Deep Learning for Computer Vision

Lecture 3: Probability, Bayes Theorem, and Bayes Classification

Peter Belhumeur

Computer Science
Columbia University

Probability

Should you play this game?

Game: A fair die is rolled. If the result is 2, 3, or 4, you win \$1; if it is 5, you win \$2; but if it is 1 or 6, you lose \$3.

Random Experiment

a *random* experiment is a process whose outcome is uncertain.

Examples:

Tossing a coin once or several times

Picking a card or cards from a deck

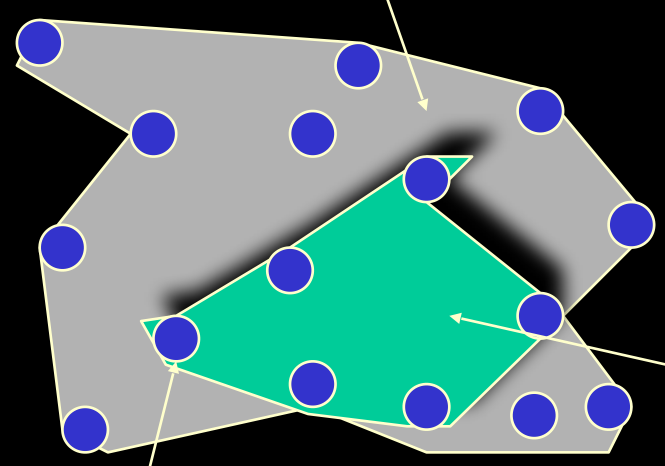
Measuring temperature of patients

...

Events & Sample Spaces

Sample Space

The sample space is the set of all possible outcomes.



Simple Events

The individual outcomes are called simple events.

Event

An event is any collection of one or more simple events.

Example

Experiment: Toss a coin 3 times.

Sample space Ω

Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}.

Examples of events include

A = {HHH, HHT, HTH, THH}
= {at least two heads}

B = {HTT, THT, TTH}
= {exactly two tails}

Basic Concepts (from Set Theory)

The *union* of two events ***A*** and ***B***, ***A*** ∪ ***B***, is the event consisting of all outcomes that are ***either*** in ***A or*** in ***B or*** in both events.

The ***complement*** of an event ***A***, ***A****c*, is the set of all outcomes in Ω that are not in ***A***.

The *intersection* of two events ***A*** and ***B***, ***A*** ∩ ***B***, is the event consisting of all outcomes that are in both events.

When two events ***A*** and ***B*** have no outcomes in common, they are said to be ***mutually exclusive,*** or ***disjoint,*** events.

Example

Experiment: toss a coin 10 times and the number of heads is observed.

Let $A = \{0, 2, 4, 6, 8, 10\}$.

$B = \{1, 3, 5, 7, 9\}$, $C = \{0, 1, 2, 3, 4, 5\}$.

$A \cup B = \{0, 1, \ldots, 10\} = \Omega$.

$A \cap B$ contains no outcomes. So A and B are mutually exclusive.

$C^c = \{6, 7, 8, 9, 10\}$, $A \cap C = \{0, 2, 4\}$.

Rules

Commutative Laws:

$A \cup B = B \cup A, \; A \cap B = B \cap A$

Associative Laws:

$(A \cup B) \cup C = A \cup (B \cup C)$

$(A \cap B) \cap C = A \cap (B \cap C)$

Distributive Laws:

