Traitement de données

E2I5

Année 2024-2025

nathalie.guyader@grenoble-inp.fr

Jusqu’à présent…

- Séance 1: analyse en composantes principales
  - TD (données Notes et Criminalité)
- Séance 2: analyse linéaire discriminante (approche descriptive)
  - TD (données Fisher et mesures de défauts sur des plaques de silicium)
- Séance 3: courbes ROC
  - TD (données «composants_electroniques.xlsx»)

Plan de la séance

Retours sur le dernier TP (courbes ROC)

Test

Cours en autonomie

TP

L'analyse discriminante peut être vue selon 2 approches:

- **une technique descriptive**: l'objectif est de proposer un nouveau système de représentation, des variables latentes formées à partir de combinaisons linéaires des variables prédictives, qui permettent de discerner le plus possible les groupes d'individus. En ce sens, elle se rapproche de l'analyse factorielle (ACP) car elle permet de proposer une représentation simplifiée (éventuellement graphique) dans un espace réduit (Cf. Chapitre 4)

- **une technique prédictive**: il s'agit dans ce cas de construire une fonction de classement (règle d'affectation, ...) qui permet de prédire le groupe d'appartenance d'un individu à partir des valeurs prises par les variables prédictives. En ce sens, cette technique se rapproche des **techniques supervisées en apprentissage automatique (Machine Learning).**

Soit une matrice X des données : *n* individus (observations) décrits par *p* variables et appartenant à K groupes (classes) différentes)

$$X = \begin{pmatrix} x_{11} & & x_{1p} \\ & \dots\dots & \\ x_{n1} & \dots\dots & x_{np} \end{pmatrix}, \qquad \text{Classe} = \begin{pmatrix} classe\ 1 \\ classe\ 2 \\ classe\ 1 \\ classe\ 3 \\ \dots \\ \dots \\ classe\ k \end{pmatrix}$$

k: nombre de classes étudiées (k<<n)

L'analyse discriminante s'envisage selon 2 axes:

- Une approche descriptive (vue au chapitre précédent-CM3 et au TD3)
- Une approche prédictive basée sur le cadre probabiliste

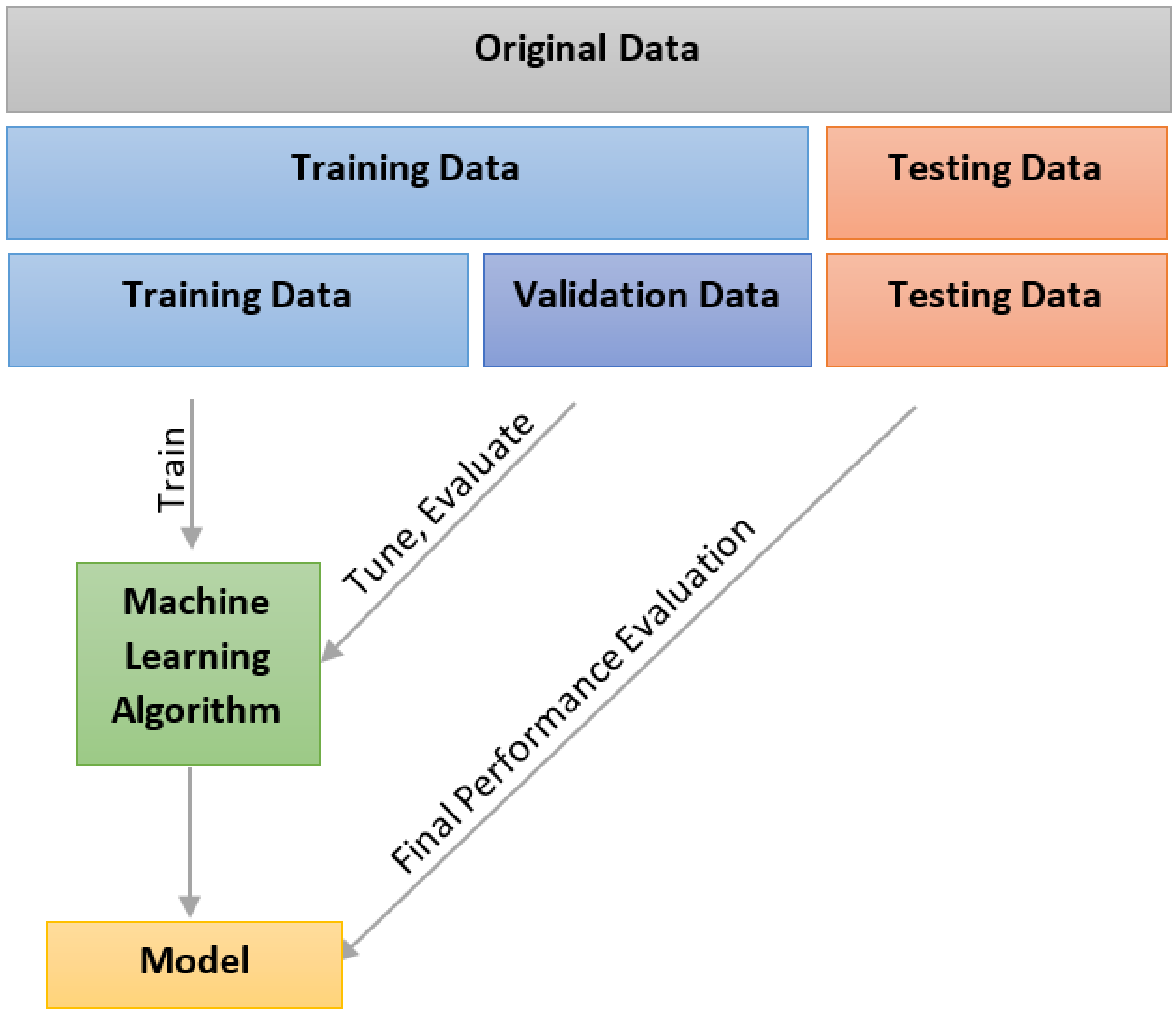
Pour la modélisation, on dispose de p variables dites prédictives (ou explicatives).

