Chapter 2 (Part 1): Bayesian Decision Theory (Sections 2.1-2.2)

- Probability Theory
- Introduction
- Bayesian Decision Theory–Continuous Features

Probability Theory

- Definitions: If *x* is a discrete random variable that can assume any of the values in the finite set *X={v1,v2,…,vn}*, then we denote by *pk* the probability of *x* assuming the value *vk*.

  $0 \leq pk \leq 1$ and $\Sigma pk=1$ for $k=1,\ldots,n$

- In general, the probability of event "x" occurring is *P(x)*:

  $0 \leq P(x) \leq 1$ and $\Sigma_{x \hat{I} X} P(x)=1$

- *If y is a random variable that can take values in Y={w1,w2,…,wm}, then the joint probability of x =vi and y =wj is pij or P(xÇy). In general:*

  $0 \leq P(x \text{Ç} y) \leq 1$ and $\Sigma_{x \hat{I} X} \Sigma_{y \hat{I} y Y} P(x \text{ Ç} y)=1$

Probabilistic Events

- Independent events occur independently of each other:

  $P(e1 \cap e2 \ldots \cap en) = P(e1)P(e2) \ldots P(en)$

- Mutually exclusive events cannot occur at the same time:

  $P(e1 \cap e2) = 0$

  $P(e1 \cup e2 \ldots \cup en) = P(e1) + P(e2) \ldots + P(en)$

- Exhaustive events are a set of events from which at least one of the events will occur:

  $P(e1 \cup e2 \ldots \cup en) = 1$

- Mutually exclusive and exhaustive events:

  $P(e1) + P(e2) \ldots + P(en) = 1$

Random Variables

Let x be random variables that can take values in the domain X ($x \in X$), and let $P(x)$ be its probability of occurrence.

- Mean (or expected value or average) is defined as:

$$m = \Sigma_{x \hat{I} X} \, xP(x)$$

- Variance is defined as:

$$Var(x) = s^2 = \Sigma_{x \hat{I} X} \, (x - m)^2 P(x)$$

- Covariance of x and y is defined as:

$$s_{xy} = \Sigma_{x \hat{I} X} \, \Sigma_{y \hat{I} Y} \, (x - m_x)\,(y - m_y)\, P(x \text{Ç} y)$$

- Chauchy-Schwarz inequality:

$$s_{xy}^2 \leq s_x^2 \, s_y^2$$

- Correlation coefficient is defined as:

$$r = s_{xy} / \, s_x \, s_y$$

Gaussian Distribution

- Central Limit Theorem: the aggregate effect of a large number of random disturbances produces a normal (or Gaussian) distribution
  - The convolution of an infinite number of identical distributions is Gaussian.
  - Since many patterns (including handwritten characters and voice) can be viewed as some ideal (or prototype) patterns corrupted by large number of random processes, the Gaussian is often a good model for the actual probability distribution.
- The normal (Gaussian) distribution is completely described by its mean μ and variance σ^2 (or standard deviation σ).

$$P(x) = \frac{1}{\sqrt{2p}\,s} exp\left[-\frac{1}{2}\left(\frac{x-m}{s}\right)^2\right],$$

Conditional Probability

- Definitions: If *x* and *y* are random variables that can take values in X and Y respectively, then the probability of *x* occurring in the condition of *y* (known as the conditional probability of *x* given y) is:

  $$P(x/y)=P(x \cap y)/P(y)$$

- Properties:
  - $P(\sim x/y)=1-P(x/y)$
  - $P(x/y)=P(x)$, if, and only if, *x* and *y* are statistically independent events (i.e. events occurring independently of each other).
- Law of Total Probability: if random variable *x* can take values in X and y1,y2,…,yn be mutually exclusive and exhaustive events, then:

  $$P(x)= \Sigma_i P(x \cap yi)$$

  $$P(x)= \Sigma_i P(x/yi)P(yi)$$

Introduction

- The sea bass/salmon example
  - State of nature, prior
    - State of nature is a random variable
    - Let *P(w)* be the probability of occurrence for the pattern w within the population of training samples. This is the *a priori* or prior probability that can be determined without extracting any feature.
    - The catch of salmon w_1 and sea bass w_2 is State of nature
      - $P(w_1) = P(w_2)$ (uniform priors)
      - $P(w_1) + P(w_2) = 1$ (exclusivity and exhaustivity)

