# Reconnaissance de formes I

## Approches statistiques

Yves Goussard

GBM6103A

19 novembre 2014

# Plan

# Exemple I : évaluation du risque de maladies vasculaires

## Position du problème

Existence de bases de données épidémiologiques

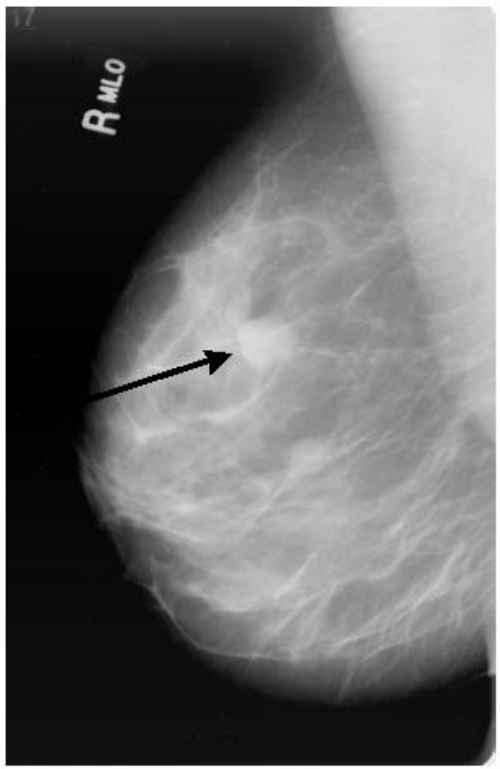
Paramètres morphologiques, paramètres géographiques, historique médical, mode de vie...

## Démarche pour évaluation du risque couru par un individu

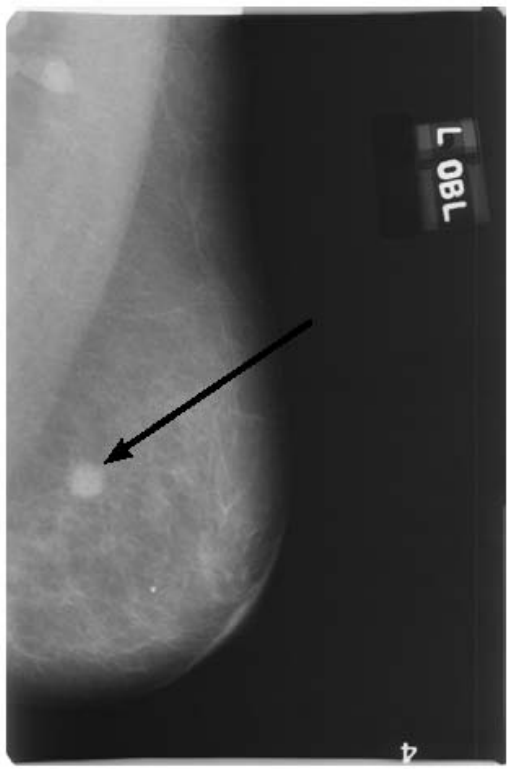
- Fixer les objectifs (classes) : risque faible, moyen, élevé...
- Déterminer des *attributs* permettant d'effectuer la classification
  - taille, âge, poids
  - lieu de résidence (ville, banlieue, milieu rural, région nordique...)
  - nombre d'heures d'activité physique par semaine, consommation quotidienne de tabac, d'alcool...
- Analyser le comportement des attributs dans chaque classe choisie, à partir des bases de données existantes
- En fonction de l'analyse, classer la population à évaluer

# Exemple II : aide à la classification de mammogrammes (1)

## Exemples de mammogrammes




Benign mass




Malignant mass

# Exemple II : aide à la classification de mammogrammes (2)

## Position du problème

- Bases de données mammographiques examinées par des experts (présence de tumeurs bénignes, malignes, mammogrammes normaux)
- Objectif : aide au diagnostic

## Démarche

- Fixer les objectifs (deux classes, ou plus ?).
- Déterminer des *attributs* dans les mammogrammes qui soient caractéristiques des classes choisies. Difficultés :
  - variabilité des mammogrammes
  - fiabilité des traitements permettant le calcul des attributs
  - immunité aux questions de normalisation, d'orientation...
- Analyser le comportement des attributs dans chaque classe choisie, à partir des bases de données existantes
- En fonction de l'analyse, classer les mammogrammes à traiter

# Exemple III : classification automatique des ECG

## Position du problème

- Objectifs possibles : surveillance de patients, aide au diagnostic
- Démarche : voir exemple II
- Particularité : structure forte dans un ECG
  - présence de plusieurs ondes
  - durée stable de chaque onde
  - ordre significatif