On utilise alors la probabilité a priori d'une classe et on cherche la probabilité d'appartenance à une classe sachant une observation; cette probabilité est appelée probabilité a posteriori.
Cette approche est basée sur les probabilités conditionnelles et la règle de Bayes. On parle également de classifier de Bayes ; nous n'aurons pas le temps de rentrer dans les détails de la méthode mais nous utiliserons les classifiers de Bayes dans le projet 1.

Données d'apprentissage (training):
L'approche utilise des données appelées d'apprentissage pour « apprendre » les différentes classes. Dans le cas des classifiers bayésiens, chaque classe est modélisée par une loi normale multidimensionnelle (ou normale multivariée ou loi multinormale).

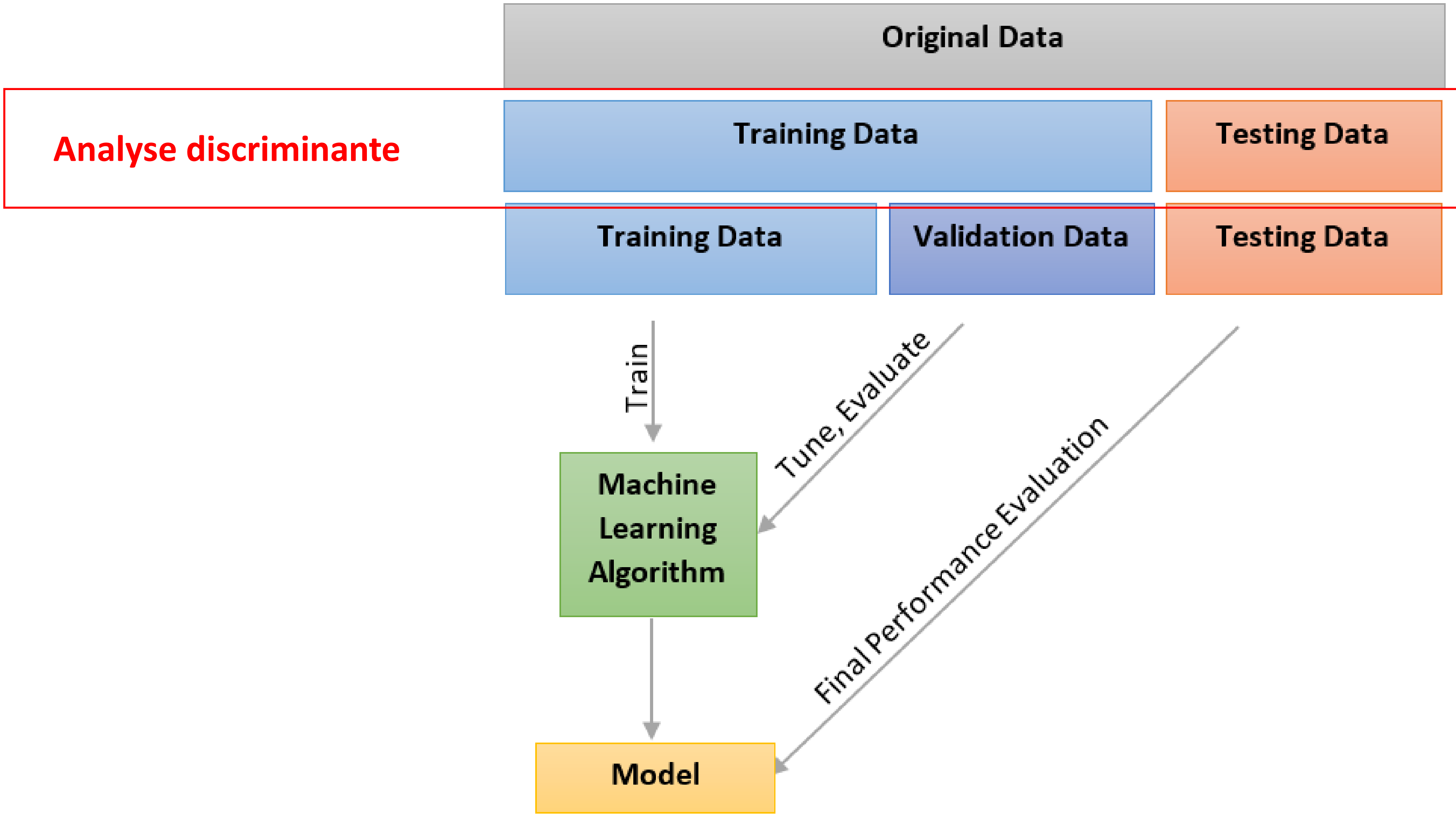
Données de test (testing):
On calcule pour chaque nouvelle observation la probabilité de chaque classe sachant la nouvelle observation et on attribue à l'observation la classe qui a la plus grande probabilité (règle du maximum a posteriori).

Base d'apprentissage, base de test



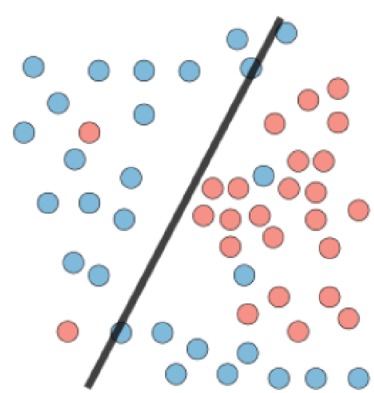
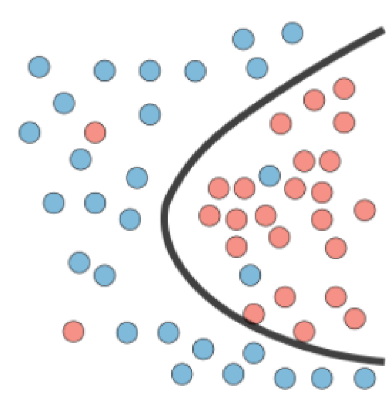
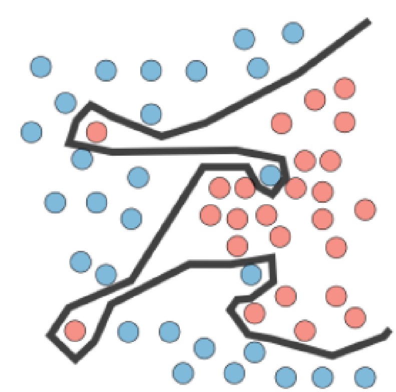
https://stanford.edu/~shervine/l/fr/teaching/cs-229/pense-bete-machine-learning-petites-astuces

