Creative Computing for Engineers

Lecture 9: Bayesian Approach to Data Analysis

Machine Learning: Approaches

1) Deterministic:

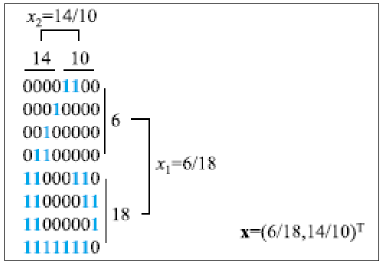
- All variables/observables are treated as certain/exact
- Example: Digit recognition
  - Find/fit a function f(X) on an image X
  - which = 0 or 1 depending on contents
  - Class label given by y= f(X)

2) Probabilistic/Bayesian/Stochastic:

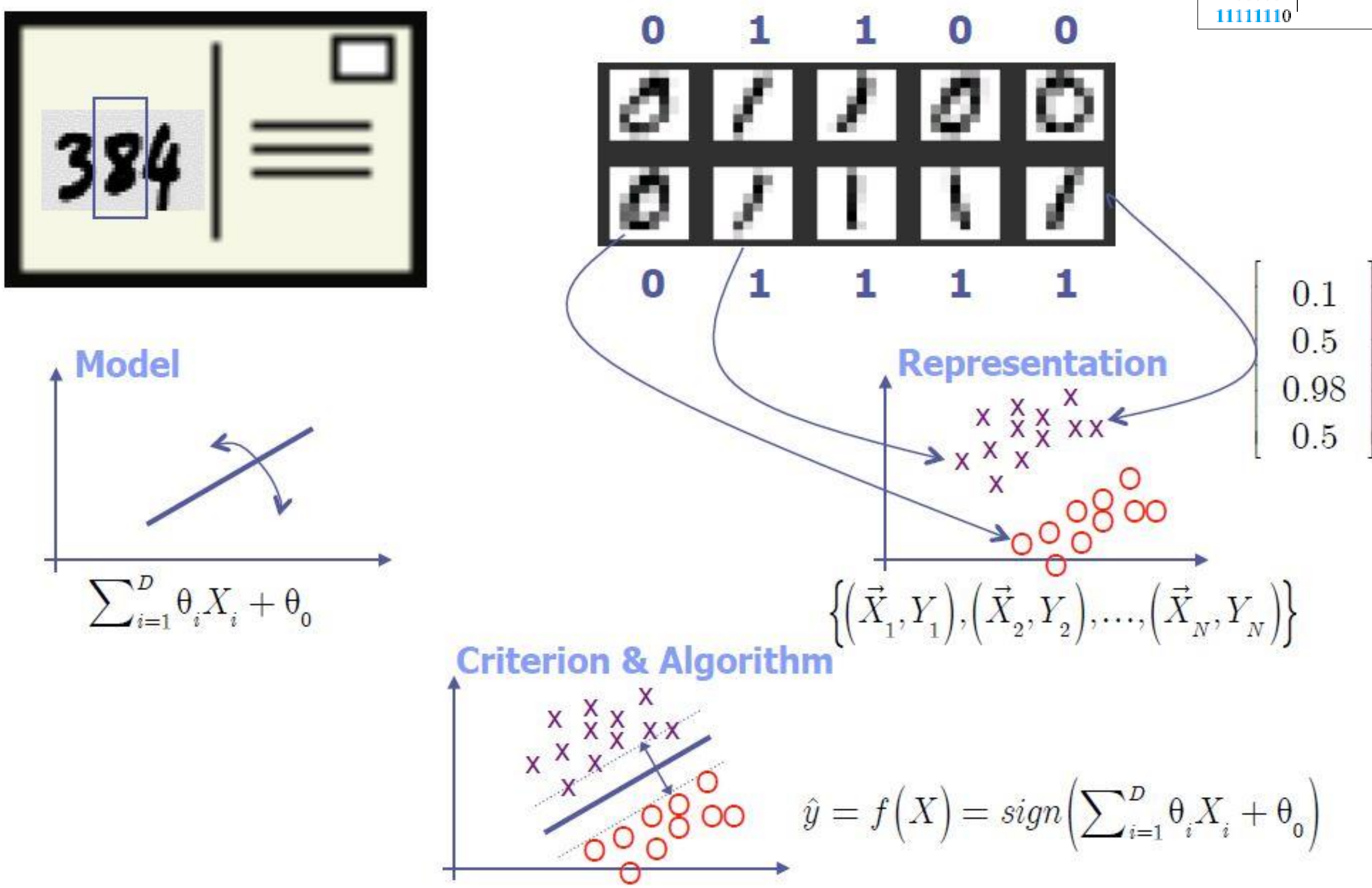
- Variables/observables are random (R.V.) and uncertain
- Example: Digit recognition
  - Probability that image is a '0' digit: $p(y=0|X) = 0.43$
  - Probability that image is a '1' digit: $p(y=1|X) = 0.57$
  - Class label given by: $p(y=0|\text{image})$ and $p(y=1|\text{image})$

Source: Tony Jebara. Machine Learning Lecture Note, Columbia University

Machine Learning: Approaches



1) Deterministic Approach

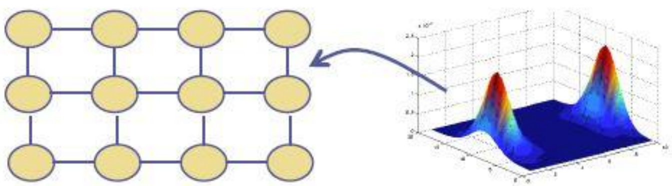


Source: Tony Jebara. Machine Learning Lecture Note, Columbia University

Machine Learning: Approaches

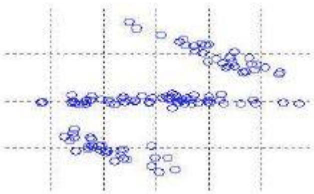
2) Probabilistic/Bayesian/Stochastic Approach

a) Provide Prior Model
Parameters & Structure



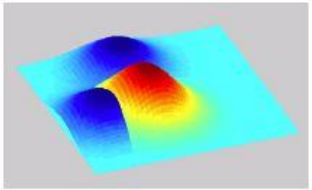
b) Obtain Data / Labels
Past experience

$$\left\{\left(X_1,Y_1\right),\ldots,\left(X_T,Y_T\right)\right\}$$



c) Learn/Refine model with data
p(all system vars)

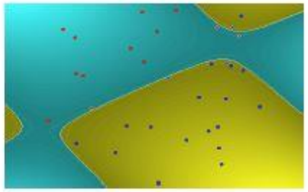
$$p\left(X,Y\right)$$



d) Use model for inference (classify/predict)

Probability image is '0': p(y=0|X)
Probability image is '1': p(y=1|X)
Output: p(y=0|X) <> p(y=1|X)

$$p\left(Y \mid X\right)$$



Source: Tony Jebara. Machine Learning Lecture Note, Columbia University

Basic Probability

Probability

: Probability is the study of randomness and uncertainty

- In the early days, probability was associated with games of chance (gambling)

Example: Simple games involving probability

A fair die is rolled.

- If the result is 2, 3, or 4, you win \$1.
- If the result is 5, you win \$2.
- If the result is 1 or 6, you lose \$3.

Should I play this game? (expected value?)

