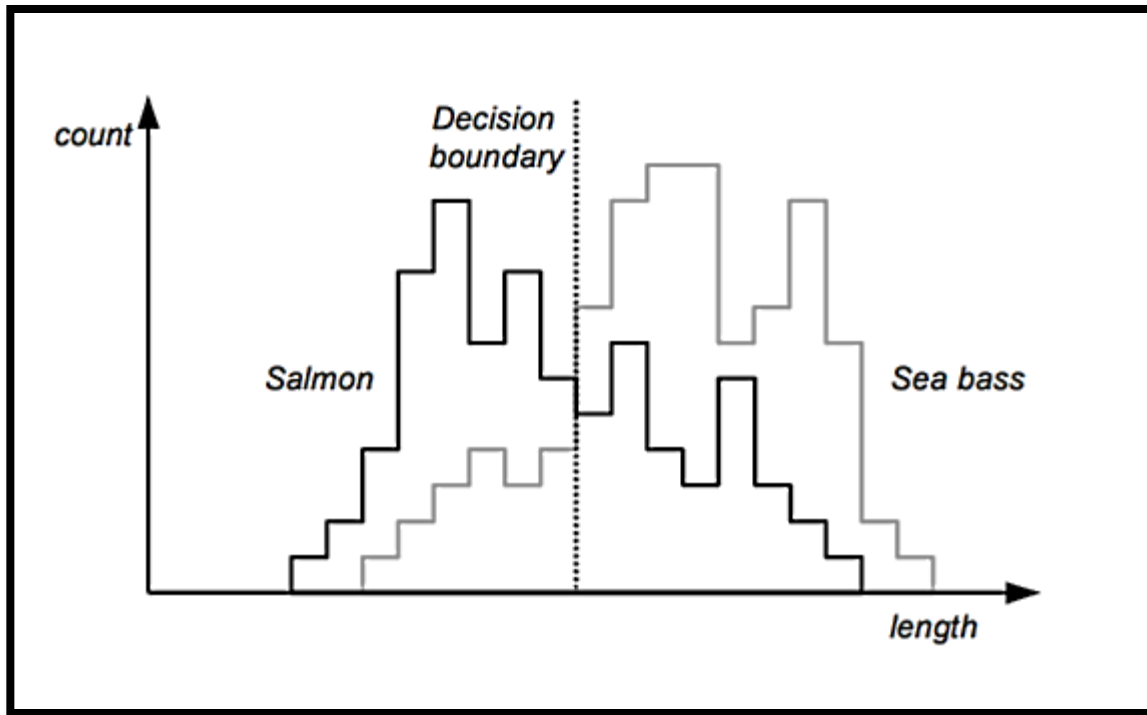


Introduction – Apriori knowledge

- The sea bass/salmon example
(a two class problem)
 - For example if we randomly catch 100 fishes and out of this if 75 are *sea bass* and 25 are *salmon*.
 - Let the rule, in this case is: For any fish say its class is *sea bass*.
 - What is the error rate of this rule?
 - This information which is independent of feature values is called **apriori** knowledge.

- Let the two classes are ω_1 and ω_2
 - $P(\omega_1) + P(\omega_2) = 1$
 - State of nature (class) is a random variable
 - If $P(\omega_1) = P(\omega_2)$, we say it is of uniform priors
 - The catch of salmon and sea bass is equi-probable

- Perhaps we considered 100 fish of Salmon and 100 fish of sea bass and found the following...



- Using only information $P(X|class)$?
- Is this good without considering the apriori knowledge?

- Decision rule with only the prior information
 - Decide ω_1 if $P(\omega_1) > P(\omega_2)$, otherwise decide ω_2
- *This is not a good classifier.*
- *We should take feature values into account !*
- *If x is the pattern we want to classify, then use the rule:*

*If $P(\omega_1 | x) > P(\omega_2 | x)$ then assign class ω_1
 Else assign class ω_2*

- *$P(\omega_1 | x)$ is called posteriori probability of class ω_1 given that the pattern is x .*

Bayes rule

- From data it might be possible for us to estimate $p(x | \omega_i)$, where $i = 1$ or 2 . These are called **class-conditional distributions**.
- Also it is easy to find apriori probabilities $P(\omega_1)$ and $P(\omega_2)$. **How this can be done?**
- Bayes rule combines apriori probability with class conditional distributions to find posteriori probabilities.