- Decision rule with only the prior information
  - Decide w_1 if $P(w_1) > P(w_2)$ otherwise decide w_2

- Use of the class –conditional information
  - x a random variable corresponds to the lightness feature
  - $P(x \mid w_1)$ and $P(x \mid w_2)$ describe the difference in lightness between populations of sea bass and salmon
  - $P(x \mid w_1)$ is the Likelihood probability that the feature value (lightness) is x while the state of nature is w_1

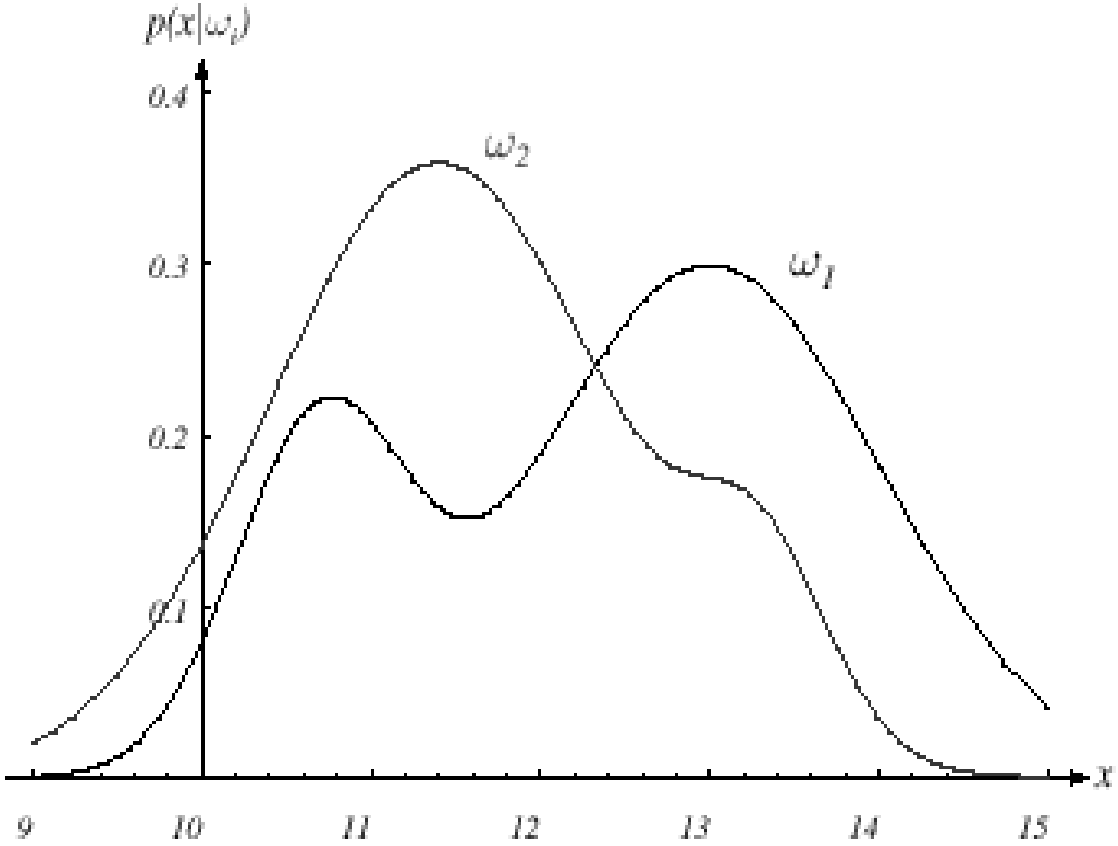

FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_j. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two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

- Decision rule with Likelihoods
  - Decide w_1 if $P(x \mid w_1) > P(x \mid w_2)$ otherwise decide w_2

Bayes' Theorem

- Posterior, likelihood, evidence
  - $P(y \text{ Ç } x) = P(y \mid x)\, P(x) = P(x \mid y)\, P(y) = P(x \text{ Ç } y)$
  - $P(w_j \mid x) = P(x \mid w_j) \,.\, P(w_j) \,/\, P(x)$
  - Where in case of two categories