## Conséquence

Les attributs pourraient refléter la structure des ECG

- Notion de chaîne
- Notion d'arbre

# Classification

## Points importants

- Choix des objectifs (classes)
- Adéquation des attributs
- Redondance et robustesse des attributs

## Définitions et notations

- **Candidat :** élément à classifier
- **Attribut** $x_i$ **:** grandeur que l'on peut associer à un candidat
- **Vecteur d'attributs** $x$ **:** ensemble des attributs associés à un candidat
- **Classe ou hypothèse** $\omega_i,\ 1 \leq i \leq I$ **:** ensemble auquel peut appartenir tout candidat et qui influe sur la valeur de ses attributs
- **Ensemble d'apprentissage :** ensemble de candidats pour lesquels le résultat de la classification est connu.

# Approches statistiques

## Théorie de la décision

### Hypothèses

- Nombre de classes connu : $\Omega = \{\omega_1, \omega_2, \ldots \omega_I\}$
- Composition du vecteur d'attributs fixée
- Information disponible sur le comportement statistique des attributs en fonction la classe à laquelle ils appartiennent (peut être estimé empiriquement)

### Approche empirique : minimum de distance

- Information statistique disponible : moyenne $\boldsymbol{m}_i$ de chaque classe
- Pour tout candidat $\boldsymbol{x}$
  - calcul d'une distance $D_i(\boldsymbol{x}) = d(\boldsymbol{x}, \boldsymbol{m}_i)$
  - classification : minimum de distance

$$\text{Classe } \omega_{i_0} \quad : \quad \forall i \ ; \ D_{i_0}(\boldsymbol{x}) \leq D_i(\boldsymbol{x})$$

# Minimum de distance (1)

## Remarques

- $\boldsymbol{m}_i$ ; $1 \leq i \leq I$ : spécifiés à l'avance ou estimés empiriquement sur un ensemble d'apprentissage
- Choix de la distance $d(\cdot, \cdot)$ laissé à l'utilisateur
- $D_i(\boldsymbol{x})$ : fonction de décision
- Points de l'espace des attributs tels que $D_i(\boldsymbol{x}) = D_j(\boldsymbol{x})$

