T2: Computational learning

Fundamentos del Aprendizaje Automático

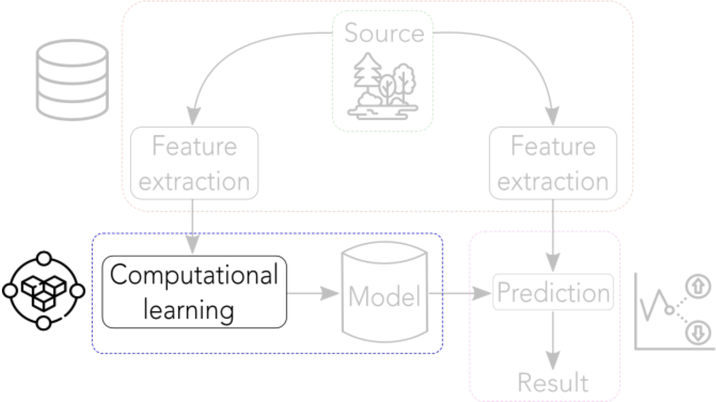
Curso 2025/2026

Structure

1. Introduction
   - Where are we?
   - Computational learning VS Decision theory
2. Bayesian decision theory
   - Two-class problem
   - General form
   - Risk
   - Discriminant functions
3. Statistical likelihood
   - Maximum likelihood estimation
4. Issues in computational learning
   - Bias-variance issues
   - Curse of dimensionality

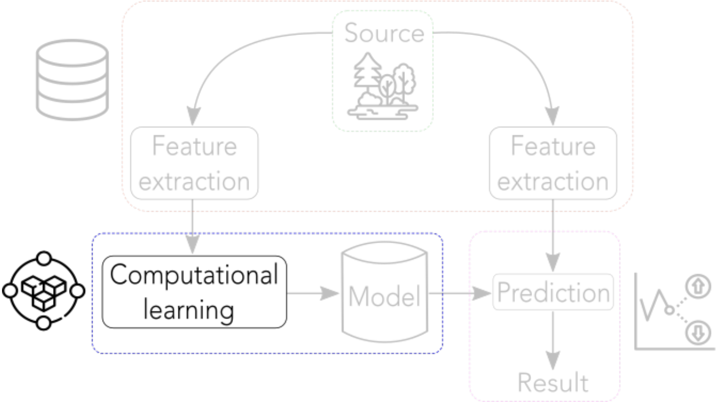
Outline

Where are we?



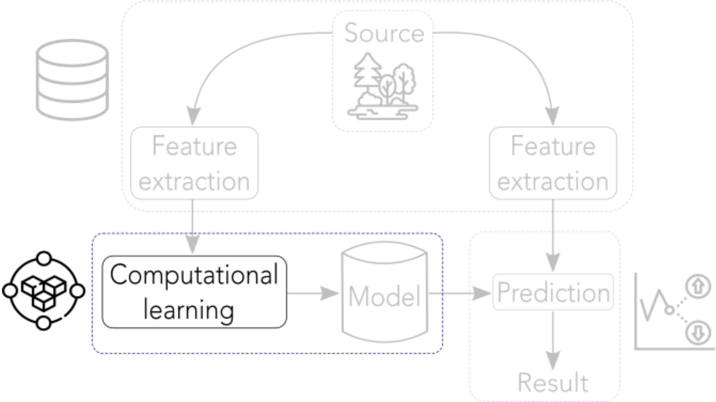
Source
Feature extraction
Feature extraction
Computational learning
Model
Prediction
Result

Where are we?