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

  - Posterior = (Likelihood. Prior) / Evidence
  - Posterior is the probability that the state of nature is w_j while the feature value of x has been measured

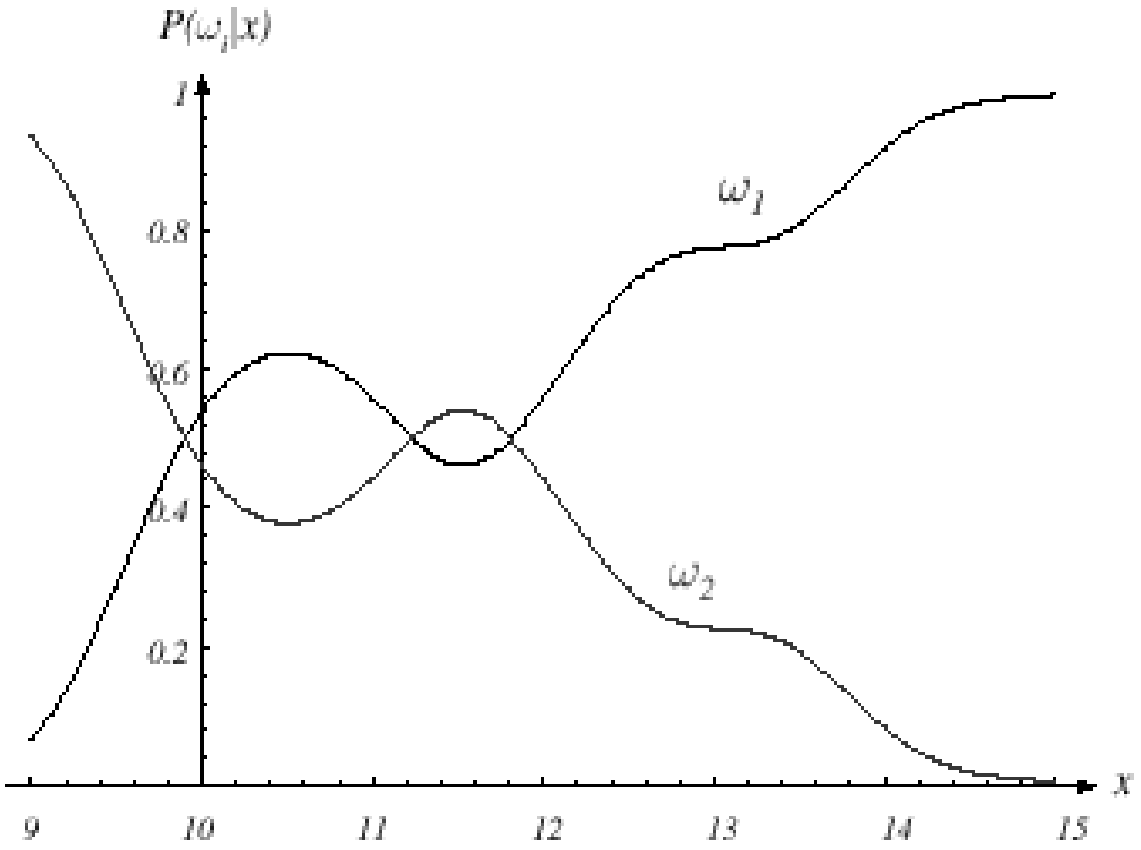

FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 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 $P(w_1 \mid x) > P(w_2 \mid x)$ ⟹ True state of nature = w_1
if $P(w_1 \mid x) < P(w_2 \mid x)$ ⟹ True state of nature = w_2

Therefore:

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

$$P(error \mid x) = P(w_1 \mid x) \text{ if we decide } w_2$$
$$P(error \mid x) = P(w_2 \mid x) \text{ if we decide } w_1$$

- Minimizing the probability of error

- Decide w_1 *if* $P(w_1 \mid x) > P(w_2 \mid x)$;
  otherwise decide w_2

  Therefore:

$$P(error \mid x) = min\ [P(w_1 \mid x), P(w_2 \mid x)]$$

(Bayes decision)

Applying Bayes' Theorem

- Desired information: probability that a set of conditions is at the origin of the observed outcome. Usually hard to get.