Random Experiment

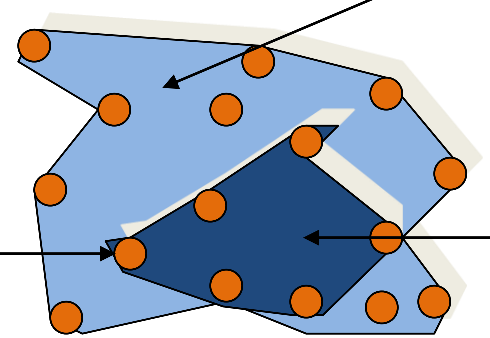
- A random experiment is a process whose outcome is uncertain.
- Examples:
  - Tossing a coin once or several times
  - Picking a card or cards from a deck
  - Measuring temperature of patients

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Basic Probability

Probability

Events and 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 = {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}

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Basic Probability

Basic Concepts: Set Theory

- The *union* of two events *A* and *B*, $A \cup B$, 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*, $A \cap B$, 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: Let A = { 0, 2, 4, 6, 8, 10}, *B* = { 1, 3, 5, 7, 9}, and *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.
- $C^c = \{6, 7, 8, 9, 10\}$, $A \cap C = \{0, 2, 4\}$

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Basic Probability

Basic Rules

- Commutative Laws:

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

- Associative Laws:

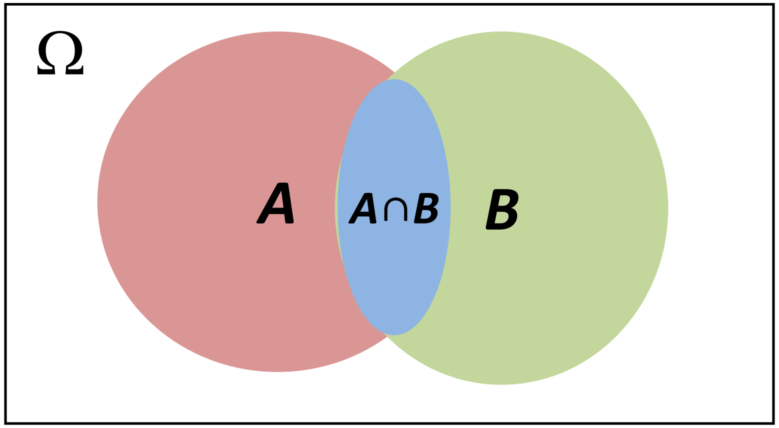
$$(A \cup B) \cup C = A \cup (B \cup C)$$
$$(A \cap B) \cap C = A \cap (B \cap C)$$

- 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:

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



Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Basic Probability

Probability

- A probability is a number assigned to each subset (events) of a sample space Ω.
- Probability distributions satisfy the following rules:

[*Axioms of Probability*]

- For any event A, $0 \leq P(A) \leq 1$
- $P(\Omega) = 1$
- If A1, A2, … An is a partition of A, then
  P(A) = P(A1) + P(A2) + ...+ P(An)

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

(A1, A2, … An is called a partition of A if A1 $\cup$ A2 $\cup$ …$\cup$ An = A and A1, A2,… An 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:
  $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Basic Probability

Probability

Intuitive Development (agrees with axioms)

- Intuitively, 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

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)

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Basic Probability

Bayes Theorem

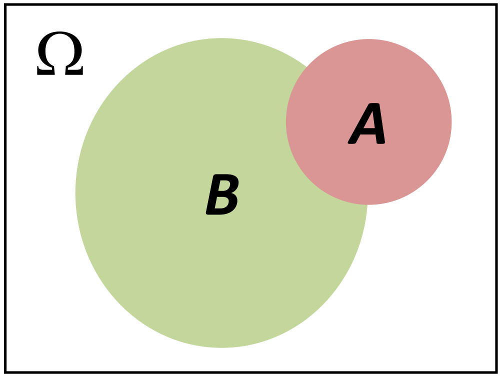
- Provides a way to convert *“a priori”* probabilities to *“a posteriori”* probabilities:

$$P(A|B)P(B) = P(B|A)P(A) = P(A \cap B)$$

Conditional Probability

- One of the most useful concept!

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



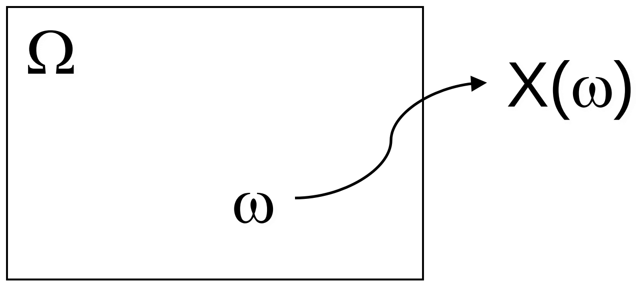
Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Basic Probability

Random Variables

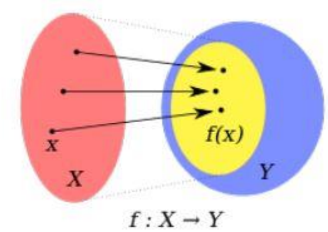
- A (scalar) random variable X is a function that maps the outcome of a random event into real scalar values (i.e., E: measurable space).

$X: \Omega \rightarrow E$



Sample Space

Example



f: function,
X: domain, Y: Range

Example:

P(X < 3) is the measure of the set outcomes $\{\omega \in \Omega: X(\omega) < 3\}$

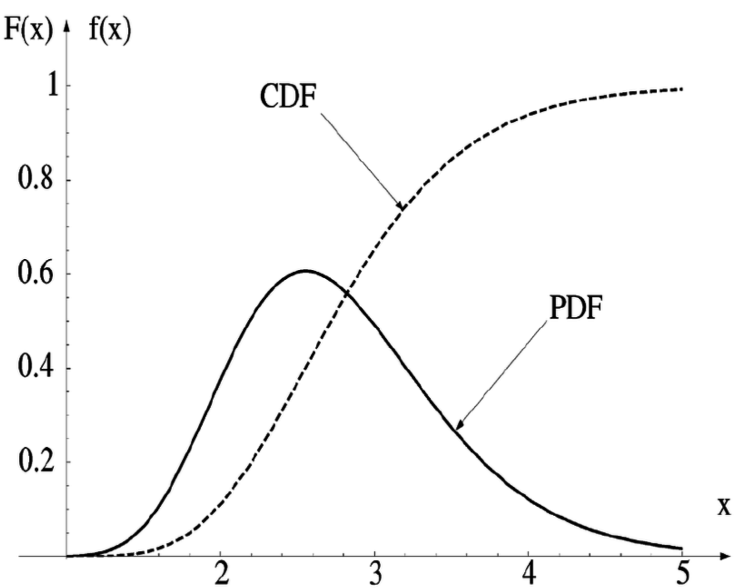
Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Basic Probability

Random Variables

Cumulative Probability Distribution (CDF): $$F_X(x) = P(X \le x)$$

Probability Density Function (PDF): $$p_X(x) = \frac{dF_X(x)}{dx}$$



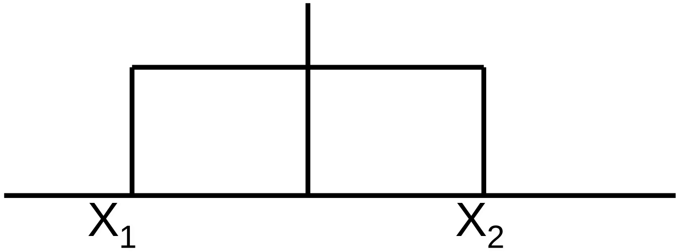
Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Distribution

Uniform Distribution

- A Random Variable X that is uniformly distributed between x_1 and x_2 has density function:

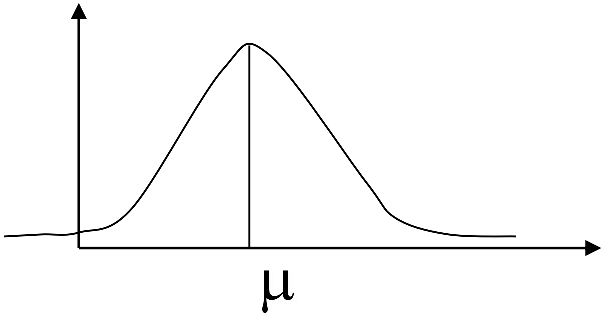
$$p_X(x) = \begin{cases} \frac{1}{x_2 - x_1} & x_1 \leq x \leq x_2 \\ 0 & otherwise \end{cases}$$



Gaussian (Normal) Distribution

- A Random Variable X that is normally distributed has density function:

$$p_X(x) = \frac{1}{2\pi\sigma} \exp - \frac{(x-\mu)^2}{2\sigma^2}$$



Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Basic Probability

Statistical Characterization

Expectation (Mean value, First Moment):

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

Second Moment:

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

Variance of X:

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

Standard Deviation of X: $\sigma_X = \sqrt{Var(X)}$

Mean

Continuous	$\boldsymbol{\mu} = \int_{R^d} \mathbf{x} p(\mathbf{x}) d\mathbf{x}$
Discrete	$\boldsymbol{\mu} = \sum_{\mathbf{x}} \mathbf{x} P(\mathbf{x})$

Covariance

Continuous	$\boldsymbol{\Sigma} = \int_{R^d} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T p(\mathbf{x}) d\mathbf{x}$
Discrete	$\boldsymbol{\Sigma} = \sum_{\mathbf{x}} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T P(\mathbf{x})$