Computational learning:

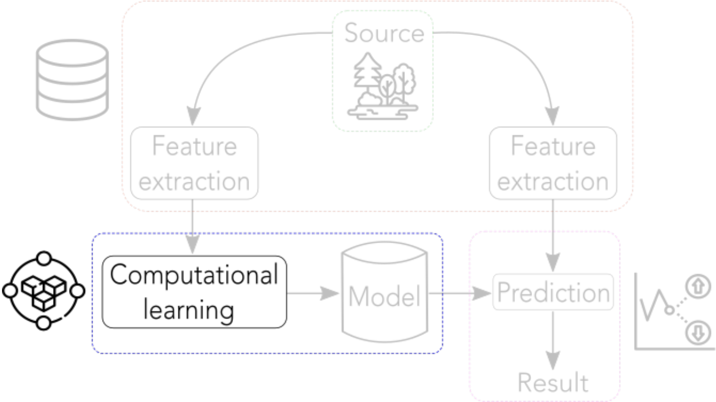
Where are we?



Computational learning:

- Infer knowledge from data

Where are we?



Computational learning:

- Infer knowledge from data
- Derive a model out of that knowledge

Computational learning VS Decision theory

Computational learning VS Decision theory

Two related, but different, concepts:

Computational learning VS Decision theory

Two related, but different, concepts:

- **Computational learning**: Studies the limits of learning from data

Computational learning VS Decision theory

Two related, but different, concepts:

- **Computational learning**: Studies the limits of learning from data
- **Decision theory**: Make optimal decisions under uncertainty

Computational learning VS Decision theory

Two related, but different, concepts:

- **Computational learning**: Studies the limits of learning from data
  - What can be learned and how efficiently
  - Also statistical learning theory
- **Decision theory**: Make optimal decisions under uncertainty
  - Foundational mathematical framework for Machine Learning
  - Pattern classification is the most important subfield

Computational learning VS Decision theory

Two related, but different, concepts:

- **Computational learning**: Studies the limits of learning from data
  - What can be learned and how efficiently
  - Also statistical learning theory
- **Decision theory**: Make optimal decisions under uncertainty
  - Foundational mathematical framework for Machine Learning
  - Pattern classification is the most important subfield

Bayesian decision theory (for classification):

Computational learning VS Decision theory

Two related, but different, concepts:

- **Computational learning**: Studies the limits of learning from data
  - What can be learned and how efficiently
  - Also statistical learning theory
- **Decision theory**: Make optimal decisions under uncertainty
  - Foundational mathematical framework for Machine Learning
  - Pattern classification is the most important subfield

Bayesian decision theory (for classification):

- Fundamental statistical approach for the pattern classification problem

Computational learning VS Decision theory

Two related, but different, concepts:

- **Computational learning**: Studies the limits of learning from data
  - What can be learned and how efficiently
  - Also statistical learning theory
- **Decision theory**: Make optimal decisions under uncertainty
  - Foundational mathematical framework for Machine Learning
  - Pattern classification is the most important subfield

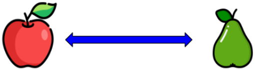
Bayesian decision theory (for classification):

- Fundamental statistical approach for the pattern classification problem
- Quantifying tradeoffs between various classification decisions using probability and the cost of such decisions

Outline

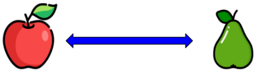
Introduction

Introduction

Task: 

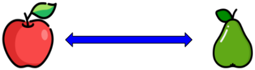
Introduction

Task:



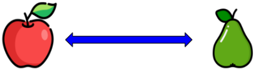
- **State of nature** (w): future event not under the control of the decision maker

Introduction

Task: 

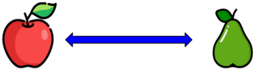
- **State of nature** (w): future event not under the control of the decision maker
  - $\rightarrow$ In our case two classes: apple (ω_1) and pear (ω_2)

Introduction

Task: 

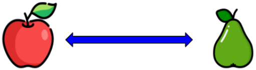
- **State of nature** (w): future event not under the control of the decision maker
  - $\rightarrow$ In our case two classes: apple (ω_1) and pear (ω_2)
- We assume the existence of some *a priori probability* / *prior*:

Introduction

Task: 