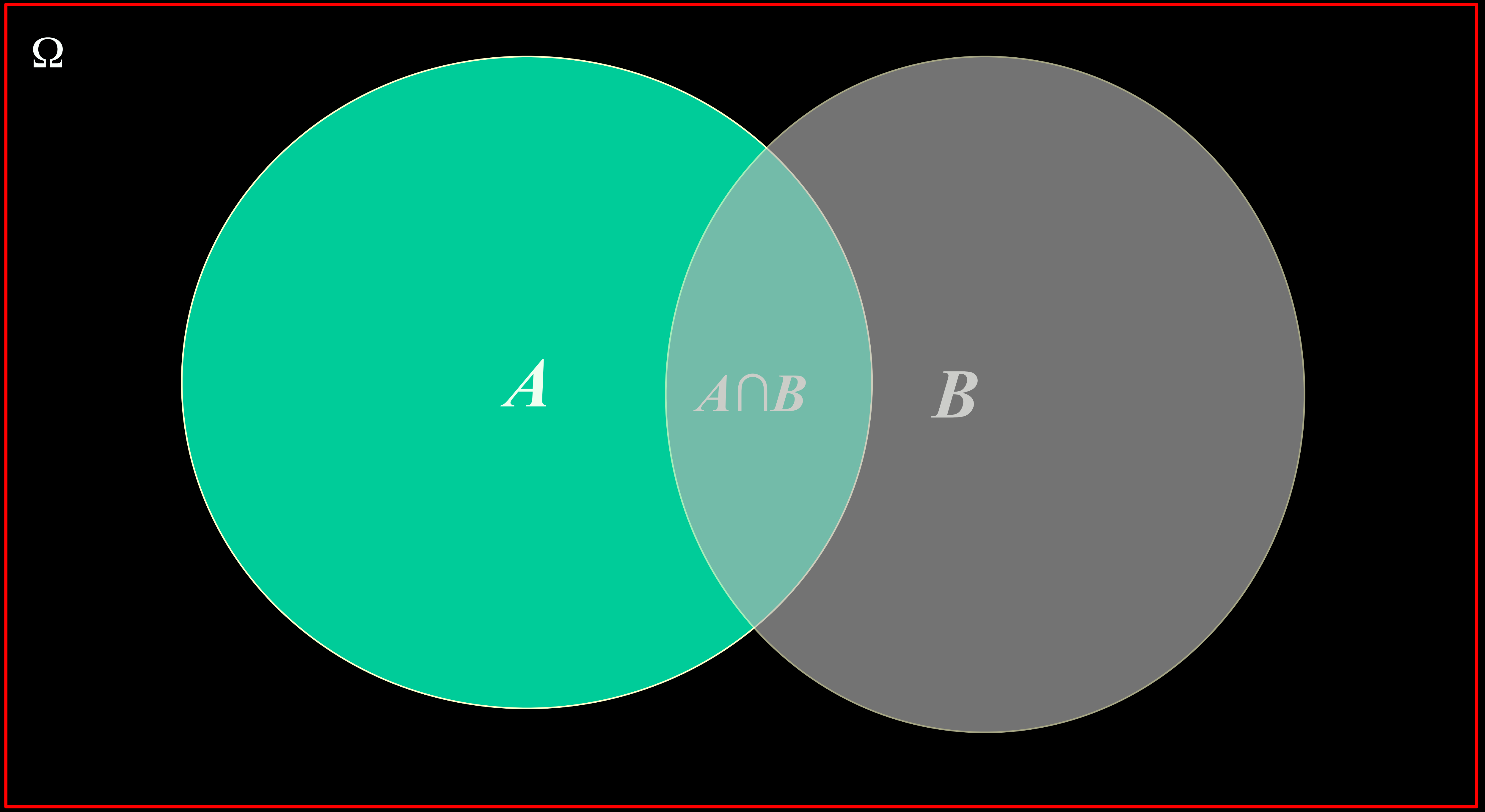
$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

DeMorgan's Laws:

$$\left(\bigcup_{i=1}^{n} A_i\right)^c = \bigcap_{i=1}^{n} A_i^c, \quad \left(\bigcap_{i=1}^{n} A_i\right)^c = \bigcup_{i=1}^{n} A_i^c.$$

Venn Diagram



A **probability** is a number assigned to each subset (events) of a sample space Ω that satifies the following rules.

Axioms of Probability

- For any event A, $0 \leq P(A) \leq 1$.

- $P(\Omega) = 1$.

- If $A_1, A_2, \ldots A_n$ is a partition of A, then

$$P(A) = P(A_1) + P(A_2) + ... + P(A_n)$$

($A_1, A_2, \ldots A_n$ is called a partition of A if $A_1 \cup A_2 \cup \ldots \cup A_n = A$ and $A_1, A_2, \ldots A_n$ are mutually exclusive.)

Properties of Probability

- For any event A, $P(A^c) = 1 - P(A)$.
- If $A \subset B$, then $P(A) \leq P(B)$.
- For any two events A and B,

  $$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

For three events, A, B, and C,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) -$$

$$P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

Frequentist View of Probability

The probability of an event **a** could be defined as:

$$P(a) = \lim_{n \to \infty} \frac{N(a)}{n}$$

Where N(a) is the number that event a happens in n trials

Here We Go Again: Not So Basic Probability

Bring on the Notation

Let Ω be the sample space, ω in Ω be a single outcome, A in Ω a set of outcomes of interest, then

1. $P(a) \geq 0 \, \forall A \in \Omega$
2. $P(\Omega) = 1$
3. $A_i \cap A_j = \emptyset \; i,j \implies P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$
4. $P(\emptyset) = 0$