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Basic Probability

Statistical Characterization

Mean Estimation from Samples

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

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

Variance Estimation from Samples

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

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

Covariance

Discrete	$\mathbf{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}}$

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

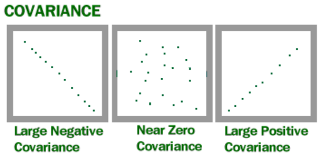
Basic Probability

Statistical Characterization

Example: Samples given

Student	$X = (x_1, x_2, x_3)^T$ (x_1: height, x_2: weight, x_3: grade)
1	$X1 = (170, 60, 4.1)^T$
2	$X2 = (165, 55, 3.0)^T$
3	$X3 = (174, 75, 2.8)^T$
4	$X4 = (169, 67, 2.9)^T$
5	$X5 = (155, 49, 3.1)^T$
6	$X6 = (172, 63, 3.6)^T$
7	$X7 = (166, 58, 3.7)^T$
8	$X8 = (168, 61, 4.0)^T$

Covariance: A measure of how much two random variables vary together.



http://www.statisticshowto.com/wp-content/uploads/2013/12/g-covariance.gif

$$\boldsymbol{\mu} = (167.375,\ 61.0,\ 3.4)^T$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 33.696 & 39.429 & 0.371 \\ 39.429 & 60.857 & -0.943 \\ 0.371 & -0.943 & 0.263 \end{pmatrix}$$

Example: Covariance $\boldsymbol{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T$

X1 (i=1)

$$(\mathbf{x}_1 - \boldsymbol{\mu})(\mathbf{x}_1 - \boldsymbol{\mu})^T = \begin{pmatrix} 170-167.375 \\ 60-61.0 \\ 4.1-3.4 \end{pmatrix}\begin{pmatrix} 170-167.375 & 60-61.0 & 4.1-3.4 \end{pmatrix}$$

$$= \begin{pmatrix} 6.891 & -2.625 & 1.838 \\ -2.625 & 1.0 & -0.7 \\ 1.838 & -0.7 & 0.49 \end{pmatrix}$$

➔ (1/7) * sum of (X1 to X8 cases)

Source: Oh Il-Seok (2008). Pattern Recognition.

Classification (re-visited)

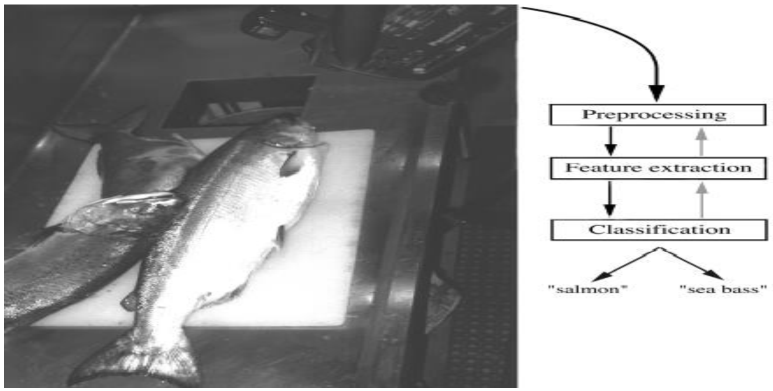
What is Classification in Machin Learning?

- Build a machine that can recognize patterns

Example: Sorting incoming Fish on a conveyor according to species using optical sensing

Species → Sea bass

Species → Salmon



- Set up a camera and take sample images to extract features:
  - Length
  - Lightness
  - Width
  - Number and shape of fins
  - Position of the mouth
  - Etc.

➔ This is the set of all suggested features to explore for use in our classifier

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

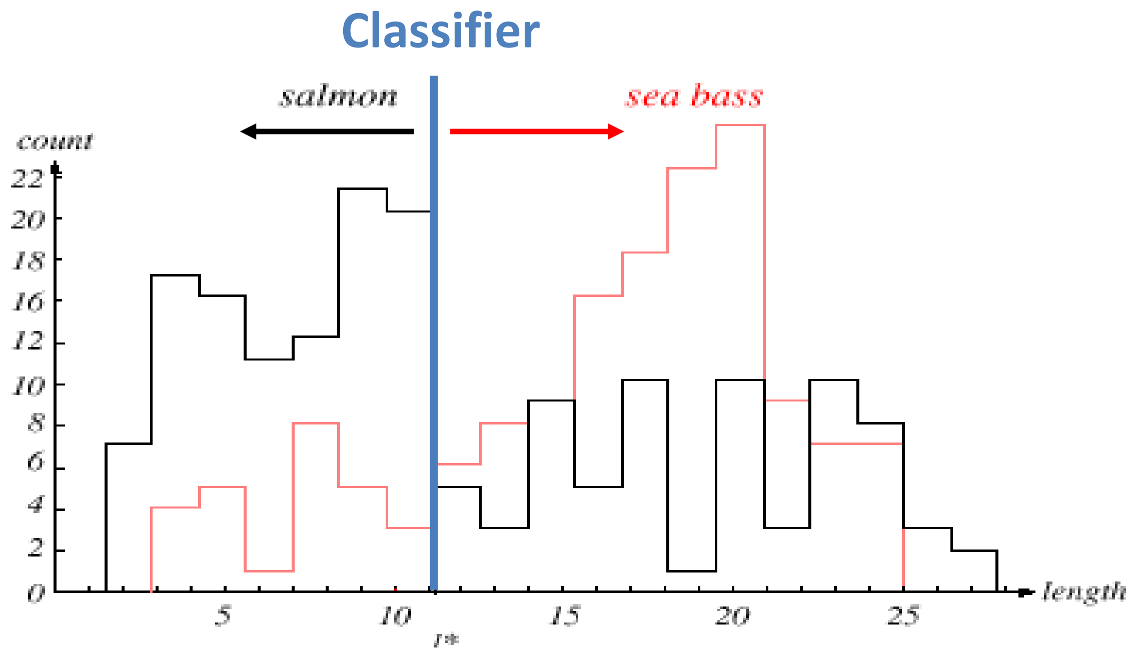
Classification (re-visited)

What is Classification in Machin Learning?

Example: Sorting incoming Fish on a conveyor according to species using optical sensing

Classification

- Select the <u>length</u> of the fish as a possible feature for discrimination



Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

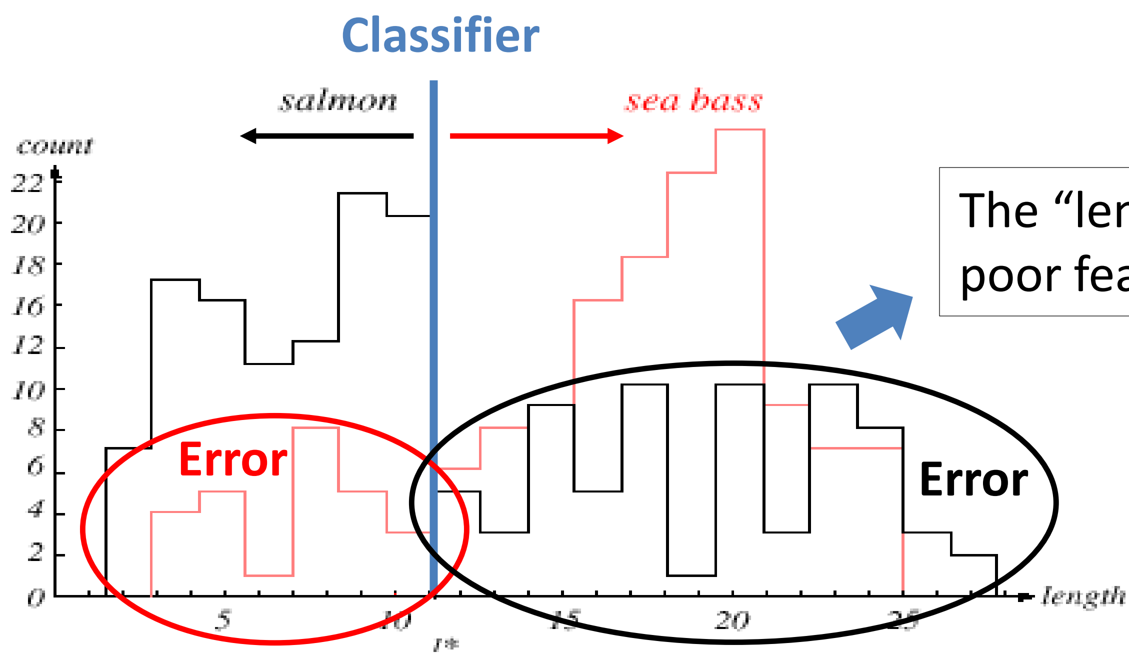
Classification (re-visited)

What is Classification in Machin Learning?