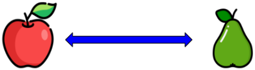
- **State of nature** (w): future event not under the control of the decision maker
  - $\rightarrow$ In our case two classes: apple (ω_1) and pear (ω_2)
- We assume the existence of some *a priori probability* / *prior*:
  - $\rightarrow$ $P(\omega_1)$ that the element is an apple (ω_1) and $P(\omega_2)$ that the element is a pear (ω_2)

Introduction

Task: 

- **State of nature** (w): future event not under the control of the decision maker
  - $\rightarrow$ In our case two classes: apple (ω_1) and pear (ω_2)
- We assume the existence of some *a priori probability* / *prior*:
  - $\rightarrow$ $P(\omega_1)$ that the element is an apple (ω_1) and $P(\omega_2)$ that the element is a pear (ω_2)
  - $\rightarrow$ Based on geographical region, season...

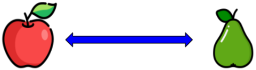
Introduction

Task: 

- **State of nature** (w): future event not under the control of the decision maker
  - $\rightarrow$ In our case two classes: apple (ω_1) and pear (ω_2)
- We assume the existence of some *a priori probability* / *prior*:
  - $\rightarrow$ $P(\omega_1)$ that the element is an apple (ω_1) and $P(\omega_2)$ that the element is a pear (ω_2)
  - $\rightarrow$ Based on geographical region, season...
- Decision rule:

$$w = \begin{cases} w_1 & \text{if } P(\omega_1) > P(\omega_2) \\ w_2 & \text{otherwise} \end{cases}$$

Introduction

Task: 

- **State of nature** (w): future event not under the control of the decision maker
  - → In our case two classes: apple (ω_1) and pear (ω_2)
- We assume the existence of some *a priori probability*/*prior*:
  - → $P(\omega_1)$ that the element is an apple (ω_1) and $P(\omega_2)$ that the element is a pear (ω_2)
  - → Based on geographical region, season...
- Decision rule:

$$w = \begin{cases} w_1 & \text{if } P(\omega_1) > P(\omega_2) \\ w_2 & \text{otherwise} \end{cases}$$

- Additional information may be incorporated!

Class-conditional probability density

- Different fruits will yield different color readings

Class-conditional probability density

- Different fruits will yield different color readings
  - Single descriptor/feature to encode color $\Rightarrow x \in \mathbb{R}$

Class-conditional probability density

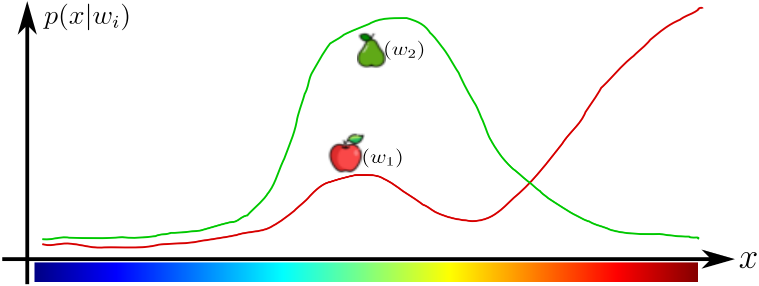
- Different fruits will yield different color readings
  - Single descriptor/feature to encode color $\Rightarrow x \in \mathbb{R}$
- Class-conditional probability density: $p(x|\omega)$

Class-conditional probability density

- Different fruits will yield different color readings
  - Single descriptor/feature to encode color $\Rightarrow x \in \mathbb{R}$
- Class-conditional probability density: $p(x|\omega)$
  - How likely is it to observe color x on fruit ω?