Base d'apprentissage, base de test

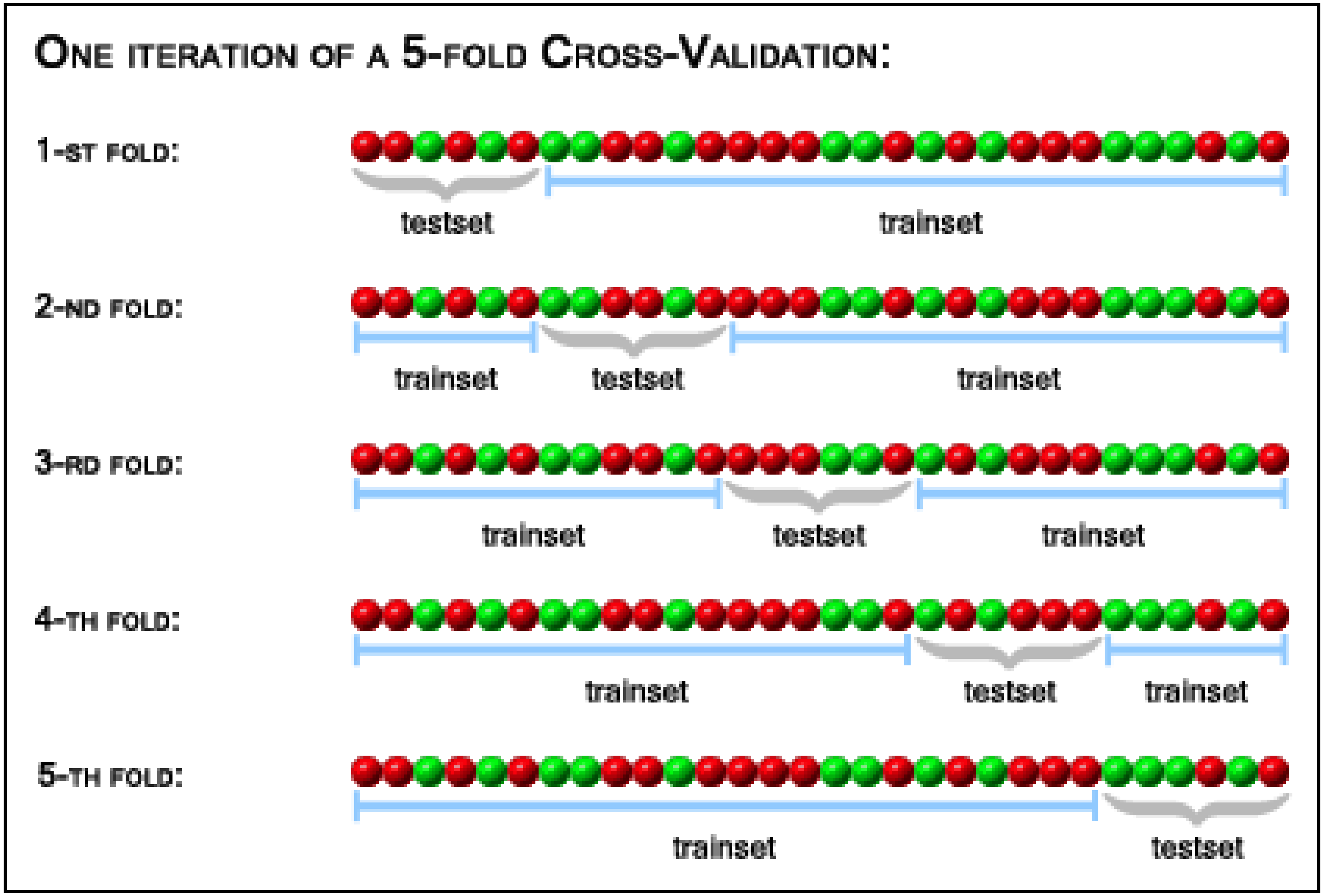


https://stanford.edu/~shervine/l/fr/teaching/cs-229/pense-bete-machine-learning-petites-astuces