Independence

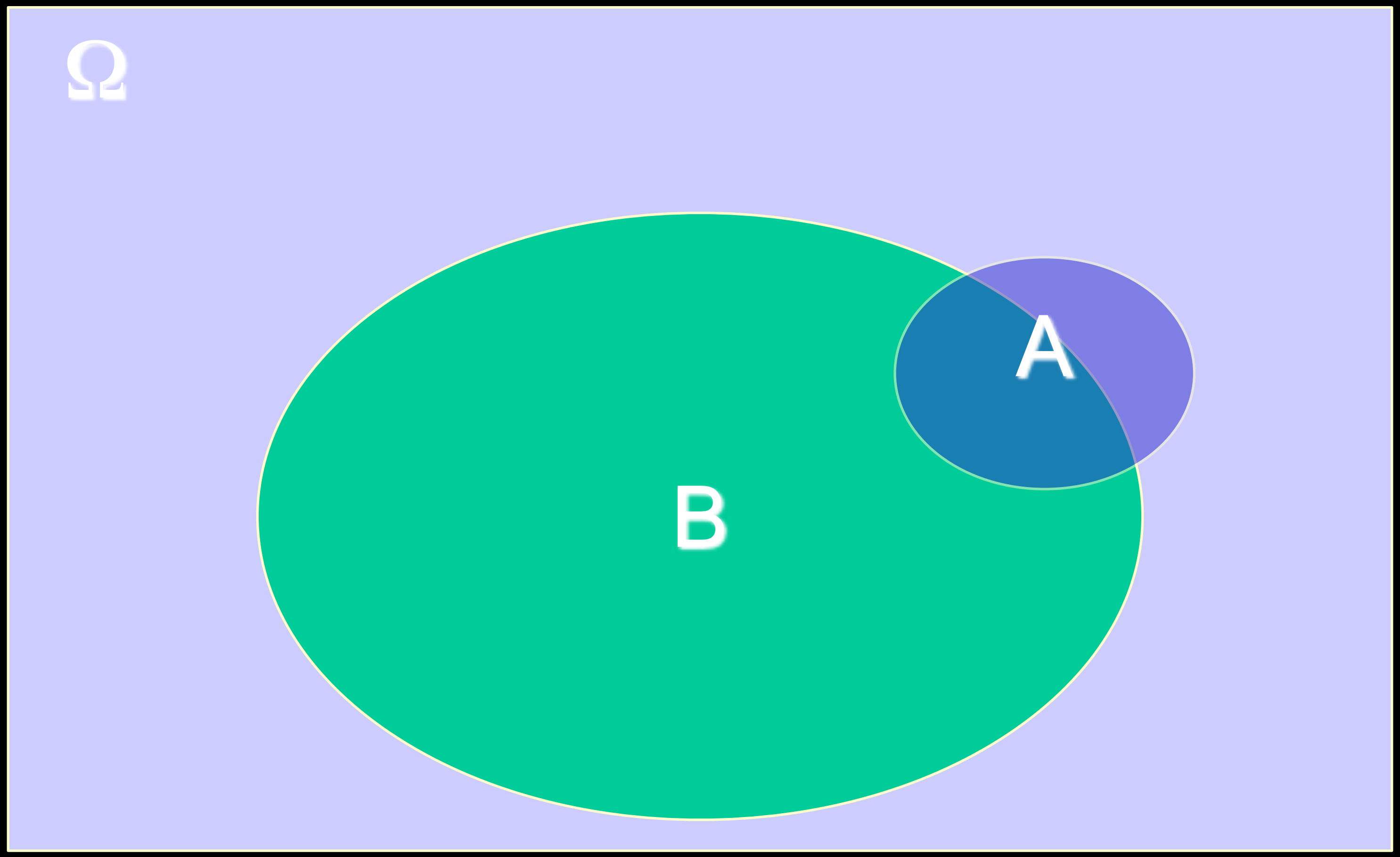
The probability of independent events A, B and C is given by:

$$P(A,B,C) = P(A)P(B)P(C)$$

A and B are independent, if knowing that A has happened does not say anything about B happening

Conditional Probability

We say “probabilty of A given B” to mean the probability of event A given that event B occurs.



Conditional Probability

So “probabilty of A given B” is the probability that both event A and B occur normalized by the probability of event B.