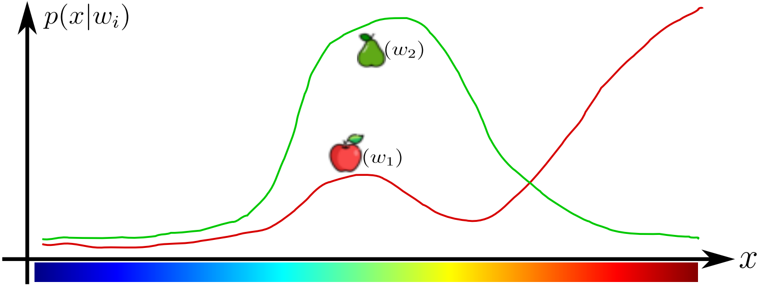
Class-conditional probability density

- Different fruits will yield different color readings
  - Single descriptor/feature to encode color $\Rightarrow x \in \mathbb{R}$
- Class-conditional probability density: $p(x|\omega)$
  - How likely is it to observe color x on fruit ω?



Class-conditional probability density

- Different fruits will yield different color readings
  - Single descriptor/feature to encode color $\Rightarrow x \in \mathbb{R}$
- Class-conditional probability density: $p(x|\omega)$
  - How likely is it to observe color x on fruit ω?



$$\int p(x|\omega_i)\, dx = 1 \quad \forall i \in \{1, 2\}$$

Joint probability density

- Likelihood for simultaneous occurrence of different variables

Joint probability density

- Likelihood for simultaneous occurrence of different variables
  - In our case $\Rightarrow P(\omega_j, x)$

Joint probability density

- Likelihood for simultaneous occurrence of different variables
  - In our case $\Rightarrow P(\omega_j, x)$
- If statistically independent: $p(\omega_j, x) = P(\omega_j) \cdot p(x)$

Joint probability density

- Likelihood for simultaneous occurrence of different variables
  - In our case $\Rightarrow P(\omega_j, x)$

✗ If statistically independent: $p(\omega_j, x) = P(\omega_j) \cdot p(x)$

✓ If statistically dependent:

Joint probability density

- Likelihood for simultaneous occurrence of different variables
  - In our case $\Rightarrow P(\omega_j, x)$

✗ If statistically independent: $p(\omega_j, x) = P(\omega_j) \cdot p(x)$

✓ If statistically dependent:

$$p(\omega_j, x)$$

Joint probability density

- Likelihood for simultaneous occurrence of different variables
  - In our case $\Rightarrow P(\omega_j, x)$

✗ If statistically independent: $p(\omega_j, x) = P(\omega_j) \cdot p(x)$

✓ If statistically dependent:

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

Joint probability density

- Likelihood for simultaneous occurrence of different variables
  - In our case $\Rightarrow P(\omega_j, x)$

✗ If statistically independent: $p(\omega_j, x) = P(\omega_j) \cdot p(x)$

✓ If statistically dependent:

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

Joint probability density

- Likelihood for simultaneous occurrence of different variables
  - In our case $\Rightarrow P(\omega_j, x)$

✗ If statistically independent: $p(\omega_j, x) = P(\omega_j) \cdot p(x)$

✓ If statistically dependent:

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

$$\downarrow$$

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

Bayes theorem

Bayes theorem

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

Bayes theorem

Bayes theorem

$$\underbrace{P(\omega_j|x)}_{posterior} = \frac{\overbrace{p(x|\omega_j)}^{likelihood} \cdot \overbrace{P(\omega_j)}^{prior}}{\underbrace{p(x)}_{evidence}}$$

Bayes theorem

Bayes theorem

$$\underbrace{P(\omega_j|x)}_{posterior} = \frac{\overbrace{p(x|\omega_j)}^{likelihood} \cdot \overbrace{P(\omega_j)}^{prior}}{\underbrace{p(x)}_{evidence}}$$

- **Posterior**: Observing state of nature w_j given feature value x

Bayes theorem

Bayes theorem

$$\underbrace{P(\omega_j|x)}_{posterior} = \frac{\overbrace{p(x|\omega_j)}^{likelihood} \cdot \overbrace{P(\omega_j)}^{prior}}{\underbrace{p(x)}_{evidence}}$$

- **Posterior**: Observing state of nature w_j given feature value x
- **Likelihood**: Observing feature value x assuming that state of nature w_j is correct

Bayes theorem

Bayes theorem

$$\underbrace{P(\omega_j|x)}_{posterior} = \frac{\overbrace{p(x|\omega_j)}^{likelihood} \cdot \overbrace{P(\omega_j)}^{prior}}{\underbrace{p(x)}_{evidence}}$$

- **Posterior**: Observing state of nature w_j given feature value x
- **Likelihood**: Observing feature value x assuming that state of nature w_j is correct
- **Prior**: Initial probability of observing state of nature w_j without further evidence (existing knowledge, historical data...)