  P(Cause | Effect) $\Rightarrow$ P(Disease | Symptoms)

- Available information: probability that the observed outcome is often associated with a given set of conditions. Usually easier to gather.

  P(Effect | Cause) $\Rightarrow$ P(Symptoms | Disease)

- Help from Bayes' theorem:

  P(Cause | Effect) =P(Effect | Cause)P(Cause)/P(Effect) $\Rightarrow$

  P(Disease | Symptoms)=P(Symptoms | Disease)P(Disease)/P(Symptoms)

- **Application:** given a set of symptoms, the most probable disease can be determined based on the known (i.e. prior) probability of the disease and the probability that the disease will be associated with given symptoms.
- Note: probability of the symptoms can be ignored (being the same for any potential disease).

Bayesian Decision Theory – Continuous Features

- Generalization of the preceding ideas
  - Use more than one feature
  - Use more than two states of nature
  - Allowing actions and not only decide on the state of nature
  - 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

Loss Function

- Purpose: provide a mathematical description of the misclassification cost.
  - Are certain classification errors more costly than others? In a detection problem (such as medical, fire alarm), false negatives may be much more costly than false positives.
- Let $\{w_1, w_2, \ldots, w_c\}$ be the set of c states of nature (or "categories").
- Let $\{a_1, a_2, \ldots, a_a\}$ be the set of possible actions.
  - Note: *a* is not necessary equal to c. For instance, one possible action might be not to make a classification decision.
- Let $l(a_i \mid w_j)$ be the loss incurred for taking action a_i when the state of nature is w_j

- $R(a_i \mid x)$ Coditional Risk: That is the loss expected by taking action a_i when the observed evidence (i.e. the feature vector extracted) is x.

$$\boldsymbol{R(a_i \mid x) = \sum_{j=1}^{j=c} l(a_i \mid w_j) P(w_j \mid x)}$$

Select the action α_i *(i = 1,…,a)* for which $R(a_i \mid x)$ is minimum

R is minimum and R in this case is called the
Bayes risk = best performance that can be achieved!

- Two-category classification

  a_1 : deciding w_1

  a_2 : deciding w_2

  $l_{ij} = l(a_i \mid w_j)$:

  loss incurred for deciding w_i when the true state of nature is w_j

  Conditional risk:

$$R(a_1 \mid x) = l_{11}P(w_1 \mid x) + l_{12}P(w_2 \mid x)$$
$$R(a_2 \mid x) = l_{21}P(w_1 \mid x) + l_{22}P(w_2 \mid x)$$

Our rule is the following:

if $R(a_1 \mid x) < R(a_2 \mid x)$

action a_1: "decide w_1" is taken

This results in the equivalent rule :

decide w_1 if:

$$(l_{21} - l_{11})\, P(x \mid w_1)\, P(w_1) > (l_{12} - l_{22})\, P(x \mid w_2)\, P(w_2)$$

and decide w_2 otherwise

Likelihood ratio:

The preceding rule is equivalent to the following rule:

$$\textbf{if } \frac{P(x|w_1)}{P(x|w_2)} > \frac{l_{12} - l_{22}}{l_{21} - l_{11}} \cdot \frac{P(w_2)}{P(w_1)}$$

Then take action a_1 (decide w_1)

Otherwise take action a_2 (decide w_2)

Optimal decision property

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

THOMAS BAYES

- At the time of his death, Thomas Bayes (1702 – 1761) left behind two unpublished essays attempting to determine the probabilities of causes from observed effects. Forwarded to the British Royal Society, the essays had little impact and were soon forgotten.
- When several years later, the French mathematician Laplace independently rediscovered a very similar concept, the English scientists quickly reclaimed the ownership of what is now known as the “Bayes Theorem”.

Exercise

Select the optimal decision where:

$\Omega = \{w_1, w_2\}$

$P(x \mid w_1)$ ⟹ N(2, 0.5) (Normal distribution)

$P(x \mid w_2)$ ⟹ N(1.5, 0.2)

$P(w_1) = 2/3$

$P(w_2) = 1/3$

$$l = \begin{bmatrix} 0 & 3 \\ 4 & 0 \end{bmatrix}$$