Bayes Rule

$$P(B|A) = \frac{P(A, B)}{P(A)} = \frac{P(A|B) * P(B)}{P(A)}$$

This is Bayes Rule



Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

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

– Where in case of two categories

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

– Posterior = $\frac{\text{Likelihood} \cdot \text{Prior}}{\text{Evidence}}$

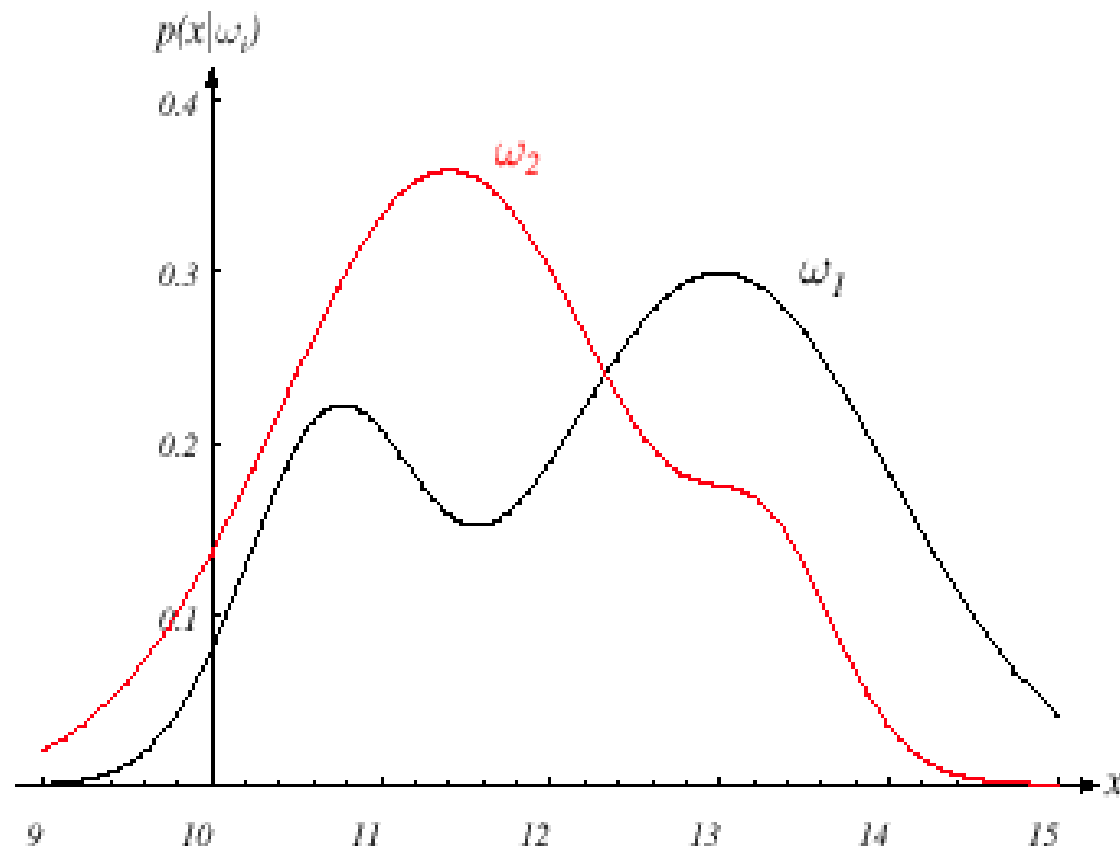


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i . If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two 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.

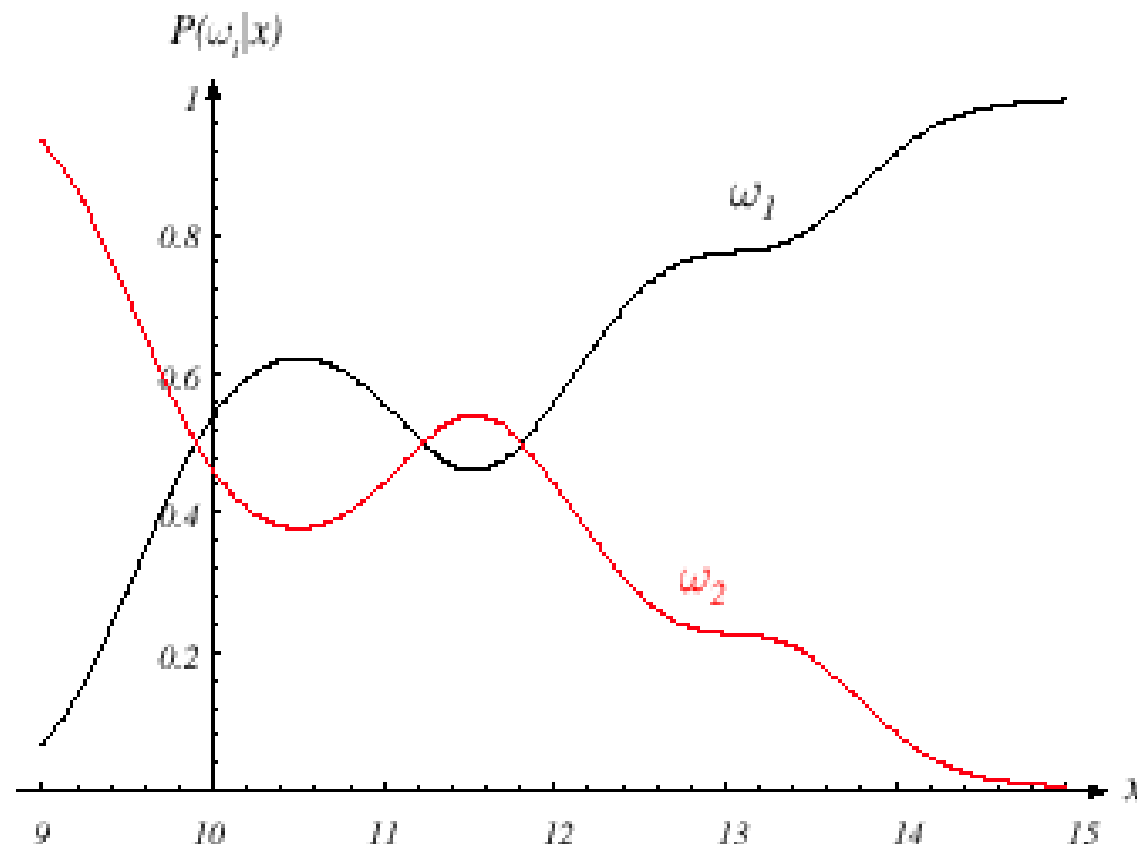
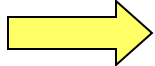
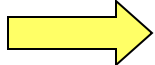


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 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*. Copyright © 2001 by John Wiley & Sons, Inc.