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

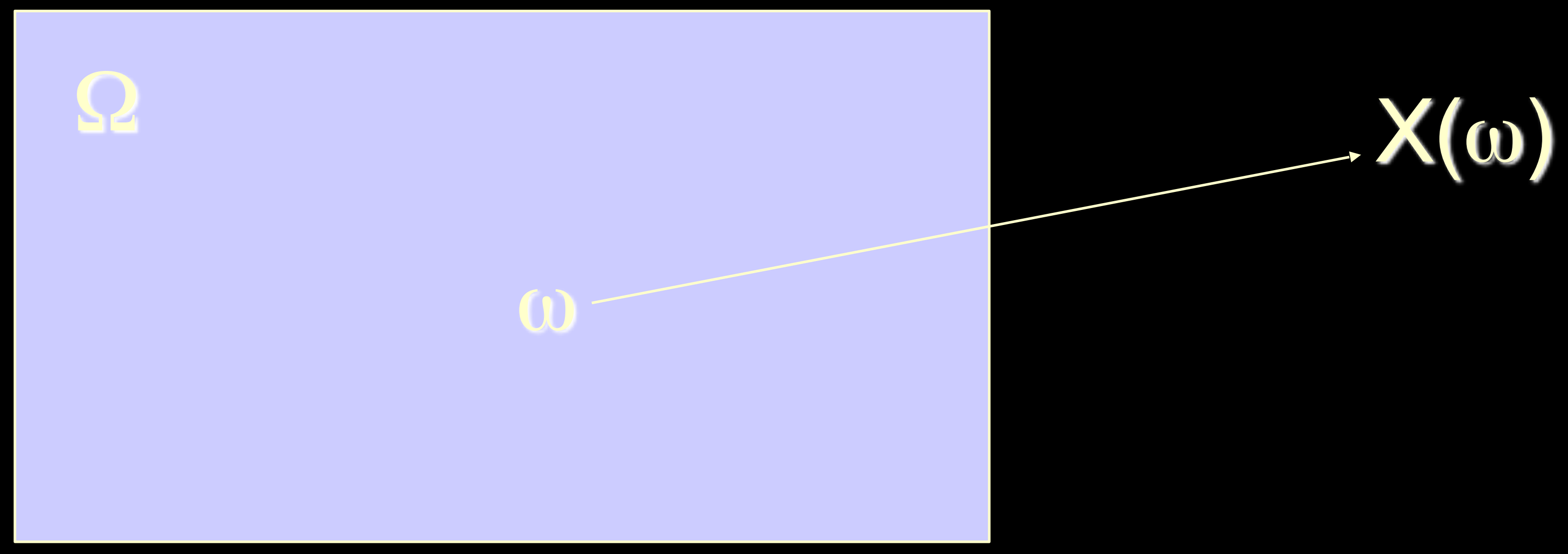
Bayes Theorem

Provides a way to convert *a-priori* probabilities to *a-posteriori* probabilities:

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

Random Variables

A (scalar) random variable X is a function that maps the outcome of a random event into real scalar values



Random Variable's Distributions

Cumulative Probability Distribution (CDF):

$$F_X(x) = P(X \leq x)$$

Probability Density Function (PDF):

$$p_X(x) = \frac{dF_X(x)}{dx}$$

The PDF integrates to 1

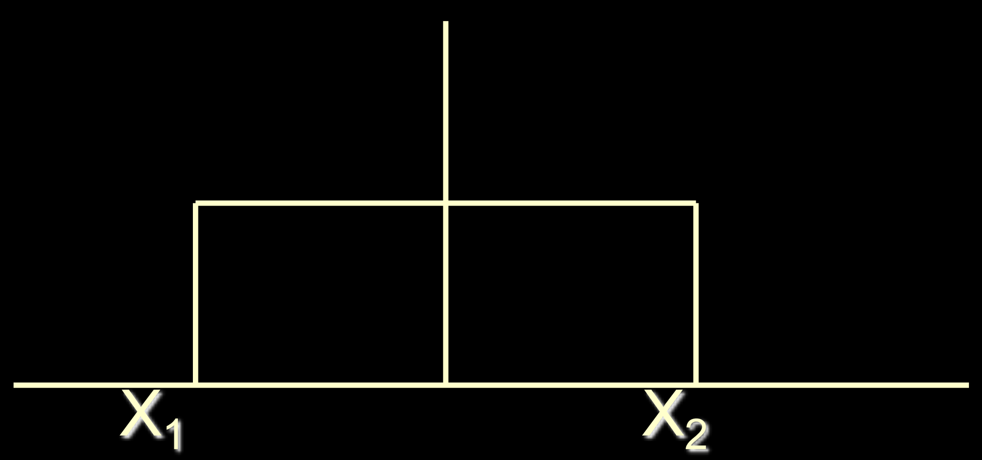
So as you would expect:

$$\int_{-\infty}^{\infty} p_X(x)dx = 1.0$$

Uniform Distribution

A R.V. X that is uniformly distributed between x_1 and x_2 has density function:

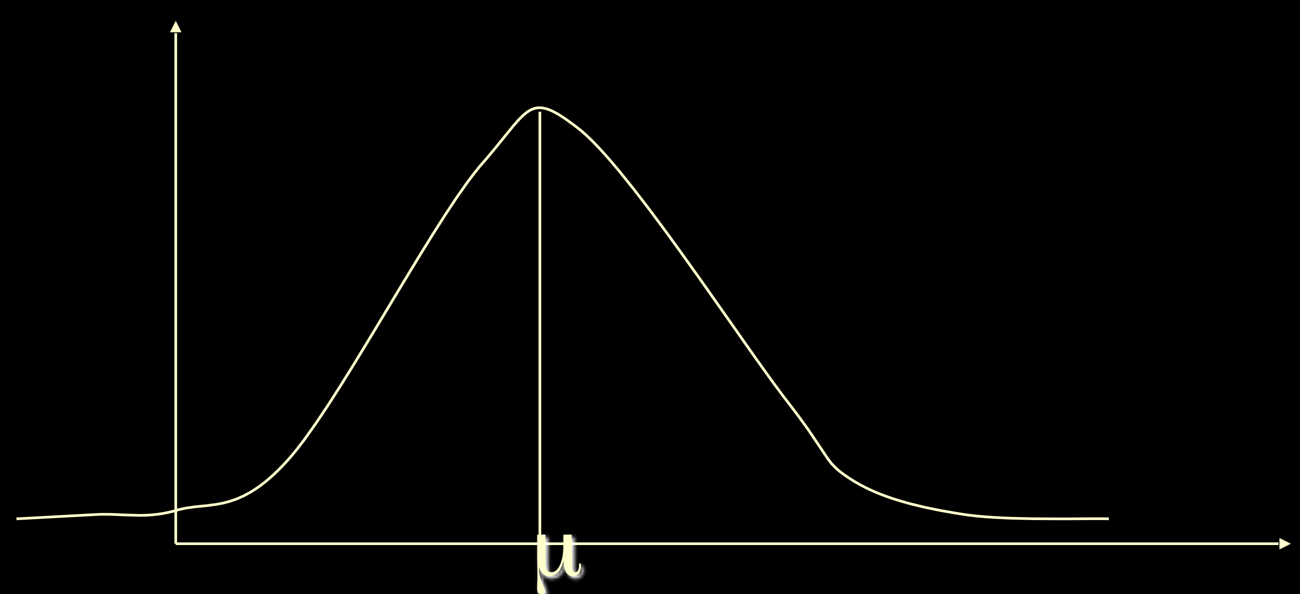
$$
\begin{aligned}
p_X(x) &= \frac{1}{x_2 - x_1} \quad & x_1 \leq x \leq x_2 \\
&= 0 & otherwise
\end{aligned}
$$



Gaussian (Normal) Distribution

A R.V. X that is normally distributed has density function:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



Simple Statistics

Expectation (Mean or First Moment):

$$E(X) = \int_{-\infty}^{\infty} x\, p(x)\, dx$$

Second Moment:

$$E(X^2) = \int_{-\infty}^{\infty} x^2\, p(x)\, dx$$

Simple Statistics

Variance of X:

$$\begin{aligned} Var(X) &= E[(X - E[X])^2] \\ &= \int_{-\infty}^{\infty} (x - E[X])^2 \, p(x) \, dx \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

Standard Deviation of X:

$$Std(X) = \sqrt{Var(X)}$$

Sample Mean

Given a set of N samples from a distribution, we can estimate the mean of the distribution by:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Sample Variance

Given a set of N samples from a distribution, we can estimate the variance of the distribution by:

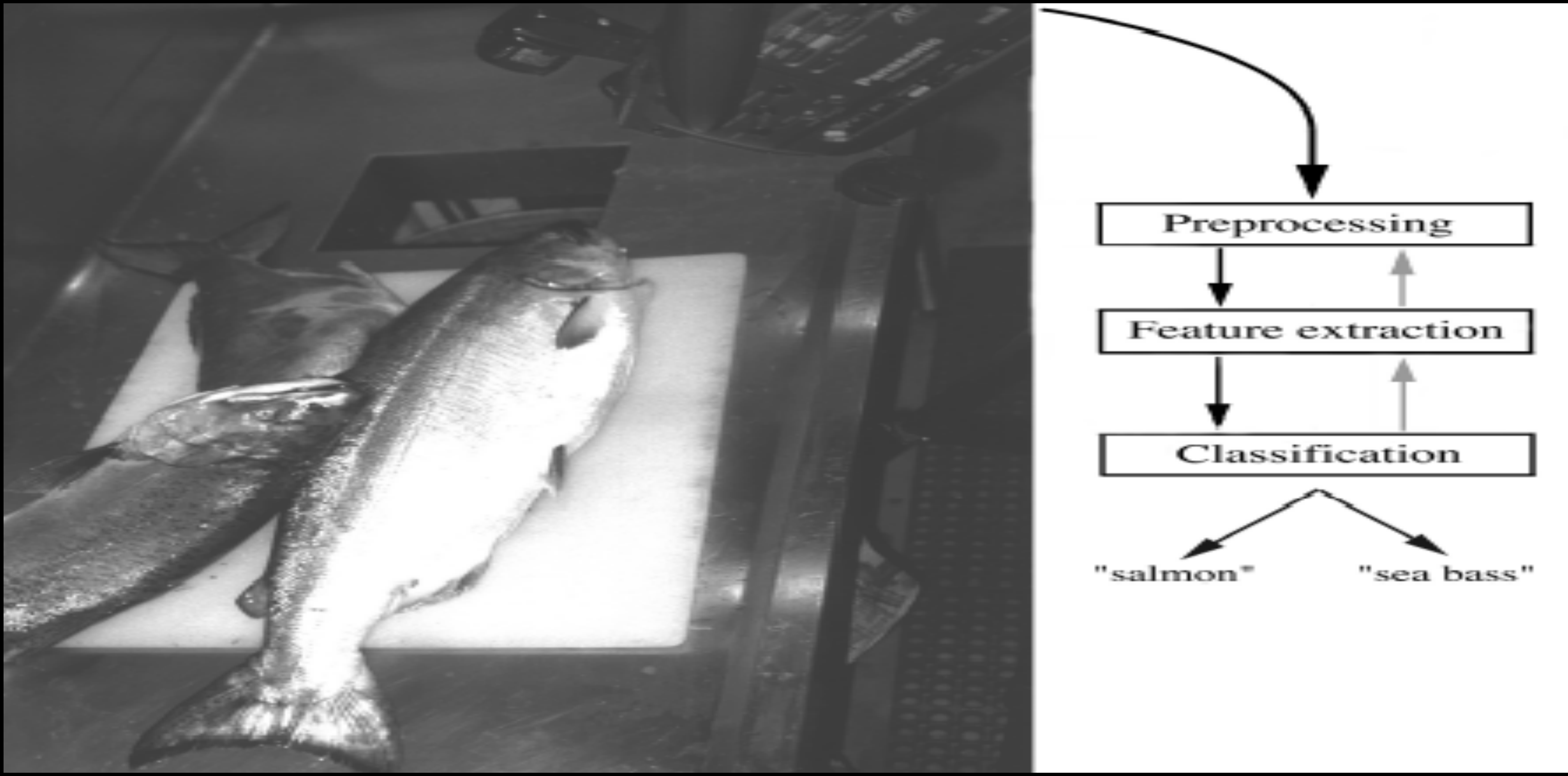
$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2$$

Bayesian Classifiers

Classification: An Example

Classify fish species at an Alaskan Canning Factory

Species
- Sea bass
- Salmon


Preprocessing
Feature extraction
Classification
"salmon"
"sea bass"

Priors

The sea bass/salmon example:

Let ω_1 be the state or “class” that the fish is a salmon

Let ω_2 be the state or “class” that the fish is a sea bass

Let $P(\omega_1)$ be the prior probability that a fish is salmon

Let $P(\omega_2)$ be the prior probability that a fish is sea bass

$P(\omega_1) + P(\omega_2) = 1$ (no other species are possible)

Dumb Classifier

Decision rule with only the prior information:

Decide ω_1 if $P(\omega_1) > P(\omega_2)$ otherwise decide ω_2

This does not use any of the class–conditional information or “features”

Our features are “lightness” and the width of the fish

Fish $\longrightarrow$ $x^T = [x_1, x_2]$

Lightness **Width**

How should we use our “features”?

Minimum Error Rate Classifier

Probability of error given $\boldsymbol{x}$:

$$P(error \mid x) = \min [P(\omega_1 \mid x), P(\omega_2 \mid x)]$$

Minimizing the probability of error:

Decide ω_1 *if* $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise decide ω_2

How do we compute $P(\omega_i \mid x)$?

Bayes Theorem

$$P(\omega_i|x) = \frac{\rho(x|\omega_i)P(\omega_i)}{P(x)}$$

$$= \frac{\rho(x|\omega_i)P(\omega_i)}{\sum_i \rho(x|\omega_i)P(\omega_i)}$$

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

Likelihood (Class-conditional Density)

Need the class–conditional information:

$$p(x \mid \omega_1) \text{ and } p(x \mid \omega_2)$$

describe the difference in “lightness” between populations of sea-bass and salmon

These are also known as *likelihood* functions.