Example: Sorting incoming Fish on a conveyor according to species using optical sensing

Classification

- Select the <u>length</u> of the fish as a possible feature for discrimination



The "length" alone is a poor feature.

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

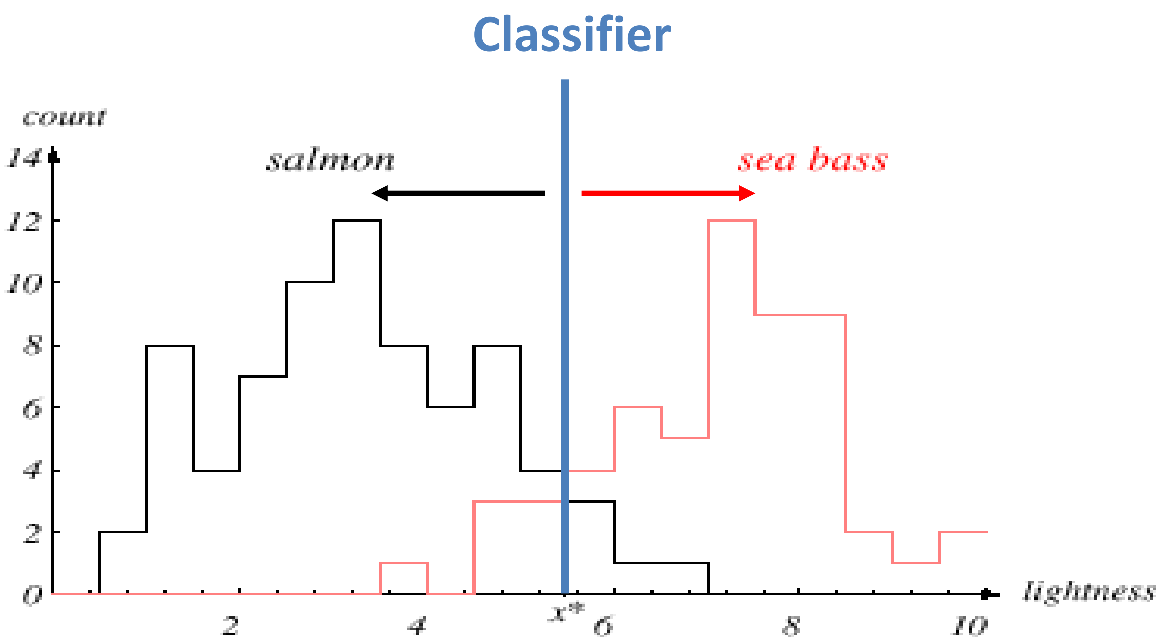
Classification (re-visited)

What is Classification in Machin Learning?

Example: Sorting incoming Fish on a conveyor according to species using optical sensing

Classification

- Select the <u>lightness</u> of the fish as a possible feature for discrimination



Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Classification (re-visited)

What is Classification in Machin Learning?

Example: Sorting incoming Fish on a conveyor according to species using optical sensing

Classification

- Select "good" feature(s) for discrimination
  - Length? Lightness? Or {Length, Lightness}? Anything else?

- Threshold decision boundary and cost relationship
  - Move our decision boundary toward smaller values of lightness in order to reduce the number of sea basses that are classified as a salmon (assuming this can minimize the cost).

  ➔ Task of decision theory

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

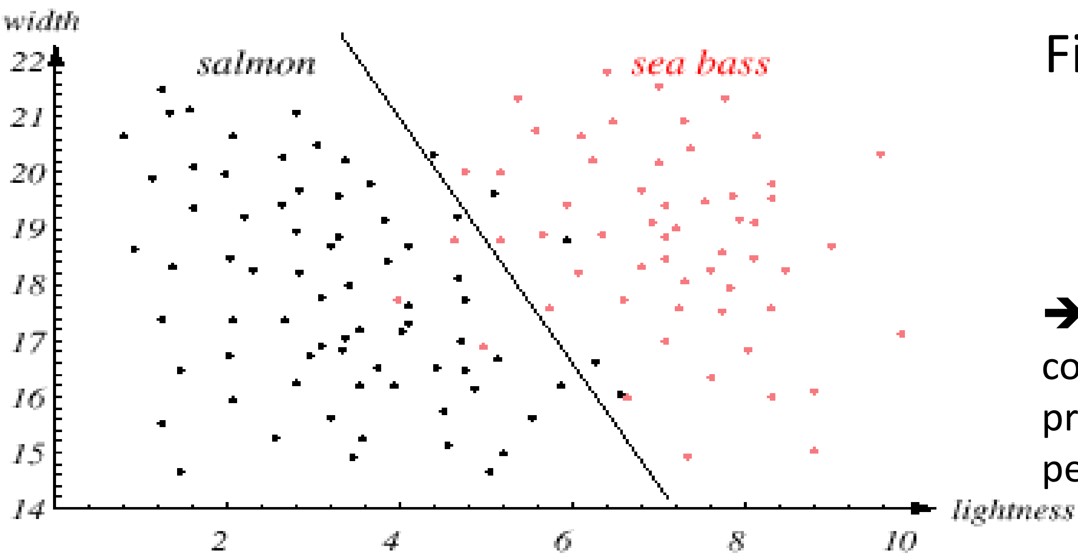
Classification (re-visited)

What is Classification in Machin Learning?

Example: Sorting incoming Fish on a conveyor according to species using optical sensing

Classification

- Select “good” feature(s) for discrimination
  - Adopt the lightness and add the width of the fish



Fish ➔ $x^T = [x_1, x_2]$

➔ We might add other features that are not correlated with the ones we already have. A precaution should be taken not to reduce the performance by adding “noisy features”

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Classification (re-visited)

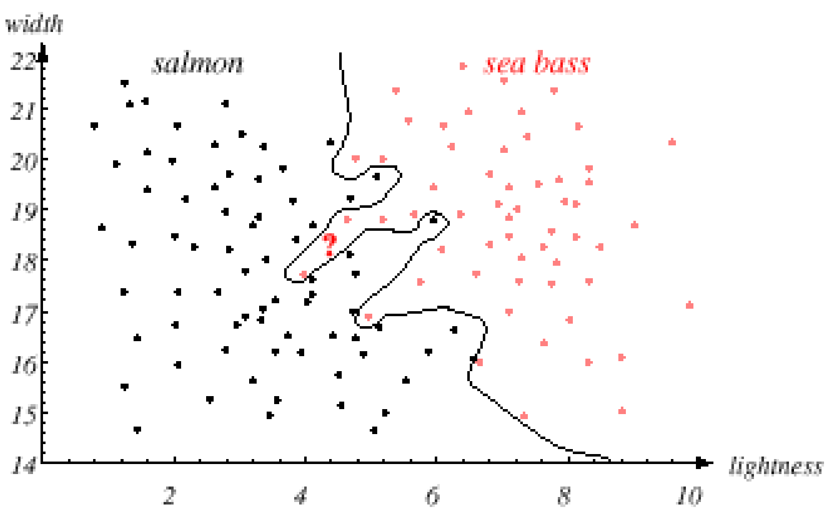
What is Classification in Machin Learning?