- Decision given the posterior probabilities

X is an observation for which:

if $P(\omega_1 | x) > P(\omega_2 | x)$  True state of nature = ω_1

if $P(\omega_1 | x) < P(\omega_2 | x)$  True state of nature = ω_2

Therefore:

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

$P(\text{error} | x) = P(\omega_1 | x)$ if we decide ω_2

$P(\text{error} | x) = P(\omega_2 | x)$ if we decide ω_1

- Minimizing the probability of error
- Decide ω_1 if $P(\omega_1 | x) > P(\omega_2 | x)$;
otherwise decide ω_2

Therefore:

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

(error of Bayes decision)

Average error rate

Average probability of error, $P(\text{error})$ is :

$$\int P(\text{error} | x) p(x) dx$$

This is the expected value of $P(\text{error}/x)$ w.r.t. x ,

i.e., $E_x[P(\text{error} / x)]$

1. Consider a one dimensional two class problem. The feature used is color of fish. Color can be either white or dark $P(\omega_1) = 0.75$, $P(\omega_2) = 0.25$, $P(\text{white} | \omega_1) = 0.2$, $P(\text{white} | \omega_2) = 0.6$, $P(\text{dark} | \omega_1) = 0.8$, $P(\text{dark} | \omega_2) = 0.4$ Find $P(\text{error})$ of the Bayes Classifier.

1. Consider a one dimensional two class problem. The feature used is color of fish. Color can be either white or dark $P(\omega_1) = 0.75$, $P(\omega_2) = 0.25$, $P(\text{white} | \omega_1) = 0.2$, $P(\text{white} | \omega_2) = 0.6$, $P(\text{dark} | \omega_1) = 0.8$, $P(\text{dark} | \omega_2) = 0.4$ Find $P(\text{error})$ of the Bayes Classifier.

$$P(\text{white}) = P(\text{white} | \omega_1)P(\omega_1) + P(\text{white} | \omega_2)P(\omega_2)$$

$$P(\text{white}) = 0.2 * 0.75 + 0.6 * 0.25 = 0.3$$

1. Consider a one dimensional two class problem. The feature used is color of fish. Color can be either white or dark $P(\omega_1) = 0.75$, $P(\omega_2) = 0.25$, $P(\text{white} | \omega_1) = 0.2$, $P(\text{white} | \omega_2) = 0.6$, $P(\text{dark} | \omega_1) = 0.8$, $P(\text{dark} | \omega_2) = 0.4$ Find $P(\text{error})$ of the Bayes Classifier.

$$P(\text{white}) = P(\text{white} | \omega_1)P(\omega_1) + P(\text{white} | \omega_2)P(\omega_2)$$

$$P(\text{white}) = 0.2 * 0.75 + 0.6 * 0.25 = 0.3$$

$$P(\text{dark}) = P(\text{dark} | \omega_1)P(\omega_1) + P(\text{dark} | \omega_2)P(\omega_2)$$

$$P(\text{dark}) = 0.8 * 0.75 + 0.4 * 0.25 = 0.7$$

1. Consider a one dimensional two class problem. The feature used is color of fish. Color can be either white or dark $P(\omega_1) = 0.75$, $P(\omega_2) = 0.25$, $P(\text{white} | \omega_1) = 0.2$, $P(\text{white} | \omega_2) = 0.6$, $P(\text{dark} | \omega_1) = 0.8$, $P(\text{dark} | \omega_2) = 0.4$ Find $P(\text{error})$ of the Bayes Classifier.

$$P(\text{white}) = P(\text{white} | \omega_1)P(\omega_1) + P(\text{white} | \omega_2)P(\omega_2)$$

$$P(\text{white}) = 0.2 * 0.75 + 0.6 * 0.25 = 0.3$$

$$P(\text{dark}) = P(\text{dark} | \omega_1)P(\omega_1) + P(\text{dark} | \omega_2)P(\omega_2)$$

$$P(\text{dark}) = 0.8 * 0.75 + 0.4 * 0.25 = 0.7$$

$$P(\omega_1 | \text{white}) = \frac{P(\text{white} | \omega_1)P(\omega_1)}{P(\text{white})} = \frac{0.2 * 0.75}{0.3} = 0.5$$

$$P(\omega_2 | \text{white}) = \frac{P(\text{white} | \omega_2)P(\omega_2)}{P(\text{white})} = \frac{0.6 * 0.25}{0.3} = 0.5$$

1. Consider a one dimensional two class problem. The feature used is color of fish. Color can be either white or dark $P(\omega_1) = 0.75$, $P(\omega_2) = 0.25$, $P(\text{white} | \omega_1) = 0.2$, $P(\text{white} | \omega_2) = 0.6$, $P(\text{dark} | \omega_1) = 0.8$, $P(\text{dark} | \omega_2) = 0.4$ Find $P(\text{error})$ of the Bayes Classifier.