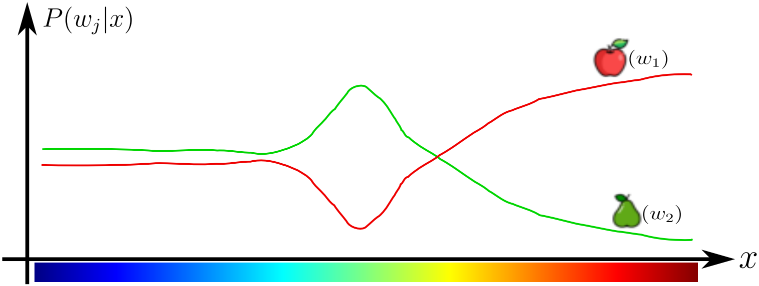
Bayes theorem

Bayes theorem

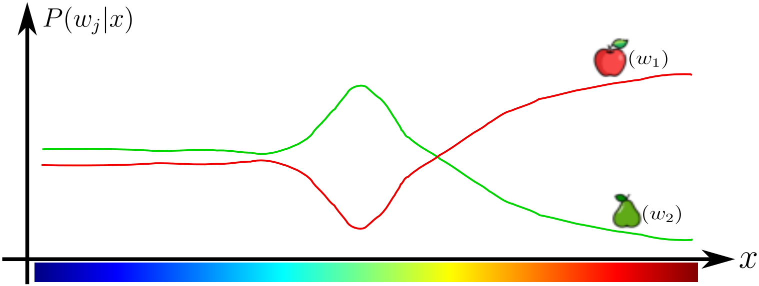
$$\underbrace{P(\omega_j|x)}_{posterior} = \frac{\overbrace{p(x|\omega_j)}^{likelihood} \cdot \overbrace{P(\omega_j)}^{prior}}{\underbrace{p(x)}_{evidence}}$$

- **Posterior**: Observing state of nature w_j given feature value x
- **Likelihood**: Observing feature value x assuming that state of nature w_j is correct
- **Prior**: Initial probability of observing state of nature w_j without further evidence (existing knowledge, historical data...)
- **Evidence**: Scale factor $\Rightarrow p(x) = \sum_{\forall j} p(x|\omega_j) \cdot P(\omega_j)$

Posterior probability



Posterior probability



For any color $x = x_0 \longrightarrow P(\omega_1|x_0) + P(\omega_2|x_0) = 1$

Practical exercise

Estimate the probability of obtaining a particular fruit (state of nature) given *priors* and its color

1. Historically, the harvest of apples represents the 70% of the fruit, whereas that of pears is the 30%
2. The distribution of color for each fruit is:

Color ($[0, 100]$)	Blue (0)	Green (40)	Yellow (50)	Orange (70)	Red (100)
$p(x \mid w_1)$	0.05	0.25	0.15	0.1	0.45
$p(x \mid w_2)$	0.05	0.6	0.25	0.05	0.05

Q1. How likely is it to obtain a yellow apple? And a same-colored pear?