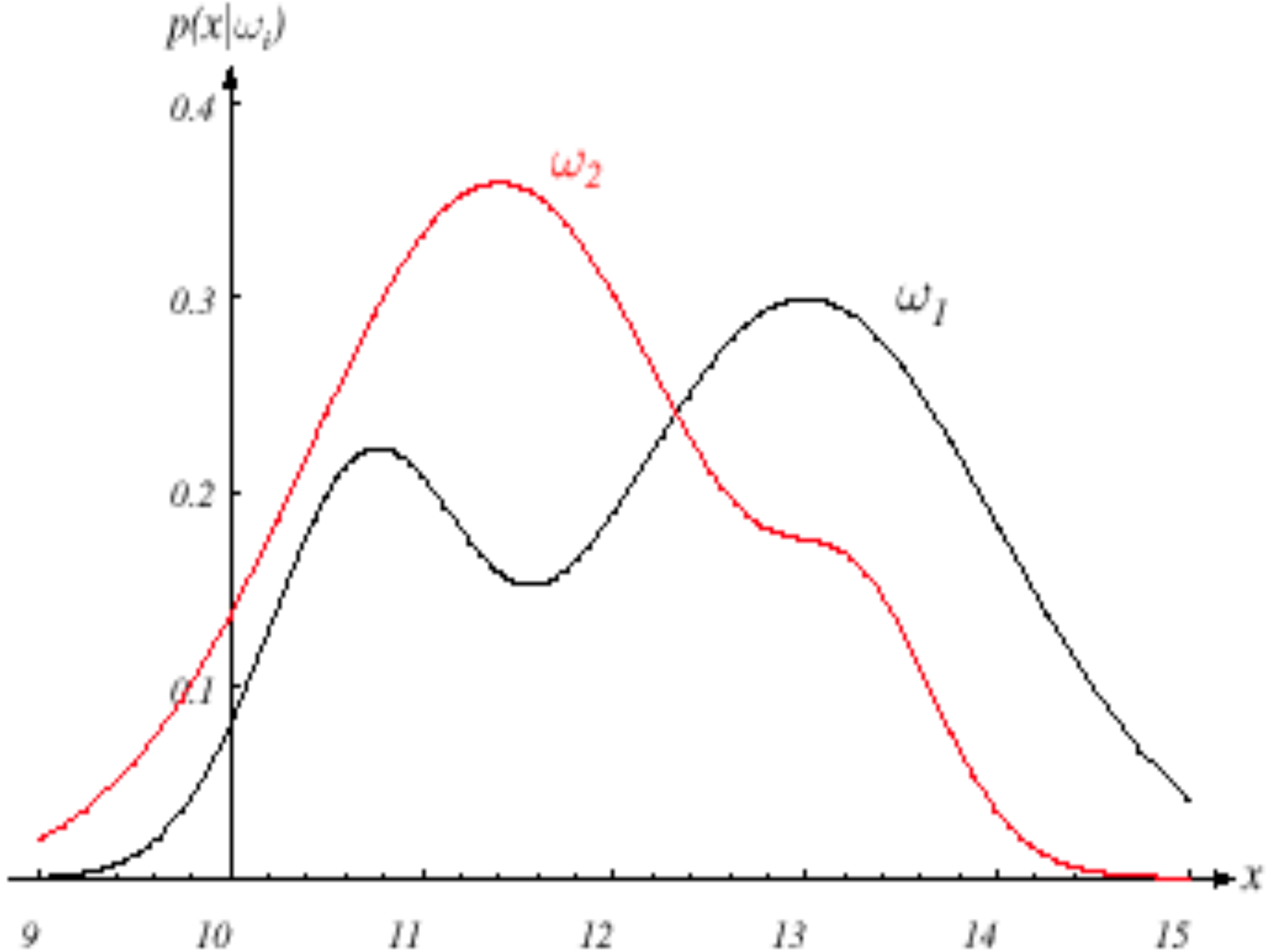


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two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