$$P(\text{white}) = P(\text{white} | \omega_1)P(\omega_1) + P(\text{white} | \omega_2)P(\omega_2)$$

$$P(\text{white}) = 0.2 * 0.75 + 0.6 * 0.25 = 0.3$$

$$P(\text{dark}) = P(\text{dark} | \omega_1)P(\omega_1) + P(\text{dark} | \omega_2)P(\omega_2)$$

$$P(\text{dark}) = 0.8 * 0.75 + 0.4 * 0.25 = 0.7$$

$$P(\omega_1 | \text{white}) = \frac{P(\text{white} | \omega_1)P(\omega_1)}{P(\text{white})} = \frac{0.2 * 0.75}{0.3} = 0.5$$

$$P(\omega_2 | \text{white}) = \frac{P(\text{white} | \omega_2)P(\omega_2)}{P(\text{white})} = \frac{0.6 * 0.25}{0.3} = 0.5$$

$$P(\omega_1 | \text{dark}) = \frac{P(\text{dark} | \omega_1)P(\omega_1)}{P(\text{dark})} = \frac{0.8 * 0.75}{0.7} = \frac{6}{7}$$

$$P(\omega_2 | \text{dark}) = \frac{P(\text{dark} | \omega_2)P(\omega_2)}{P(\text{dark})} = \frac{0.4 * 0.25}{0.7} = \frac{1}{7}$$

$$P(error) = P(error|white)P(white) + P(error|dark)P(dark)$$

$$P(error) = 0.5 * 0.3 + \frac{1}{7} * 0.7 = 0.25$$

$$P(error) = P(error|white)P(white) + P(error|dark)P(dark)$$

$$P(error) = 0.5 * 0.3 + \frac{1}{7} * 0.7 = 0.25$$

- But, what is the error, if we use only apriori probabilities?

Since, $P(\omega_1) = 0.75$, $P(\omega_2) = 0.25$, every pattern is assigned to ω_1 , So the error,

$$P(error) = P(\omega_2|white)P(white) + P(\omega_2|dark)P(dark)$$

Since, $P(\omega_1) = 0.75$, $P(\omega_2) = 0.25$, every pattern is assigned to ω_1 , So the error,

$$P(error) = P(\omega_2|white)P(white) + P(\omega_2|dark)P(dark)$$

$$P(error) = \frac{P(white|\omega_2)P(\omega_2)}{P(white)}P(white) + \frac{P(dark|\omega_2)P(\omega_2)}{P(dark)}P(dark)$$

Since, $P(\omega_1) = 0.75$, $P(\omega_2) = 0.25$, every pattern is assigned to ω_1 , So the error,

$$P(error) = P(\omega_2|white)P(white) + P(\omega_2|dark)P(dark)$$

$$P(error) = \frac{P(white|\omega_2)P(\omega_2)}{P(white)}P(white) + \frac{P(dark|\omega_2)P(\omega_2)}{P(dark)}P(dark)$$

$$P(error) = (P(white|\omega_2) + P(dark|\omega_2))P(\omega_2)$$

$$P(error) = P(\omega_2) = 0.25$$

Since, $P(\omega_1) = 0.75$, $P(\omega_2) = 0.25$, every pattern is assigned to ω_1 , So the error,

$$P(error) = P(\omega_2|white)P(white) + P(\omega_2|dark)P(dark)$$

$$P(error) = \frac{P(white|\omega_2)P(\omega_2)}{P(white)}P(white) + \frac{P(dark|\omega_2)P(\omega_2)}{P(dark)}P(dark)$$

$$P(error) = (P(white|\omega_2) + P(dark|\omega_2))P(\omega_2)$$

$$P(error) = P(\omega_2) = 0.25$$

- Same error? Where is the advantage?!

Consider $P(\omega_1) = 0.5$, $P(\omega_2) = 0.5$

$$P(\text{white}) = P(\text{white}|\omega_1)P(\omega_1) + P(\text{white}|\omega_2)P(\omega_2)$$

$$P(\text{white}) = 0.2 * 0.5 + 0.6 * 0.5 = 0.4$$

$$P(\text{dark}) = P(\text{dark}|\omega_1)P(\omega_1) + P(\text{dark}|\omega_2)P(\omega_2)$$

$$P(\text{dark}) = 0.8 * 0.5 + 0.4 * 0.5 = 0.6$$

$$P(\omega_1|\text{white}) = \frac{P(\text{white}|\omega_1)P(\omega_1)}{P(\text{white})} = \frac{0.2 * 0.5}{0.4} = 0.25$$

$$P(\omega_2|\text{white}) = \frac{P(\text{white}|\omega_2)P(\omega_2)}{P(\text{white})} = \frac{0.6 * 0.5}{0.4} = 0.75$$

$$P(\omega_1|\text{dark}) = \frac{P(\text{dark}|\omega_1)P(\omega_1)}{P(\text{dark})} = \frac{0.8 * 0.5}{0.6} = \frac{2}{3}$$

$$P(\omega_2|\text{dark}) = \frac{P(\text{dark}|\omega_2)P(\omega_2)}{P(\text{dark})} = \frac{0.4 * 0.5}{0.6} = \frac{1}{3}$$

$$P(error) = P(error|white)P(white) + P(error|dark)P(dark)$$

$$P(error) = 0.25 * 0.4 + \frac{1}{3} * 0.6 = 0.3$$

- But, $P(error)$ based on apriori probabilities only is 0.5.
- Error based on the Bayes classifier is the lower bound.
 - Any classifier's error is greater than or equal to this.
- One can prove this!