Apprentissage, sur-apprentissage

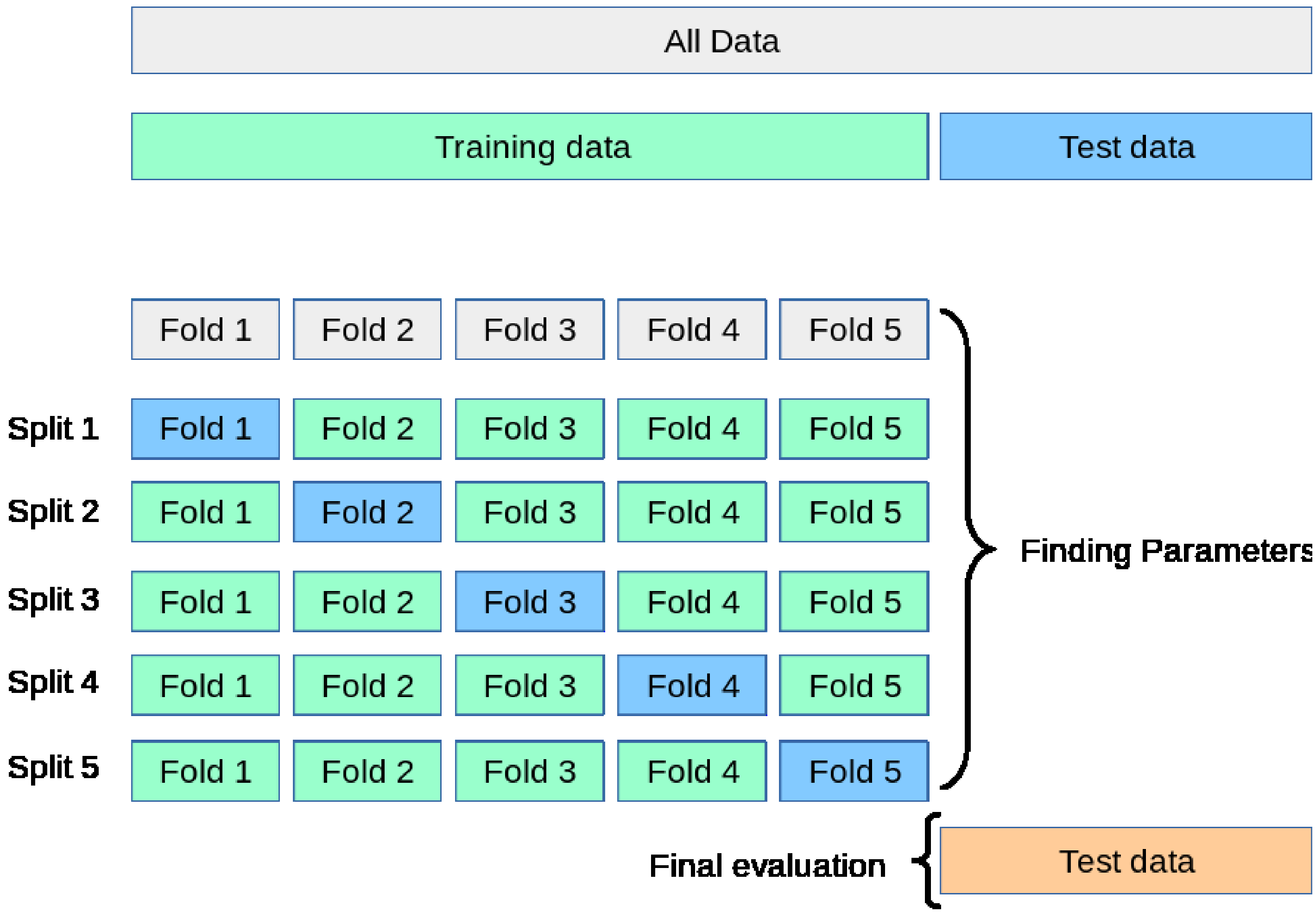
	Underfitting	Just right	Overfitting
Symptômes	• Erreur d'entrainement élevé • Erreur d'entrainement proche de l'erreur de test • Biais élevé	• Erreur d'entrainement légèrement inférieure à l'erreur de test	• Erreur d'entrainement très faible • Erreur d'entrainement beaucoup plus faible que l'erreur de test • Variance élevée
Illustration dans le cas de la classification			

Validation croisée



https://genome.tugraz.at/proclassify/help/pages/XV.html

Validation croisée pour « fitter » les paramètres du modèle



https://scikit-learn.org/stable/modules/cross_validation.html

Chapitre 6 – Bayes et kppv

- Classifiers avec hypothèse de classes gaussiennes
- k plus proches voisins

Les transparents suivants sont repris du cours de S. Charbonnier – Machine Learning IESE5

Rappel : ddp – loi normale

Densité de probabilité monodimentionnelle – loi normale : x, un scalaire, m, moyenne, σ^2 variance:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}\right)$$

Densité de probabilité multidimentionnelle – loi normale (X vecteur de dimension d), m moyenne, Σ matrice de covariance

$$p(X) = \frac{1}{(2\pi)^{d/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(X-m)^T\Sigma^{-1}(X-m)\right)$$

Règle de Bayes

Soient A et B, deux évènements.