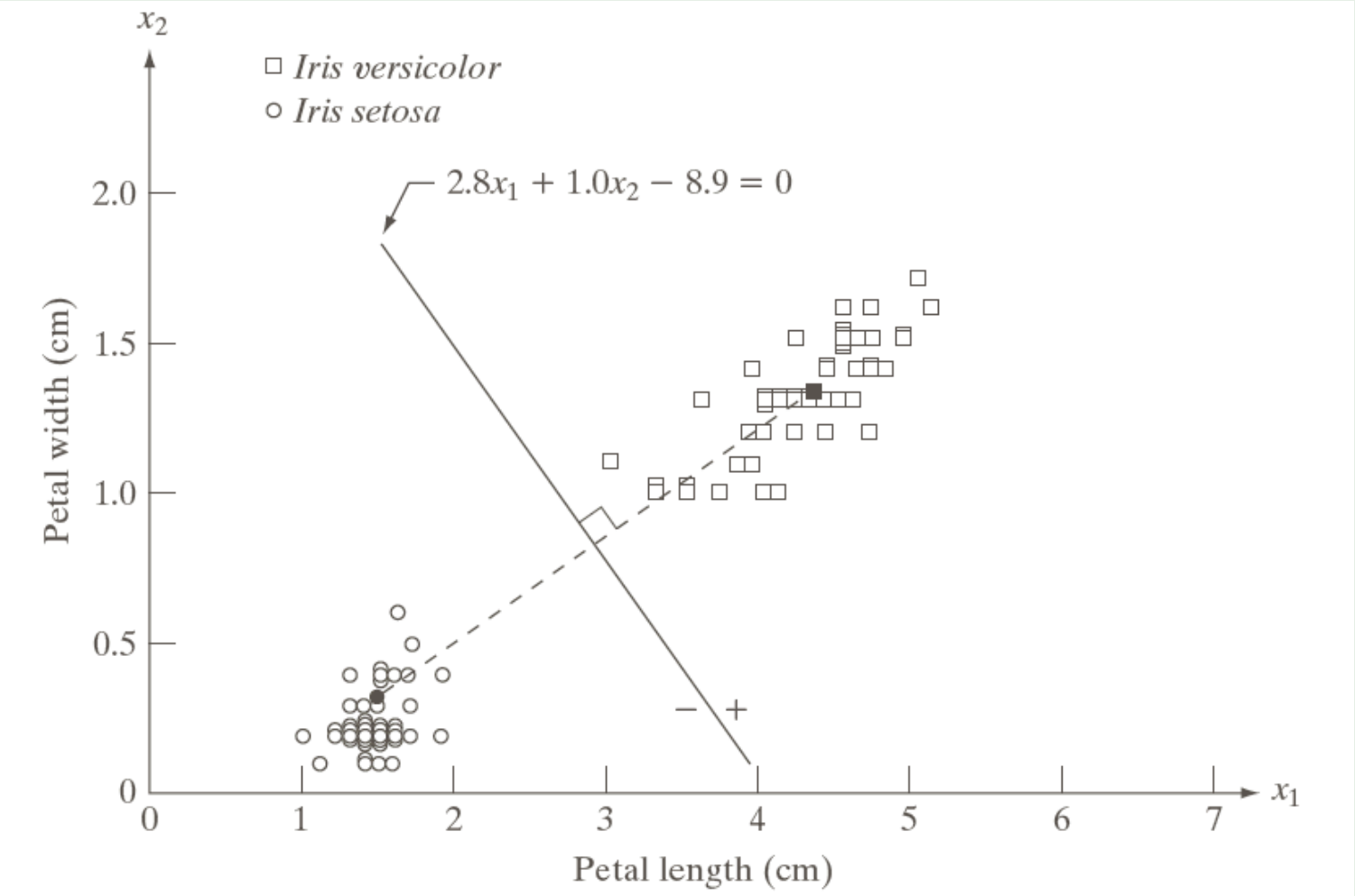
Frontières de décision

- Si $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$ (norme euclidienne) :

Les frontières de décision sont des hyperplans

# Minimum de distance (2)

## Illustration : exemple de Fisher



# Classification bayésienne

## Hypothèses

On connaît :

- $f(\boldsymbol{x}|\omega_i)$ : distribution des attributs dans chaque classe
- $f(\omega_i)$ : probabilité de chaque classe

## Approche par maximum *a posteriori*

- Recherche de l'indice $i$ de la classe qui maximise :
$$f(\omega_i|\boldsymbol{x}) \propto f(\boldsymbol{x}|\omega_i)f(\omega_i)$$
- $f(\omega_i|\boldsymbol{x})$ : fonctions de décision $D_i(\boldsymbol{x})$

## Limitation

- Le « prix à payer » pour une mauvaise classification varie
- Cet aspect n'est pas pris en compte

Classification à minimum de risque

# Classification à minimum de risque (1)

## Démarche

- Définition empirique du coût $C(\hat{\omega}_i) = d(\hat{\omega}_i, \omega_j)$ pour chaque couple $(i, j)$
- Risque bayésien : pour chaque classe $\omega_i$
$$\mathrm{E}\left[C(\hat{\omega}_i)|\boldsymbol{x}\right] = r_i(\boldsymbol{x}) = \sum_{j=1}^{I} d(\hat{\omega}_i, \omega_j) f(\boldsymbol{x}|\omega_j) f(\omega_j)$$
- $r_i(\boldsymbol{x})$ fonction de décision

## Principales caractéristiques

- Calcul aisé en général
- Spécification de $d(\hat{\omega}_i, \omega_j)$, $f(\boldsymbol{x}|\omega_j)$, $f(\omega_j)$ plus délicate

# Classification à minimum de risque (2)

## Cas particulier : $f(\boldsymbol{x}|\omega_j)$ gaussienne

- Coût $d(\hat{\omega}_i, \omega_j)$ quelconque : pas de simplification
- Coût $d(\hat{\omega}_i, \omega_j) = 1 - \delta_{ij}$ (classification MAP) :

$$r_i(\boldsymbol{x}) = \underbrace{f(\boldsymbol{x}|\omega_i)}_{\mathcal{N}(\boldsymbol{m}_i, \boldsymbol{R}_i)} \underbrace{f(\omega_i)}_{\text{scalaire}}$$

- Frontières de décision : quadriques
- Si, de plus, $\forall i : \boldsymbol{R}_i = \boldsymbol{I}$, alors $\log r_i(\boldsymbol{x}) \propto \boldsymbol{m}_i^{\mathrm{t}} \boldsymbol{x} - \frac{1}{2} \boldsymbol{m}_i^{\mathrm{t}} \boldsymbol{m}_i$

Frontières de décision : hyperplans
(classification par minimum de distance euclidienne)

# Estimation des caractéristiques des classes

## Position du problème

Comment spécifier $f(\boldsymbol{x}|\omega_i)$ et $f(\omega_i)$ ?

- $f(\omega_i)$ : peu de difficultés (connaissance du problème, ensemble d'apprentissage)
- $f(\boldsymbol{x}|\omega_i)$ : souvent délicat
  - approches non paramétriques (estimateurs à noyaux)
  - approches paramétriques (souvent empiriques)
  - exemple : loi gaussienne multivariée