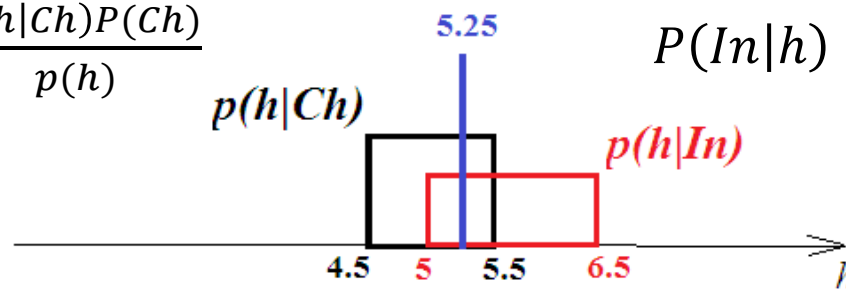
2. We know that 30% of people in Sri City are from China. Remaining are Indians. We know that Chinese height is a uniformly distributed random variable with parameters 4.5 and 5.5. We also know that Indians height is also uniformly distributed with parameters 5 and 6.5. By measuring his/her height we want to classify the person as "Chinese" or "Indian" (We know that the person is living in Sri City). We follow the rule "if height is 5.25 or below classify the person as Chinese, otherwise classify the person as Indian". There are two types of mistakes in this classification, (1) Chinese being classified as Indians, (2) Indians being classified as Chinese. Find the probability of making each of these two types of mistakes.

$$P(Ch) = 0.3$$

$$P(In) = 0.7$$

$$P(Ch|h) = \frac{p(h|Ch)P(Ch)}{p(h)}$$

$$P(In|h) = \frac{p(h|In)P(In)}{p(h)}$$



$$p(h) = p(h|Ch)P(Ch) + p(h|In)P(In)$$

$$= \begin{cases} 0.3, & \text{for } h \text{ in } [4.5, 5] \\ 1 * 0.3 + \frac{2}{3} * 0.7 = 0.767, & \text{for } h \text{ in } [5, 5.5] \\ \frac{2}{3} * 0.7 = 0.467, & \text{for } h \text{ in } [5.5, 6.5] \end{cases}$$

For $h \in [4.5, 5]$

$$P(Ch|h) = (1 * 0.3)/0.3 = 1$$

$$P(In|h) = 0$$

For $h \in [5, 5.5]$

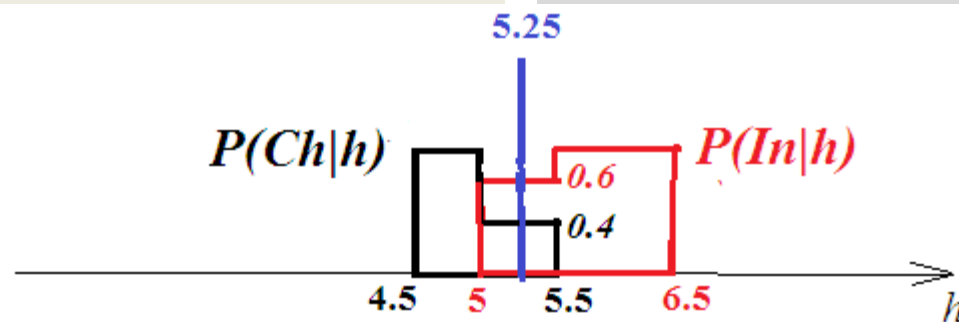
$$P(Ch|h) = (1 * 0.3)/0.767 = 0.39$$

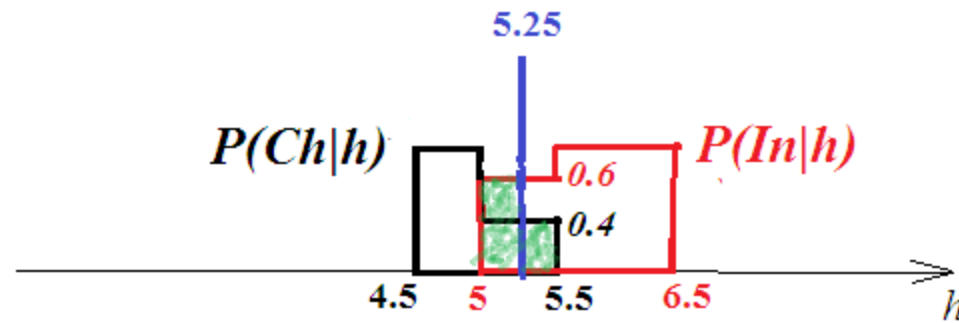
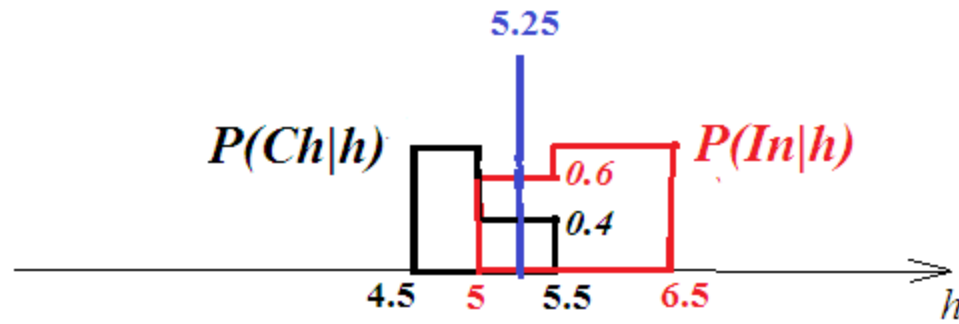
$$P(In|h) = 0.67 * 0.7/0.767 = 0.61$$

For $h \in [5.5, 6.5]$

$$P(Ch|h) = 0$$

$$P(In|h) = 1$$





Error is due to this region

$$\begin{aligned}
 P(\text{error}) &= \int_5^{5.25} 0.609 p(h) dh + \int_{5.25}^{5.5} 0.391 p(h) dh \\
 &= 0.117 + 0.075 = 0.192
 \end{aligned}$$

- But, this is not the Bayes classifier.
- What does the Bayes classifier do?
- What is the error of the Bayes classifier?