La règle de Bayes exprime que :

$$P(A/B) = \frac{P(B/A) \times P(A)}{P(B)}$$

Règle de Bayes

Soit X, un vecteur de d caractéristiques

Soit ω_j, une classe parmi M classes possibles

$$P(\omega_j / X) = \frac{p(X / \omega_j).P(\omega_j)}{p(X)} \quad avec \;\; \mathrm{p(X)} = \sum_{i=1}^{M} p(X / \omega_i) P(\omega_i)$$

$\boldsymbol{P(\omega_i/X)}$: probabilité d'avoir la classe ω_j sachant le vecteur X – Probabilité a posteriori

$\boldsymbol{p(X/\omega_i)}$: densité de probabilité du vecteur X sachant la classe ω_i

$\boldsymbol{P(\omega_i)}$: probabilité d'avoir la classe ω_i – Probabilité a priori

$\boldsymbol{p(X)}$: densité de probabilité du vecteur X

Règle du maximum a posteriori

Soit une fonction de décision D:

$$R^d \to Z^M$$
$$X \to D(X) = \omega_i \text{ avec } \omega_i \in \{\omega 1 , \omega_2, \dots ,\omega_{\mathrm{M}}\}$$

La règle du maximum a posteriori s'écrit:

$$D(X) = \omega_i \quad \text{si} \quad P(\omega_i/X) = \max_{j=1 \text{ à } M} P\,(\omega_j/X)$$

Ou, d'après la règle de Bayes

$$D(X) = \omega_i \quad \text{si} \max_{j=1 \text{ à } M} p\,(X/\omega_j).\,P(\omega_j)$$

<u>Règle de décision :</u>

Le vecteur forme X est attribué à la classe la plus probable, c'est-à-dire celle pour qui $P(X/\omega_j)$) ou $p(X/\omega_j)$ $P(\omega_j)$ est la plus grande.

Règle du maximum de vraisemblance

- Si les **P(ω_i) sont inconnues**, on considère chaque classe comme équiprobable : P(ω_i) =1/M

$$D(X) = \omega_i \quad \text{si}$$

$$P(\omega_i / X) = \max_{j=1 \text{ à } M} P(\omega_j / X) = \max_{j=1 \text{ à } M} p(X / \omega_j).P(\omega_i)$$

devient:

$$D(X) = \omega_i \quad \text{si} \quad p(X / \omega_i) = \max_{j=1 \text{ à } M} p(X / \omega_j)$$

<u>Règle de décision :</u>

Le vecteur forme X est attribué à la classe la plus vraisemblable c'est-à-dire celle pour qui $p(X/\omega_j)$ est la plus grande

Apprentissage de D

Règle de décision :
Le vecteur forme X est attribué à la classe la plus probable, c'est à dire celle qui maximise $P(\omega_j/X)$ ou $p(X/\omega_j).P(\omega_j)$

Apprentissage de la fonction de décision

- Ce que l'on connaît à priori

$$\mathbf{P(\omega_j)}$$

- Ce que l'on apprend à partir des données

$$\mathbf{p(X/\omega_j)}$$

→ On en déduit $\mathbf{P(\omega_j/X)}$

Apprentissage de D(X) : estimation de $p(X/\omega_j)$ pour tout $\omega_j \in \{\omega 1, \omega_2, \ldots, \omega_M\}$ à l'aide d'une base d'apprentissage

« Apprentissage par estimation de densité de probabilité »

Hypothèse de classes gaussiennes

Avec l'hypothèse, on écrit:

$$p(X/\omega_j) = \frac{1}{(2\pi)^{d/2}(\det \boxed{\Sigma_j})^{1/2}} \exp(-\frac{1}{2}(X - \boxed{m_j})^T \boxed{\Sigma_j^{-1}} (X - \boxed{m_j}))$$

On dispose d'une base d'apprentissage composée de N exemples X_k dont n_j exemples de classe ω_j

Apprentissage de : $p(X/\omega_j)$: on estime **pour chaque classe** ω_j

Le centre de gravité: m_j

La matrice de variance-covariance: Σ_j

Règle de décision:

$$D(X) = \omega_i \quad \text{si} \quad P(\omega_i / X) = \max_{j=1 \text{ à } M} P(\omega_j / X) = \max_{j=1 \text{ à } M} P(X/\omega_j).P(\omega_i)$$