Example: Sorting incoming Fish on a conveyor according to species using optical sensing

Classification

- Threshold decision boundary and cost relationship

  ➔ Ideally, the best decision boundary should be the one which provides an optimal performance such as in the following figure:



Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Classification (re-visited)

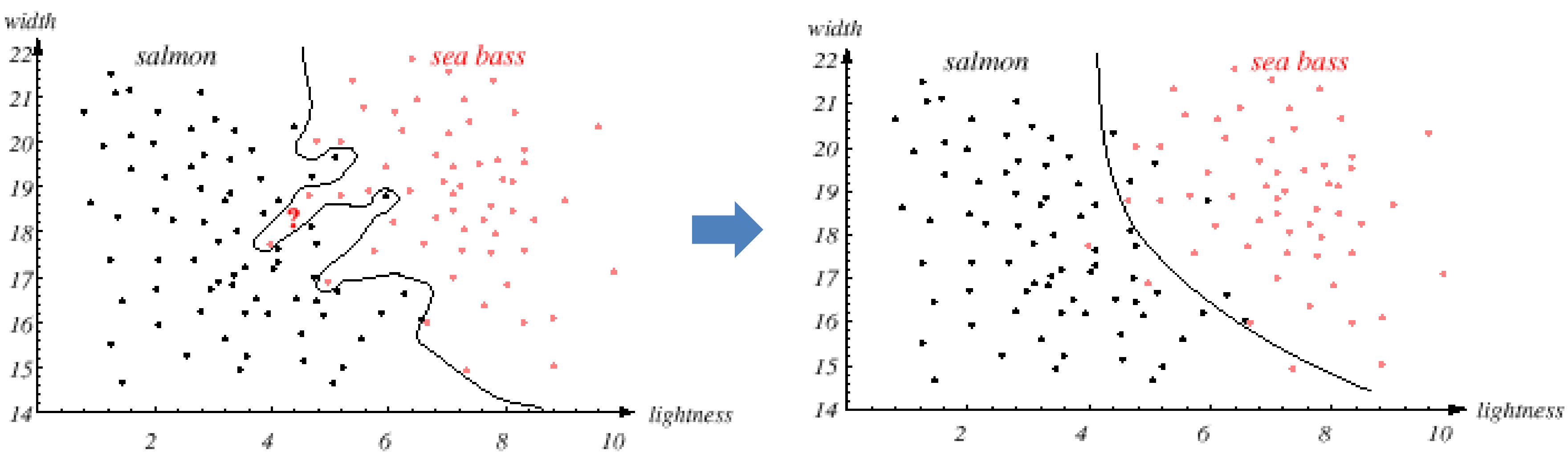
What is Classification in Machin Learning?

Example: Sorting incoming Fish on a conveyor according to species using optical sensing

Classification

- Threshold decision boundary and cost relationship

  ➔ However, our satisfaction is premature because the central aim of designing a classifier is to correctly classify novel input ➔ Issue of Generalization!



Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Bayesian Classification

Probabilistic Approach

Example: Sorting incoming Fish on a conveyor according to species using optical sensing

State of nature (prior): State of nature is a random variable

- If it is assumed that the catch of salmon and sea bass is equiprobable:
  - $P(\omega 1) = P(\omega 2)$ (uniform priors)
  - $P(\omega 1) + P(\omega 2) = 1$ (exclusivity and exhaustivity)

(1) Decision rule with only the prior information

- Decide $\omega 1$ if $P(\omega 1) > P(\omega 2)$; otherwise, decide $\omega 2$

(2) Use of the class-conditional information for classification

- $P(x \mid \omega 1)$ and $P(x \mid \omega 2)$ describe the difference in lightness between populations of sea-bass and salmon

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Bayesian Classification

Probabilistic Approach

(2) Use of the class-conditional information for classification

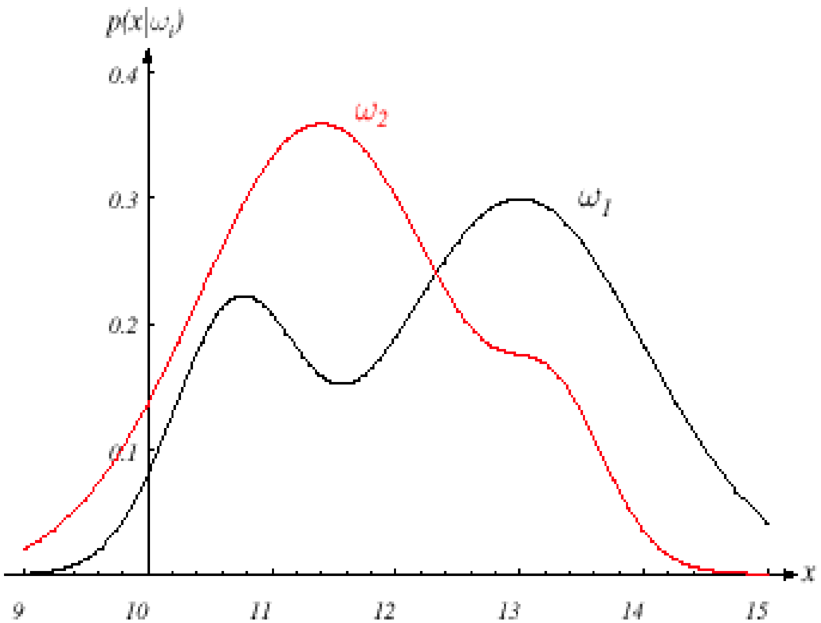


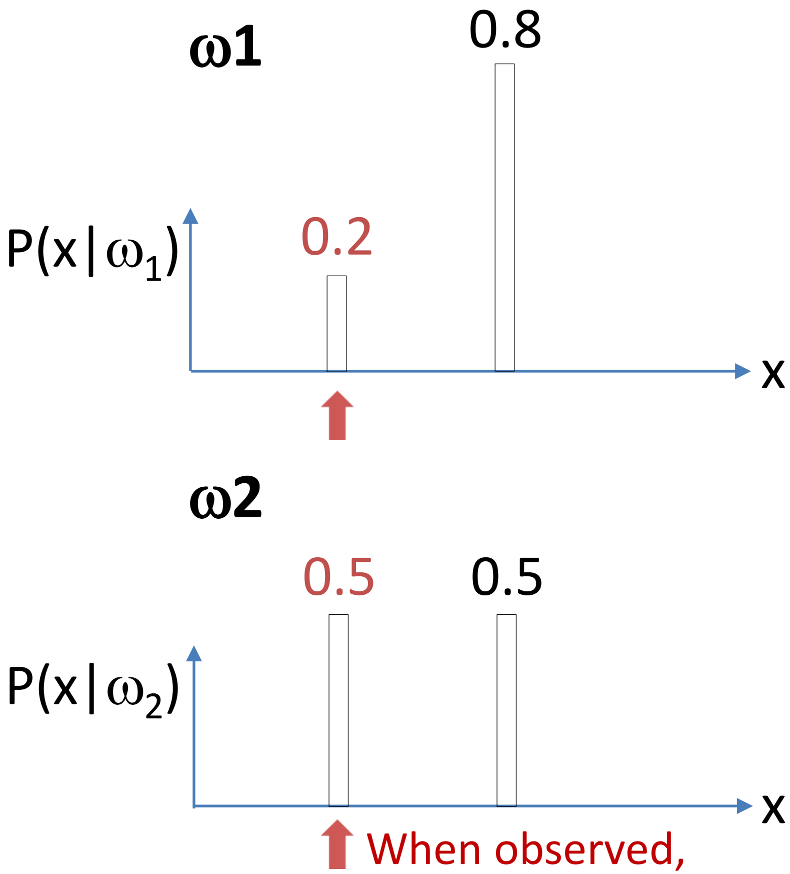
FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two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

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Bayesian Classification

Probabilistic Approach

(2) Use of the class-conditional information for classification



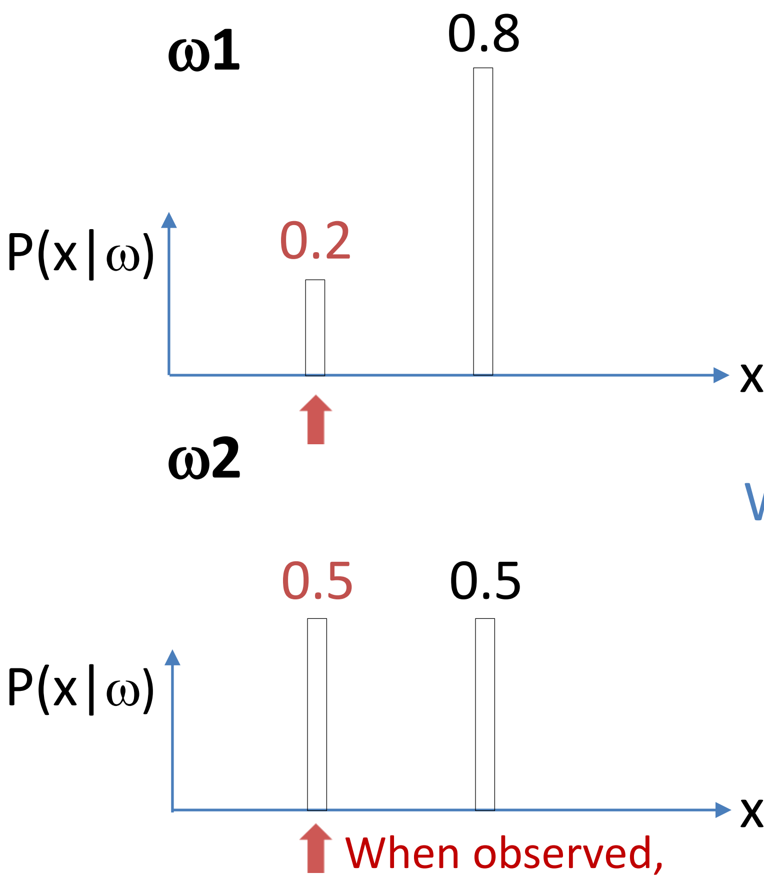
When observed,

➔ ω2 ?