Q2. Obtain all posterior probabilities for all possible color and fruit combinations

Probability of error

- Given an observation $x = x_0$, one would select as state of nature:

Probability of error

- Given an observation $x = x_0$, one would select as state of nature:
  $\rightarrow$ w_1 if $P(w_1|x_0) > P(w_2|x_0)$

Probability of error

- Given an observation $x = x_0$, one would select as state of nature:
  - $\rightarrow$ w_1 if $P(w_1|x_0) > P(w_2|x_0)$
  - $\rightarrow$ w_2 otherwise

Probability of error

- Given an observation $x = x_0$, one would select as state of nature:
  - $\rightarrow$ w_1 if $P(w_1|x_0) > P(w_2|x_0)$
  - $\rightarrow$ w_2 otherwise
- Is this correct?

Probability of error

- Given an observation $x = x_0$, one would select as state of nature:
  - $\rightarrow$ w_1 if $P(w_1|x_0) > P(w_2|x_0)$
  - $\rightarrow$ w_2 otherwise
- Is this correct? Probability of error when performing this action:

Probability of error

- Given an observation $x = x_0$, one would select as state of nature:
  - $\rightarrow$ w_1 if $P(w_1|x_0) > P(w_2|x_0)$
  - $\rightarrow$ w_2 otherwise
- Is this correct? Probability of error when performing this action:

$$P(\text{error}|x_0) = \begin{cases} P(w_1|x_0) & \text{if we decide } w_2 \\ P(w_2|x_0) & \text{if we decide } w_1 \end{cases}$$

Probability of error

- Given an observation $x = x_0$, one would select as state of nature:
  - $\rightarrow$ w_1 if $P(w_1|x_0) > P(w_2|x_0)$
  - $\rightarrow$ w_2 otherwise
- Is this correct? Probability of error when performing this action:

$$P(\text{error}|x_0) = \begin{cases} P(w_1|x_0) & \text{if we decide } w_2 \\ P(w_2|x_0) & \text{if we decide } w_1 \end{cases}$$

- We can minimize this expression by selecting:
  - $\rightarrow$ w_1 when $P(w_1|x_0) > P(w_2|x_0)$
  - $\rightarrow$ w_2 otherwise

Probability of error

- Will this rule minimize the average probability of error?

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, x)\, dx = \int_{-\infty}^{\infty} P(\text{error}|x) \cdot p(x)\, dx$$

Probability of error

- Will this rule minimize the average probability of error?

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, x)\, dx = \int_{-\infty}^{\infty} P(\text{error}|x) \cdot p(x)\, dx$$

- Minimize $P(\text{error}) \Rightarrow$ ensure $P(\text{error}|x)$ is as small as possible $\forall\ x$

Probability of error

- Will this rule minimize the average probability of error?

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, x)\, dx = \int_{-\infty}^{\infty} P(\text{error}|x) \cdot p(x)\, dx$$

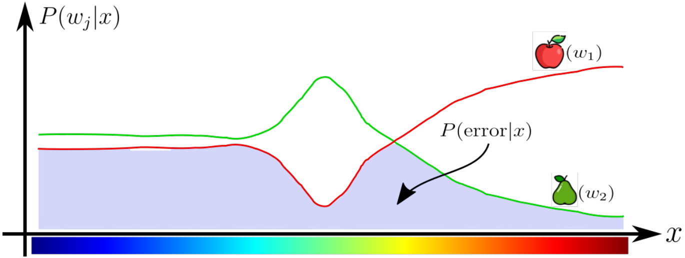
- Minimize $P(\text{error}) \Rightarrow$ ensure $P(\text{error}|x)$ is as small as possible $\forall\ x$
  $\rightarrow\ P(\text{error}|x) = \min[P(w_1|x), P(w_2|x)]$

Probability of error

- Will this rule minimize the average probability of error?