→ ***Classifieur quadratique***

Hypothèses simplificatrices

- Supposer que les Σ_i sont identiques

  Regrouper tous les points pour estimer S

  → ***Classifieur linéaire (ALD)***

- Supposer que les Σ_i sont diagonales ~ les caractéristiques sont décorrélées

  Estimer uniquement la variance des caractéristiques pour chaque classe

  → ***Classifieur bayésien naïf***

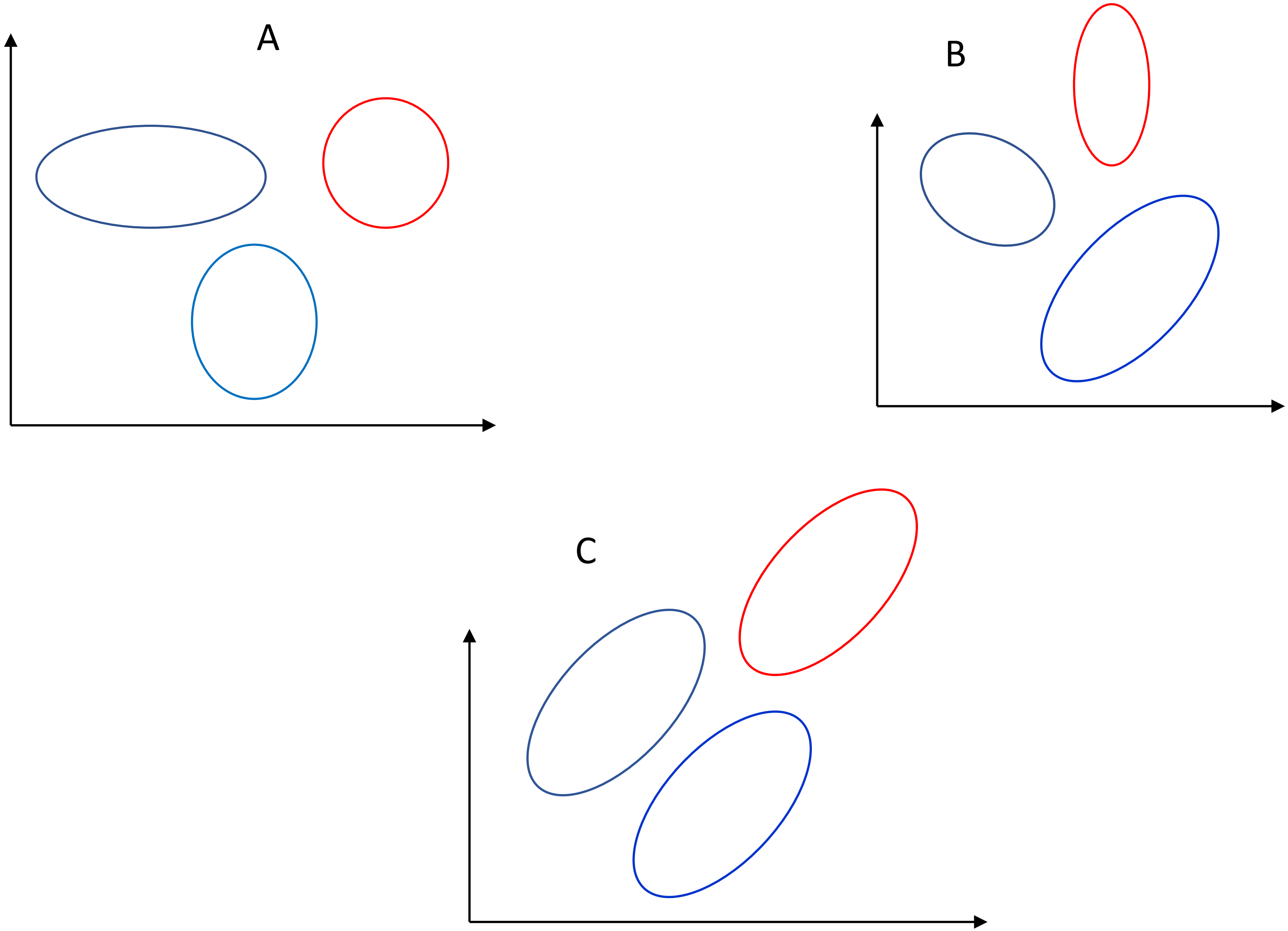
Méthodes - hypothèse gaussienne

	Classifieur quadratique	Classifieur linéaire	Classifieur bayesien naif
Matrice de covariance	Σ_i Différente pour chaque classe	Σ Identique pour toutes les classes	Σ_i Différente pour chaque classe Diagonale
Frontières	Quadratiques	Linéaires	Quadratiques