For $h \in [4.5, 5]$

$$P(Ch|h) = (1 * 0.3)/0.3 = 1$$

$$P(In|h) = 0$$

For $h \in [5, 5.5]$

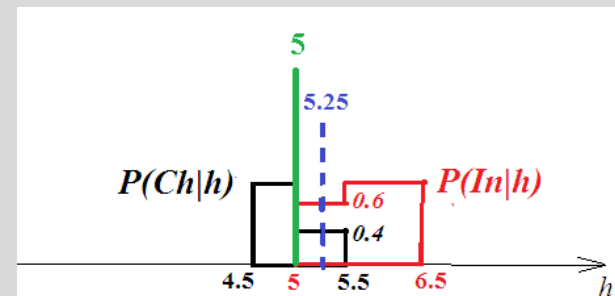
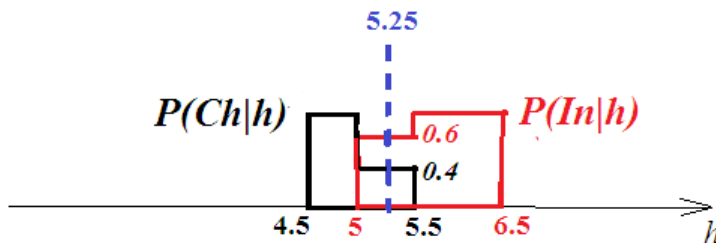
$$P(Ch|h) = (1 * 0.3)/0.767 = 0.39$$

$$P(In|h) = 0.67 * 0.7/0.767 = 0.61$$

For $h \in [5.5, 6.5]$

$$P(Ch|h) = 0$$

$$P(In|h) = 1$$



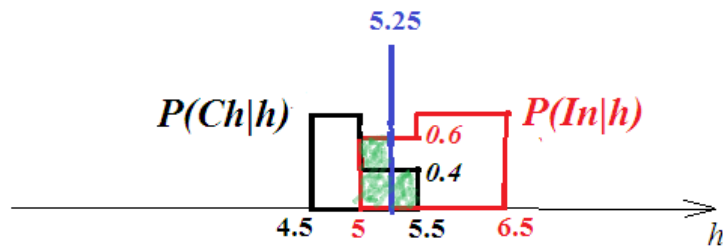
If $h > 5$ classify as Indian

Else classify as Chinese

Bayes Classifier's Error

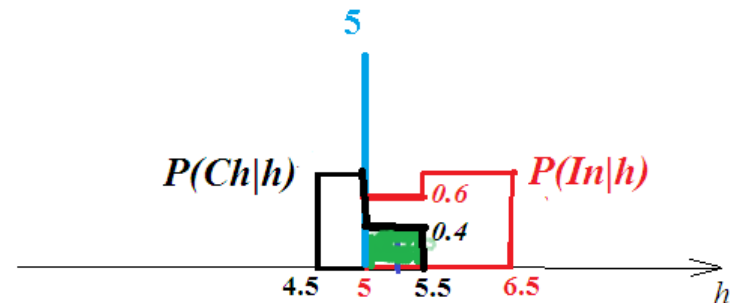
- $$\begin{aligned} P(\text{error}) &= \int_5^{5.5} 0.391 p(h) dh \\ &= (0.391) * (0.767) * (5.5 - 5) \\ &= 0.15 \end{aligned}$$

Error diagnosis



Error is due to this region

Previous classifier



Error is due to this region

The Bayes Classifier

Bayesian Decision Theory – Continuous Features

- Generalization of the preceding ideas
 - Use of more than one feature
 - Use more than two states (classes) of nature
 - Allowing actions (decisions) other than just classification.
 - Introduce a *loss function* which is more general than the probability of error.

- Allowing actions other than classification primarily allows the possibility of rejection
- Refusing to make a decision in close or bad cases!
- The loss function states how costly each action taken is

- Given the pattern X , we want to find the best possible action among actions set
$$A = \{\alpha_1, \alpha_2, \dots, \alpha_a\}$$
- For example : α_1 is the action *ring the alarm*
 α_2 is the action *shutdown the system*
- Outputting the class prediction is one action.

Problem setting

What is given to us:

Loss function : $\lambda(\alpha_i / \omega_j)$ is the loss of taking action α_i when the state of nature is ω_j

Let $\Omega = \{\omega_1, \omega_2, \dots, \omega_c\}$ be the set of c states of nature

(or “categories” or “classes”)

Let $A = \{\alpha_1, \alpha_2, \dots, \alpha_a\}$ be the set of possible actions

Let $\Omega = \{\omega_1, \omega_2, \dots, \omega_c\}$ be the set of c states of nature
(or “categories” or “classes”)

Let $A = \{\alpha_1, \alpha_2, \dots, \alpha_a\}$ be the set of possible actions.

Objective: Given a pattern x , find the action to take.

That is, to find a function $\alpha(x)$ which maps x to action.