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, x)\, dx = \int_{-\infty}^{\infty} P(\text{error}|x) \cdot p(x)\, dx$$

- Minimize $P(\text{error}) \Rightarrow$ ensure $P(\text{error}|x)$ is as small as possible $\forall\ x$
  $\rightarrow\ P(\text{error}|x) = \min[P(w_1|x), P(w_2|x)]$



Bayes decision rule

Bayes decision rule

- Optimal decision rule $\Rightarrow$ minimizes probability of error

$$\textit{Decide } w_1 \textit{ if } P(w_1|x) > P(w_2|x)\textit{; otherwise decide } w_2$$

Bayes decision rule

- Optimal decision rule $\Rightarrow$ minimizes probability of error

$$\textit{Decide } w_1 \textit{ if } P(w_1|x) > P(w_2|x)\textit{; otherwise decide } w_2$$

- **Evidence** is unimportant as far as making a decision is concerned

Bayes decision rule

- Optimal decision rule $\Rightarrow$ minimizes probability of error

$$\textit{Decide } w_1 \textit{ if } P(w_1|x) > P(w_2|x)\textit{; otherwise decide } w_2$$

- **Evidence** is unimportant as far as making a decision is concerned
  - $\rightarrow$ It may be obviated

Bayes decision rule

- Optimal decision rule $\Rightarrow$ minimizes probability of error

  Decide w_1 if $P(w_1|x) > P(w_2|x)$; otherwise decide w_2

- **Evidence** is unimportant as far as making a decision is concerned
  $\rightarrow$ It may be obviated

Bayes decision rule for two-class classification

Decide w_1 if $p(x|w_1) \cdot P(w_1) > p(x|w_2) \cdot P(w_2)$;
otherwise decide w_2

Bayes decision rule

- Optimal decision rule ⇒ minimizes probability of error

$$\textit{Decide } w_1 \textit{ if } P(w_1|x) > P(w_2|x)\textit{; otherwise decide } w_2$$

- **Evidence** is unimportant as far as making a decision is concerned
  → It may be obviated

Bayes decision rule for two-class classification

Decide w_1 if $p(x|w_1) \cdot P(w_1) > p(x|w_2) \cdot P(w_2)$;
otherwise decide w_2

- The rule may be also written as:

$$w = \underset{w \in \{w_1, w_2\}}{\arg\max}\ P(w|x) = \underset{w \in \{w_1, w_2\}}{\arg\max}\ p(x|w) \cdot P(w)$$

Practical exercise

Considering the same exercise as before where

1. Historically, the harvest of apples represents the 70% of the fruit, whereas that of pears is the 30%
2. The distribution of color for each fruit is:

Color ([0, 100])	Blue (0)	Green (40)	Yellow (50)	Orange (70)	Red (100)
$p(x\|w_1)$	0.05	0.25	0.15	0.1	0.45
$p(x\|w_2)$	0.05	0.6	0.25	0.05	0.05

Q1. Which is the minimum error in each case considering the Bayes decision rule?

Q2. What would occur is priors were equiprobable? $P(w_1) = P(w_2) = 50\%$

Definitions

We shall generalize these ideas if four different aspects:

Definitions

We shall generalize these ideas if four different aspects:

1) More than one feature

Definitions

We shall generalize these ideas if four different aspects:

1) More than one feature
2) More than two states of nature

Definitions

We shall generalize these ideas if four different aspects:

1) More than one feature
2) More than two states of nature
3) Actions beyond selecting labels

Definitions

We shall generalize these ideas if four different aspects:

1) More than one feature
2) More than two states of nature
3) Actions beyond selecting labels
4) Loss function as a generalization of the probability of error

Definitions

Definitions

- Feature space: use more than a single descriptor

Definitions

- Feature space: use more than a single descriptor
  $\rightarrow$ Fixed d-dimensional space: $\mathbf{x} \in \mathbb{R}^d$

Definitions

- Feature space: use more than a single descriptor
  - $\rightarrow$ Fixed d-dimensional space: $\mathbf{x} \in \mathbb{R}^d$
- States of nature: allow more than two states

Definitions

- Feature space: use more than a single descriptor
  - $\rightarrow$ Fixed d-dimensional space: $\mathbf{x} \in \mathbb{R}^d$
- States of nature: allow more than two states
  - $\rightarrow$ Finite set of states: $\mathcal{W} = \left\{w_1, \ldots, w_{|\mathcal{W}|}\right\}$