ALD

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis,
QuadraticDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
```

Formes des classes

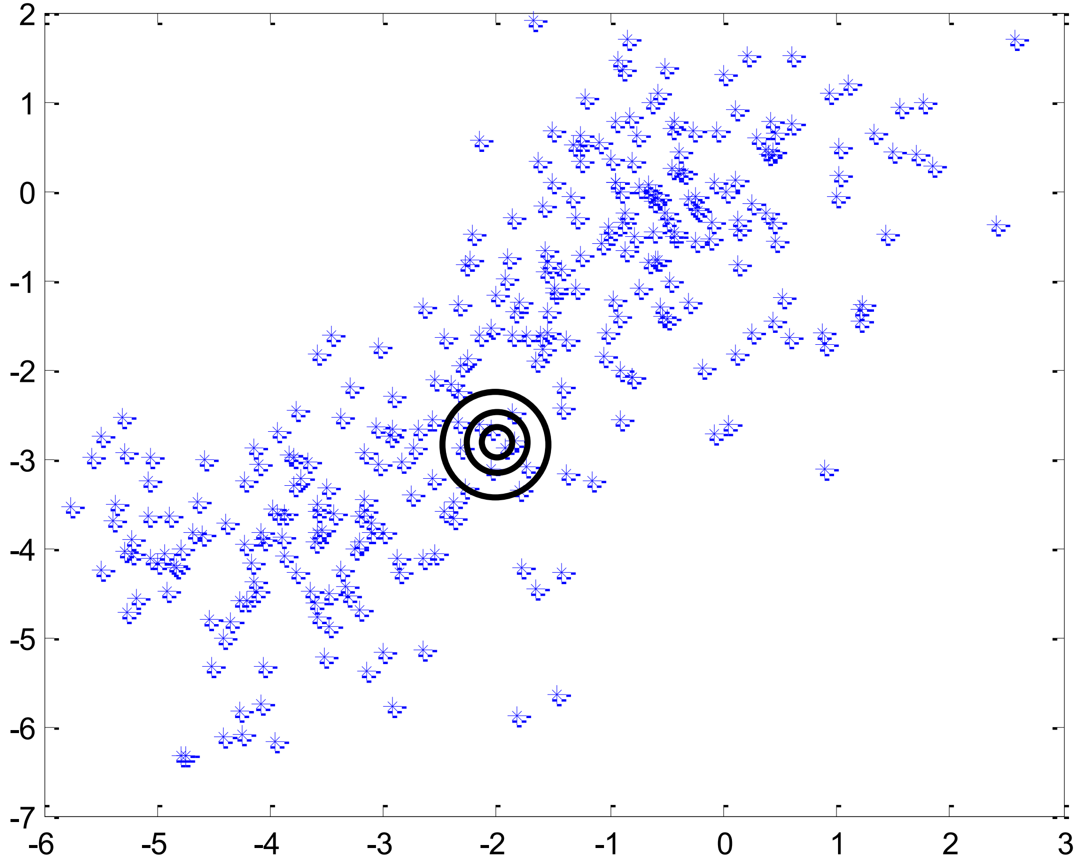


Que faire quand l'hypothèse des classes gaussiennes ne peut pas être faite ?

Estimer la densité de probabilité à partir des données

Méthode des k plus proches voisins

Les k plus proches voisins



Méthode des kppV :

On fixe un nombre de points k et on estime le volume qu'ils occupent

Les kppV en pratique

Soit X le point à classer

1. Chercher les k plus proches voisins de X dans la base d’apprentissage
2. Affecter X à la classe majoritaire parmi ses voisins

« Qui se ressemble s’assemble »

Mise en œuvre

– Calculer la distance de X à tous les points de la base d’apprentissage:

$$D(X, Xi) = (X - Xi)^T \Sigma^{-1} (X - Xi)$$

– Classer les points des plus proches vers les plus éloignés
– Sélectionner les k premiers
– Affecter la classe majoritaire

La décision est d’autant plus longue que le nombre d’exemples de la base d’apprentissage est important

Rq: si distance euclidienne réduire les données !

from sklearn.neighbors import KNeighborsClassifier

Projet 1

- Durant ce projet vous devrez tester sur un jeu de données réelles les classifiers:
  - Linéaire
  - Quadratique
  - Bayes naïf

  - K ppv