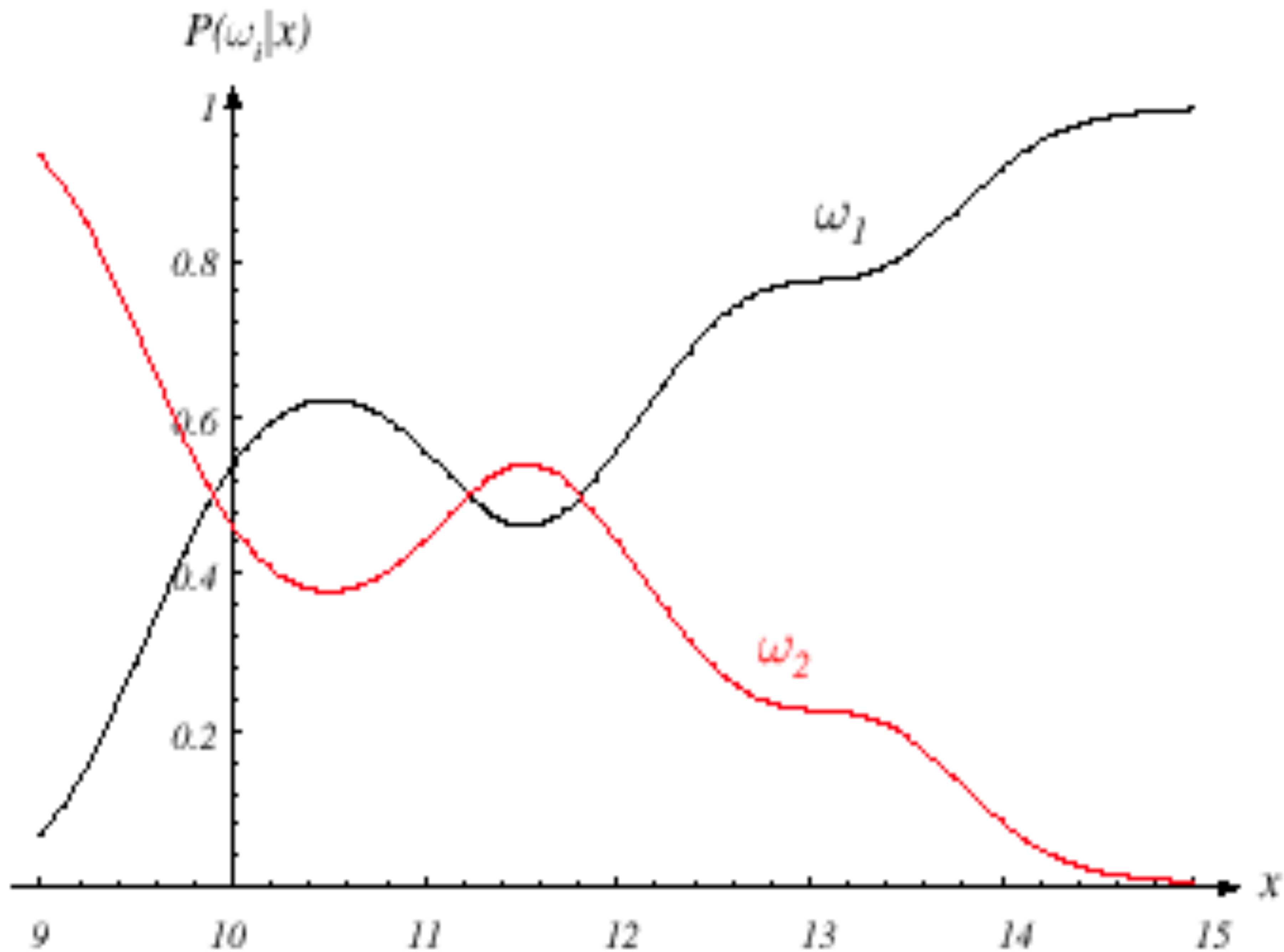


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly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

If our feature space is one dimensional then the “boundary” that separates the area assigned to one class vs. another class is a point.

But what happens as the dimensionality of our feature space increases?

Let's think a **classifier** as set of scalar functions $g_i(\mathbf{x})$ — one for each class i — that assigns a score to the vector of feature values $\mathbf{x}$ and then choses the class i with the highest score.

So a **classifier** uses the following decision rule:

Choose class i if $g_i(\mathbf{x}) > g_j(\mathbf{x}) \; \forall \, j, \; i \neq j$

So our Bayesian classifier assigns a score based on the *a posteriori* probabilities:

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

So if our feature space is n-dimensional, i.e., $\mathbf{x} \in \mathbb{R}^n$, then the boundaries separating regions that our classifier assigns to the same class is n-1 dimensional surface.

This is something that I find harder to imagine. For example, if the feature space was two dimensional, we claim that a line will separate it and if the feature space was three dimensional, a plane will separate it into classes. While that does make sense, I find it hard to imagine with the class conditional probabilities. Hmm.