Definitions

- Feature space: use more than a single descriptor
  - $\rightarrow$ Fixed d-dimensional space: $\mathbf{x} \in \mathbb{R}^d$
- States of nature: allow more than two states
  - $\rightarrow$ Finite set of states: $\mathcal{W} = \{w_1, \ldots, w_{|\mathcal{W}|}\}$
- Actions: possibilitate other beyond classification

Definitions

- Feature space: use more than a single descriptor
  - $\rightarrow$ Fixed d-dimensional space: $\mathbf{x} \in \mathbb{R}^d$
- States of nature: allow more than two states
  - $\rightarrow$ Finite set of states: $\mathcal{W} = \left\{w_1, \ldots, w_{|\mathcal{W}|}\right\}$
- Actions: possibilitate other beyond classification
  - $\rightarrow$ Finite set of actions: $\mathcal{A} = \left\{\alpha_1, \ldots, \alpha_{|\mathcal{A}|}\right\}$ s.t. $\alpha_i : \mathbb{R}^d \rightarrow \mathcal{W}$

Definitions

- Feature space: use more than a single descriptor
  - $\rightarrow$ Fixed d-dimensional space: $\mathbf{x} \in \mathbb{R}^d$
- States of nature: allow more than two states
  - $\rightarrow$ Finite set of states: $\mathcal{W} = \left\{ w_1, \ldots, w_{|\mathcal{W}|} \right\}$
- Actions: possibilitate other beyond classification
  - $\rightarrow$ Finite set of actions: $\mathcal{A} = \left\{ \alpha_1, \ldots, \alpha_{|\mathcal{A}|} \right\}$ s.t. $\alpha_i : \mathbb{R}^d \rightarrow \mathcal{W}$
- Loss function: Generalization of probability of error $\Rightarrow$ errors are not equally important

Definitions

- Feature space: use more than a single descriptor
  - $\rightarrow$ Fixed d-dimensional space: $\mathbf{x} \in \mathbb{R}^d$
- States of nature: allow more than two states
  - $\rightarrow$ Finite set of states: $\mathcal{W} = \left\{w_1, \ldots, w_{|\mathcal{W}|}\right\}$
- Actions: possibilitate other beyond classification
  - $\rightarrow$ Finite set of actions: $\mathcal{A} = \left\{\alpha_1, \ldots, \alpha_{|\mathcal{A}|}\right\}$ s.t. $\alpha_i : \mathbb{R}^d \rightarrow \mathcal{W}$
- Loss function: Generalization of probability of error $\Rightarrow$ errors are not equally important
  - $\rightarrow$ Cost of performing α_i when true state is $w_j \Rightarrow \lambda(\alpha_i | w_j)$

Definitions

- Feature space: use more than a single descriptor
  - $\rightarrow$ Fixed d-dimensional space: $\mathbf{x} \in \mathbb{R}^d$
- States of nature: allow more than two states
  - $\rightarrow$ Finite set of states: $\mathcal{W} = \left\{w_1, \ldots, w_{|\mathcal{W}|}\right\}$
- Actions: possibilitate other beyond classification
  - $\rightarrow$ Finite set of actions: $\mathcal{A} = \left\{\alpha_1, \ldots, \alpha_{|\mathcal{A}|}\right\}$ s.t. $\alpha_i : \mathbb{R}^d \rightarrow \mathcal{W}$
- Loss function: Generalization of probability of error $\Rightarrow$ errors are not equally important
  - $\rightarrow$ Cost of performing α_i when true state is $w_j \Rightarrow \lambda(\alpha_i|w_j)$

Bayes theorem (general form)

$$P(w_j|\mathbf{x}) = \frac{p(\mathbf{x}|w_j) \cdot P(w_j)}{p(\mathbf{x})} \quad \text{with} \quad p(\mathbf{x}) = \sum_{j=1}^{|\mathcal{W}|} p(\mathbf{x}|w_j) \cdot P(w_j)$$