Input

X



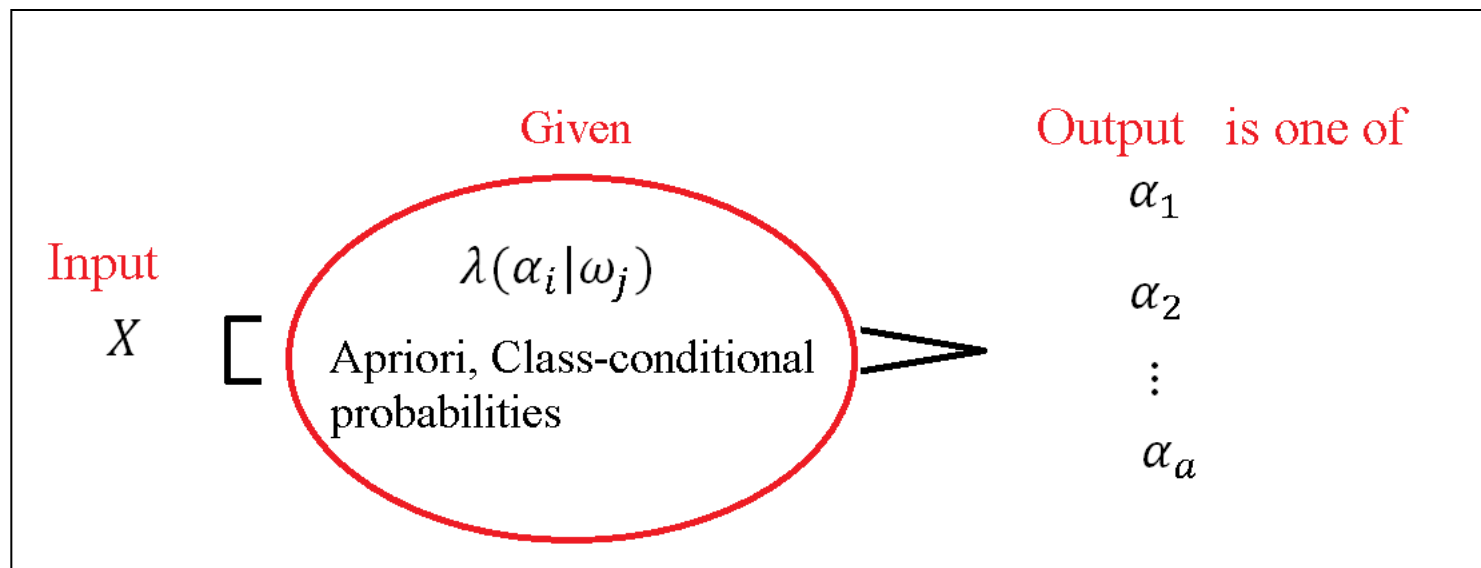
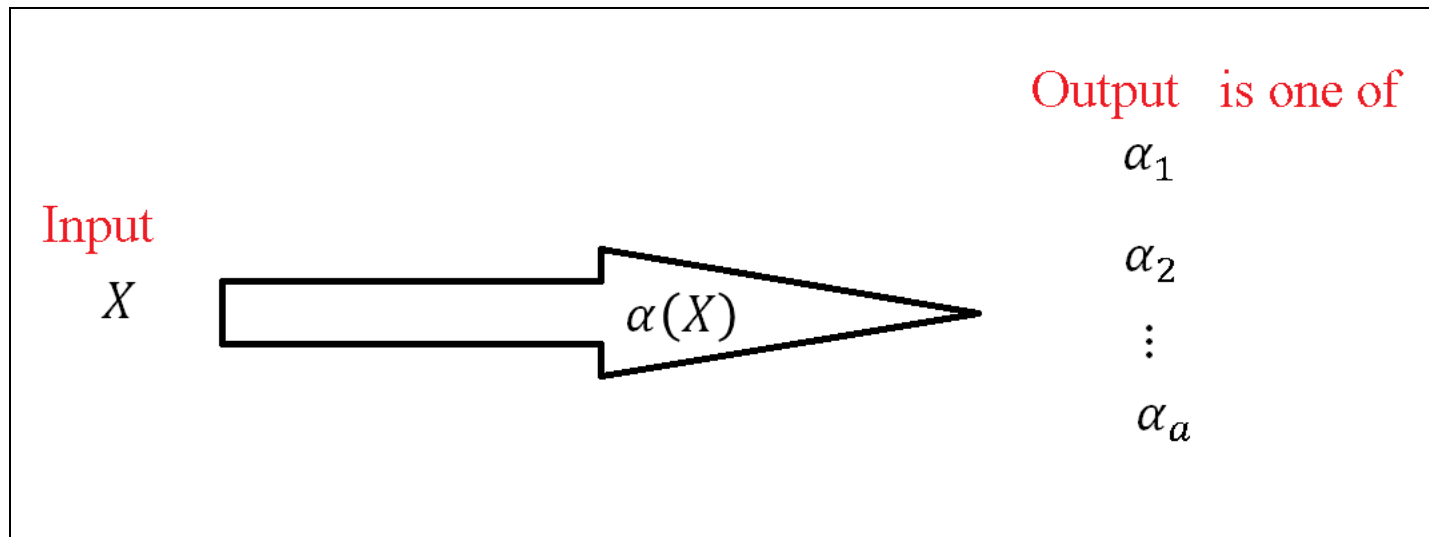
Output is one of

α_1

α_2

\vdots

α_a



How to find the best action?

Let $R(\alpha_i | x)$ is the risk of taking action α_i when the given pattern is x (this is called conditional risk).

We can take the action α_k provided $R(\alpha_k | x)$ is minimum in $\{ R(\alpha_1 | x), R(\alpha_2 | x), \dots, R(\alpha_a | x) \}$

How to relate $R(\alpha_i | x)$ with $\lambda(\alpha_i | \omega_j)$ values.

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

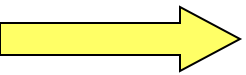
$\alpha(x)$ is given, how good is this?

- Overall risk of this rule can be found

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

- The smaller this quantity, better the decision rule.

Select the action α_i for which $R(\alpha_i / x)$ is minimum



R is minimum and R in this case is called the

Bayes risk = best performance that can be achieved!