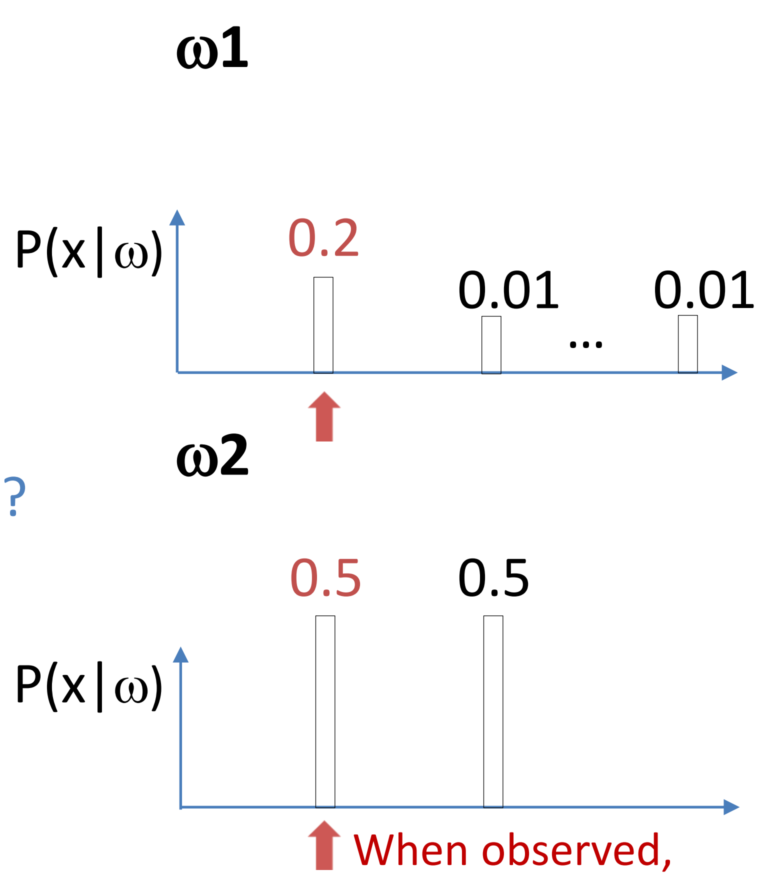
Bayesian Classification

Probabilistic Approach

(2) Use of the class-conditional information for classification



What about?



→ ω2 ?

→ Still ω2 ?

Bayesian Classification

Probabilistic Approach

Example: Sorting incoming Fish on a conveyor according to species using optical sensing

(3) Posterior, likelihood, and evidence

- $$P(\omega_j \mid x) = \frac{P(x \mid \omega_j) * P(\omega_j)}{P(x)} \qquad \text{(BAYES RULE)}$$

$\rightarrow$ $$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}}$$

- Where in case of two categories

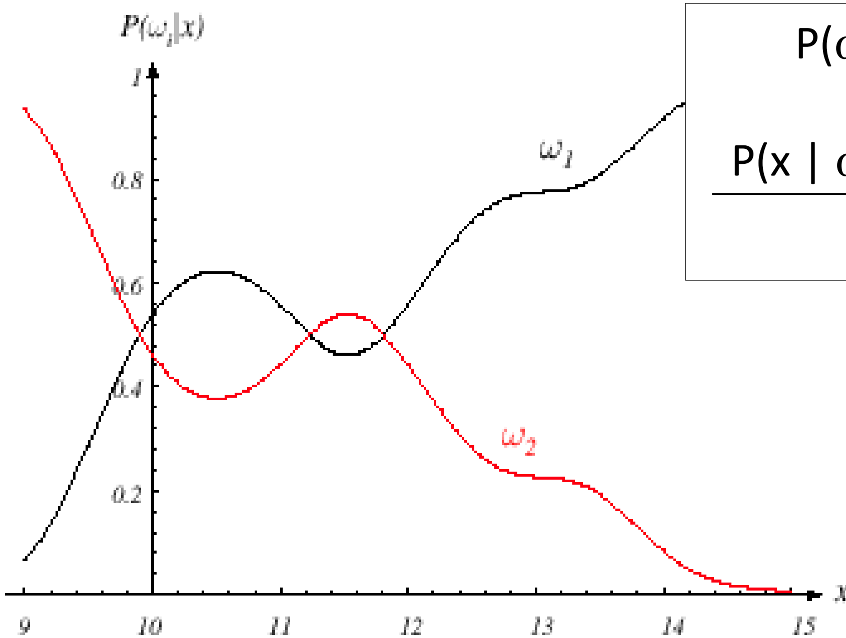
$$P(x) = \sum_{j=1}^{j=2} P(x \mid \omega_j) P(\omega_j)$$

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Bayesian Classification

Probabilistic Approach

(3) Posterior, likelihood, and evidence



$$P(\omega_1 \mid x) = \frac{P(x \mid \omega_1) * P(\omega_1)}{P(x)} \quad \overset{?}{> \text{ or } <} \quad P(\omega_2 \mid x) = \frac{P(x \mid \omega_2) * P(\omega_2)}{P(x)}$$

Decision Rule

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.

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Bayesian Classification

Probabilistic Approach

Example: Sorting incoming Fish on a conveyor according to species using optical sensing

(3) Posterior, likelihood, and evidence

- Intuitive decision rule given the posterior probabilities:

Given x:

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

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

- Why do this? Whenever we observe a particular x, the probability of error is :

$P(error \mid x) = P(\omega_1 \mid x)$ *if we decide* ω_2

$P(error \mid x) = P(\omega_2 \mid x)$ *if we decide* ω_1

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Bayesian Classification

Probabilistic Approach

Example: Sorting incoming Fish on a conveyor according to species using optical sensing

(3) Posterior, likelihood, and evidence

- Minimizing the probability of error

  Decide $\omega 1$ if $P(\omega 1 \mid x) > P(\omega 2 \mid x)$;

  otherwise, decide $\omega 2$

  Therefore:

  $$P(\text{error} \mid x) = \min [P(\omega 1 \mid x), P(\omega 2 \mid x)]$$

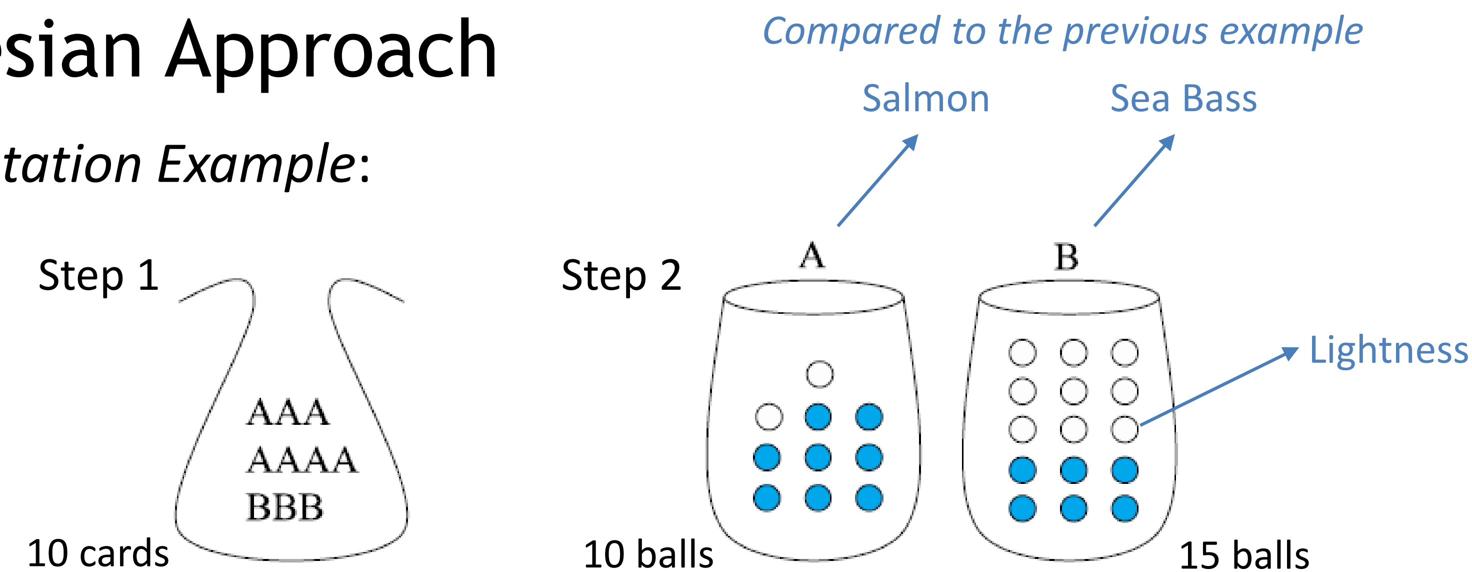
  (Bayes decision)

Source: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification.

Bayesian Classification

Bayesian Approach

Computation Example:



- Select a card (A or B in Step 1) and Pick a ball (white or blue in Step 2)
- Random variable: X ϵ {A, B}, Y ϵ {white, blue}

Task:

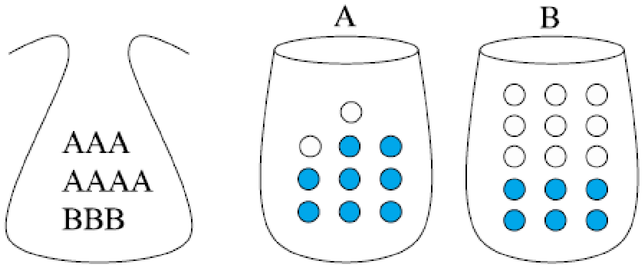
When a white ball is picked, which box (A or B) may the ball come from?

Source: Oh Il-Seok (2008). Pattern Recognition.

Bayesian Classification

Bayesian Approach

Computation Example: Basic probability



- Probability of selecting “A”
  - $P(X{=}A) = P(A) = 7/10$
- Probability of picking “white” in “A”?
  - $P(Y{=}white|X{=}A) = P(white|A) = 2/10$
- Probability of selecting “A” and “white”?
  - $P(A, white) = P(white|A)P(A) = (2/10)(7/10) = 7/50$

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

- Probability of picking “white”?
  - $P(white) = P(white|A)P(A)+P(white|B)P(B)$
    $= (2/10)(7/10)+(9/15)(3/10)=8/25$
- If $P(X,Y)=P(X)P(Y)$, X and Y is independent
- $P(X)$: *prior probability*

Source: Oh Il-Seok (2008). Pattern Recognition.

Bayesian Classification

Bayesian Approach

Computation Example:

When a white ball is picked, which box (A or B) may the ball come from?

- Approach: Calculating the probabilities that the ball may come from "A" and "B", then selecting the one with higher probability.

(1) Decision rule with only the prior information

- Comparing the probabilities of selecting "A" and "B"
- P(A)=7/10 > P(B)=3/10 ➔ Result = "A"

(Perhaps reasonable when P(white|A) ≈ P(white|B). But what if P(white|A)=0?)

(2) Use of the class-conditional information for classification

- Comparing the probabilities of selecting "white" in A and B
- P(white|A)=2/10 < P(white|B)=9/15 ➔ Result = "B"

(Perhaps reasonable when P(A) ≈ P(B). But what if P(A)=0.9999?)

Source: Oh Il-Seok (2008). Pattern Recognition.

Bayesian Classification

Bayesian Approach

Computation Example:

When a white ball is picked, which box (A or B) may the ball come from?

- Approach: Calculating the probabilities that the ball may come from “A” and “B”, then selecting the one with higher probability.

(3) Bayesian: Posterior, likelihood, and evidence

- Comparing the probabilities that the ball may come from “A” and “B” when the ball is “white”
- P(A|white)=0.4375 < P(B|white)=0.5625 ➔ Result = “B”

$$P(A|\text{white}) = \frac{P(\text{white}|A) * P(A)}{P(\text{white})} = \frac{(2/10) * (7/10)}{(8/25)} = 0.4375$$

$$P(B|\text{white}) = \frac{P(\text{white}|B) * P(B)}{P(\text{white})} = \frac{(9/15) * (3/10)}{(8/25)} = 0.5625$$

Source: Oh Il-Seok (2008). Pattern Recognition.

Bayesian Classification: General Model

- Training dataset

  $X=\{(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2), \ldots, (\mathbf{x}_N, t_N)\}$ (N: number of data points)

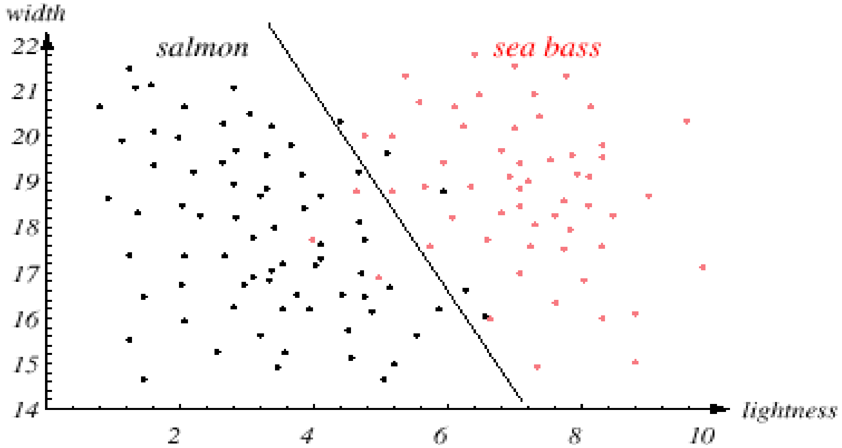
  - Feature vector: $\mathbf{x}_i = (x_1, x_2, \ldots, x_d)^T$ (d: dimension)
  - Label: $t_i \in \{\omega_1, \omega_2, \ldots, \omega_M\}$ (M: number of class)

Example:

$\mathbf{x}_i = (x_1, x_2)$ (x_1 = lightness, x_2 = width)

- $\mathbf{x}_1 = (5, 20)^T$, $t_1 = \omega_1$ (salmon)
- $\mathbf{x}_2 = (8, 15)^T$, $t_2 = \omega_2$ (sea bass)
- $\mathbf{x}_3 = (11, 2)^T$, $t_3 = \omega_2$ (sea bass)

  ...
- $\mathbf{x}_{100} = (2, 8)^T$, $t_{100} = \omega_1$ (salmon)



Source: Oh Il-Seok (2008). Pattern Recognition.

Bayesian Classification

Bayesian Classification: General Model

- Classification: decision rule (M = 2 classes)

Given a feature vector **x**:

If $P(\omega_1 \mid \mathbf{x}) > P(\omega_2 \mid \mathbf{x})$, classify **x** into ω_1

If $P(\omega_1 \mid \mathbf{x}) < P(\omega_2 \mid \mathbf{x})$, classify **x** into ω_2

To calculate $P(\omega_i \mid \mathbf{x})$,

$$P(\omega_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \omega_i) * P(\omega_i)}{P(\mathbf{x})} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}}$$

- Likelihood: Estimation based on samples of ω_i in training dataset
- Prior: Sample (e.g., $P(\omega_1)=n_1/N$, $P(\omega_2)=n_2/N$) (Note: N↑ → actual value)
- Evidence: In general, not necessary (we will compare)

Source: Oh Il-Seok (2008). Pattern Recognition.