- Two-category classification

α_1 : deciding ω_1

α_2 : deciding ω_2

$$\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$$

loss incurred for deciding ω_i when the true state of nature is ω_j

Conditional risk:

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

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

Our rule is the following:

if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$
action α_1 : “decide ω_1 ” is taken

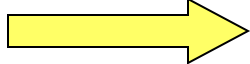
This results in the equivalent rule :
decide ω_1 if:

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

and decide ω_2 otherwise

An Example

- Let the two actions are:

α_1  ***x is criminal***

α_2  ***x is innocent***

Let $\lambda(\alpha_1 | \omega_1) = 0$; $\lambda(\alpha_1 | \omega_2) = 10$;

$\lambda(\alpha_2 | \omega_1) = 1$; $\lambda(\alpha_2 | \omega_2) = 0$;

Assume equal priors, and

Let $p(x | \omega_1) = 0.8$, $p(x | \omega_2) = 0.6$

What action you will take?

ω_1 is class of criminals,

ω_2 is class of innocents.

Example: Contd...

- $$P(\omega_j | x) = \frac{p(x | \omega_j) \cdot P(\omega_j)}{p(x)}$$
- $$P(\omega_1 | x) = \frac{0.8(0.5)}{0.8(0.5) + 0.6(0.5)} = 8/14 = 0.57$$
- $$P(\omega_2 | x) = 0.43$$
- *As per plain Bayes rule we declare the person to be a criminal.*

Example: Contd ...

- $R(\alpha_1 | x) = \lambda(\alpha_1 | \omega_1) P(\omega_1 | x) + \lambda(\alpha_1 | \omega_2) P(\omega_2 | x)$
 $= 0 (0.57) + 10 (0.43)$
 $= 4.3$
- $R(\alpha_2 | x) = \lambda(\alpha_2 | \omega_1) P(\omega_1 | x) + \lambda(\alpha_2 | \omega_2) P(\omega_2 | x)$
 $= 1 (0.57) + 0 (0.43)$
 $= 0.57$

Action taken: x is innocent

Likelihood ratio:

The preceding rule is equivalent to the following rule:

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

Then take action α_1 (decide ω_1)

Otherwise take action α_2 (decide ω_2)

Example 1

- Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$, $A = \{\alpha_1, \alpha_2, \alpha_3\}$
 $P(\omega_1 | x) = \mathbf{0.1}$, $P(\omega_2 | x) = \mathbf{0.4}$, $P(\omega_3 | x) = \mathbf{0.5}$
Loss function is

Loss	ω_1	ω_2	ω_3
α_1	0	1	2
α_2	1	0	2
α_3	3	10	0

Find $R(\alpha_k | x)$ for $k = 1, 2, 3$.
Find what is the best action?

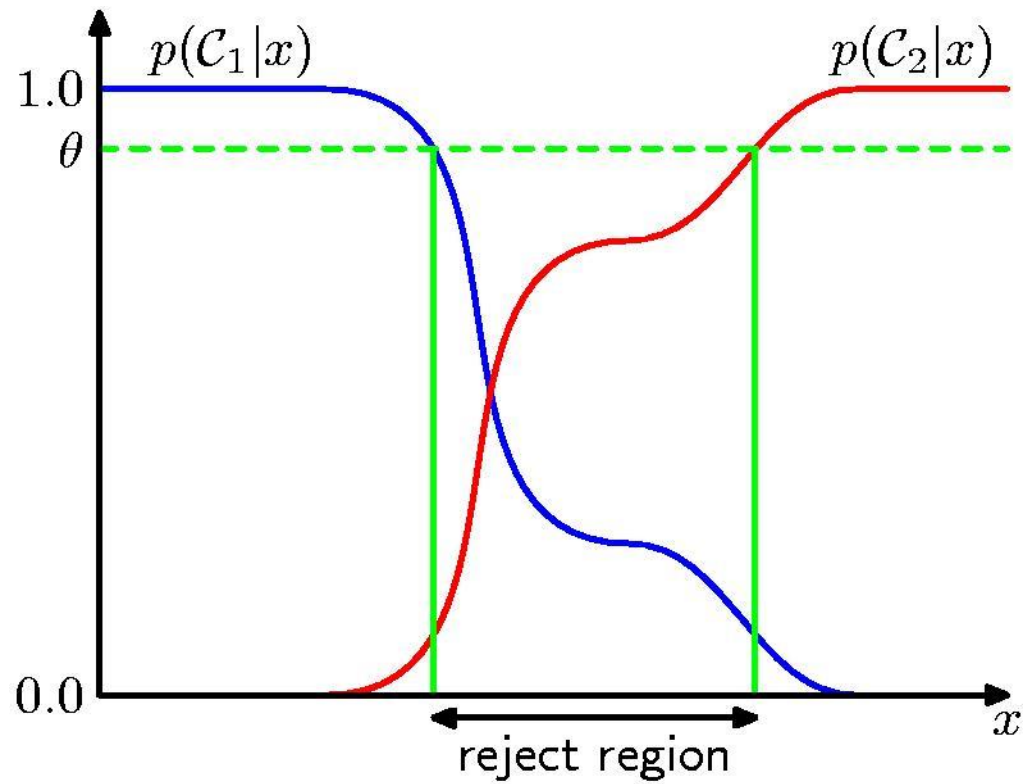
Reject option (I do not want to classify)

Actions are : Assign a class label or reject

Sometimes, when misclassification is costly, we can reject to classify it.

May be some other expert can look into it and can take appropriate action.

Reject Option



- Read Duda and Hart book and try to solve some problems related to this.