Bayesian Classification

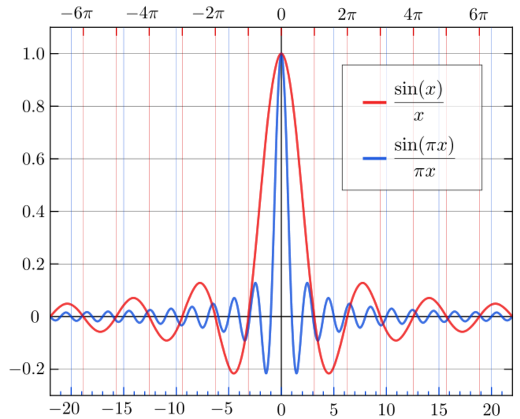
Bayesian Classification: General Model

- Classification: decision rule (M classes)
  - Label: $t_i \in \{\omega_1, \omega_2, \ldots, \omega_M\}$ (M: number of class)

Given a feature vector **x**:

If $k = \underset{i}{\arg\max}\, P(\omega_i \mid \mathbf{x})$, classify **x** into ω_k

(← If $P(\omega_k \mid \mathbf{x}) > P(\omega_{\text{any others}} \mid \mathbf{x})$, classify **x** into ω_k)



[Note]
arg max: arguments of the maxima are the points of the domain of some function at which the function values are maximized.

Example: both functions (i.e., blue and red) have arg max of {0}.

Source: Oh Il-Seok (2008). Pattern Recognition.

Bayesian Classification

Bayesian Classification: General Model

- Classification: decision rule (M = 2 classes)

To calculate $P(\omega_i \mid \mathbf{x})$,

$$P(\omega_i \mid \mathbf{x})$$

$$= \frac{P(\mathbf{x} \mid \omega_i) * P(\omega_i)}{P(\mathbf{x})}$$

```
# calculate the probability that each label occurs
probability_label_0 = len(separated[0]) / len(trainSet)
probability_label_1 = len(separated[1]) / len(trainSet)
probability_label= [probability_label_0, probability_label_1]
```

```
def calculateProbability(x, mean, stdev): # to calculate the probability, given x
        exponent = math.exp(-(math.pow(x-mean,2)/(2*math.pow(stdev,2))))
        return (1 / (math.sqrt(2*math.pi) * stdev)) * exponent

# to predict the label with datasets (testSet)
# testSet = trainSet # un-comment to switch the datasets when calculting the accuracy for trainSet
predictions = []
for h in range(len(testSet)):
    probabilities = {}
    for classValue, classSummaries in summaries.items():
        probabilities[classValue] = 1 * probability_label[int(classValue)] # initialization

        for i in range(len(classSummaries)): # len(classSummaries) = the number of features
            mean, stdev = classSummaries[i]
            x = testSet[h][i] # [0] for the first datapoint
            probabilities[classValue] *= calculateProbability(x, mean, stdev)
```

In general, not necessary
(we will compare)

Bayesian Classification

Bayesian Classification: General Model

- Classification: decision rule (M = 2 classes)

Given a feature vector **x**:

If $P(\omega_1 \mid \mathbf{x}) > P(\omega_2 \mid \mathbf{x})$, classify **x** into ω_1

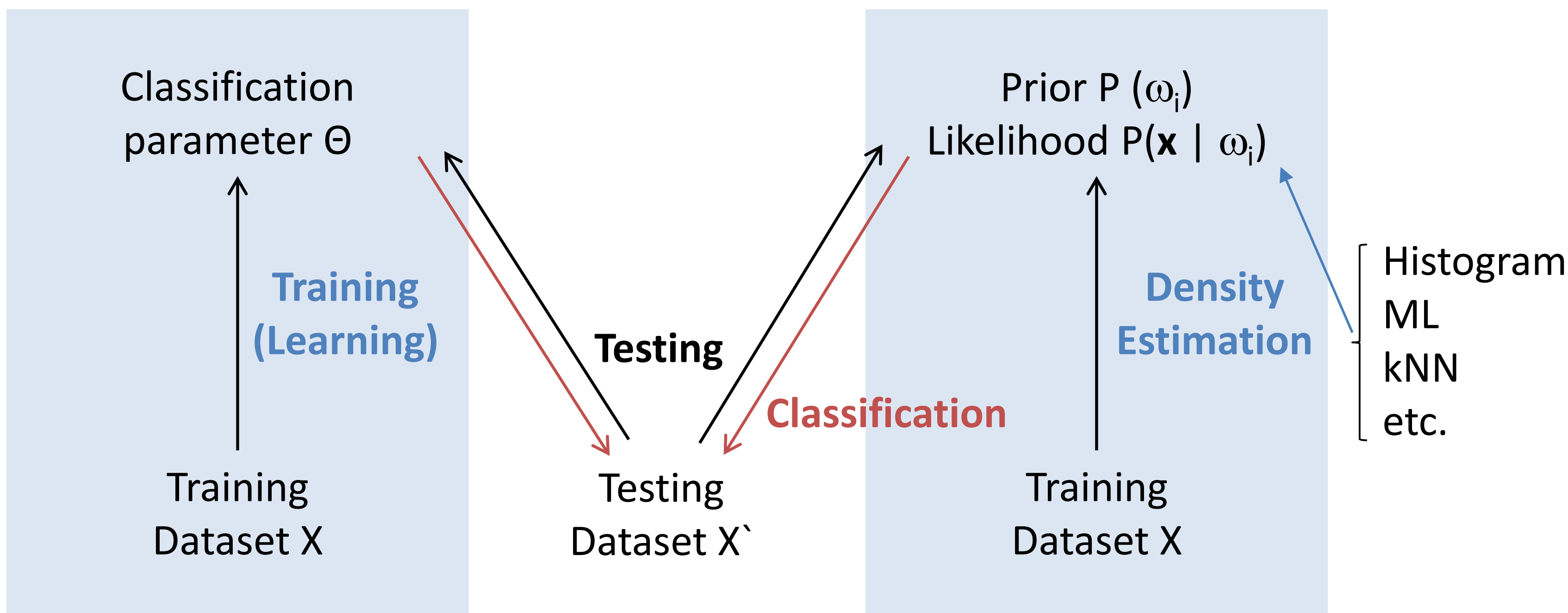
If $P(\omega_1 \mid \mathbf{x}) < P(\omega_2 \mid \mathbf{x})$, classify **x** into ω_2

```
for h in range(len(testSet)):

    …

    # predict the label: Decision Rule
    if probabilities[0] > probabilities[1]:
        predictions.append(0)
    else:
        predictions.append(1)
```

Source: Oh Il-Seok (2008). Pattern Recognition.

Bayesian Classification: General Model

Learning in general vs. in Bayesian classification



General Learning Process

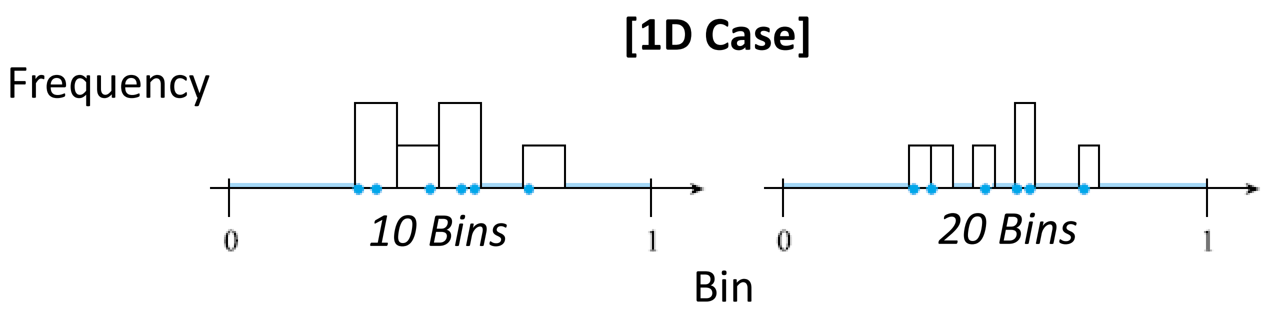
Bayesian Learning Process

Source: Oh Il-Seok (2008). Pattern Recognition.

Bayesian Classification

Probability Distribution Estimation

Histogram



- Histogram may work when the dimension is low and a large amount of sample (X) is available.
- However, s^d bins are needed when there are d dimensions and s bins per dimensions (*Curse of Dimensionality*). For this issue, other methods can be applied.

Source: Oh Il-Seok (2008). Pattern Recognition.