Science des données III : module 1

Classification supervisée (partie 1)

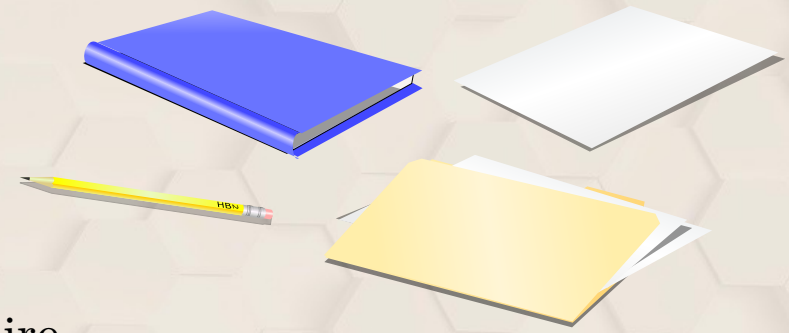
Philippe Grosjean & Guyliann Engels

Université de Mons, Belgique
Laboratoire d'Écologie numérique des Milieux aquatiques

http://biodatascience-course.sciviews.org
sdd@sciviews.org

UMONS

Objectifs du cours



- Découvrir la classification supervisée
- Savoir utiliser l'Analyse Discriminante Linéaire
- Calculer une matrice de confusions et des métriques dérivées pour quantifier les performances d'un outil de classification

UMONS

Classification supervisée *versus* non supervisée

- La **classification** sert à regrouper les individus d'un jeu de données en différents groupes ou **classes**
- La classification *non* supervisée permet de choisir les classes librement (ex.: la classification hiérarchique et le dendrogramme)
- La **classification supervisée** permet d'utiliser un ordinateur pour apprendre à classer des objets selon nos propres objectifs ("machine learning" en anglais)
- Ces techniques sont très utilisées en fouille des données ("data mining"), en 'omics et en intelligence artificielle. **Permet d'automatiser la classification d'un très grand nombre d'items**. Ex.: pages Web.

UMONS

Cadre général - 3 phases

Pour chaque item à classer, plusieurs variables sont mesurées (= **attributs**). Ces mesures peuvent être réalisées facilement sur tous les items à classer ultérieurement.

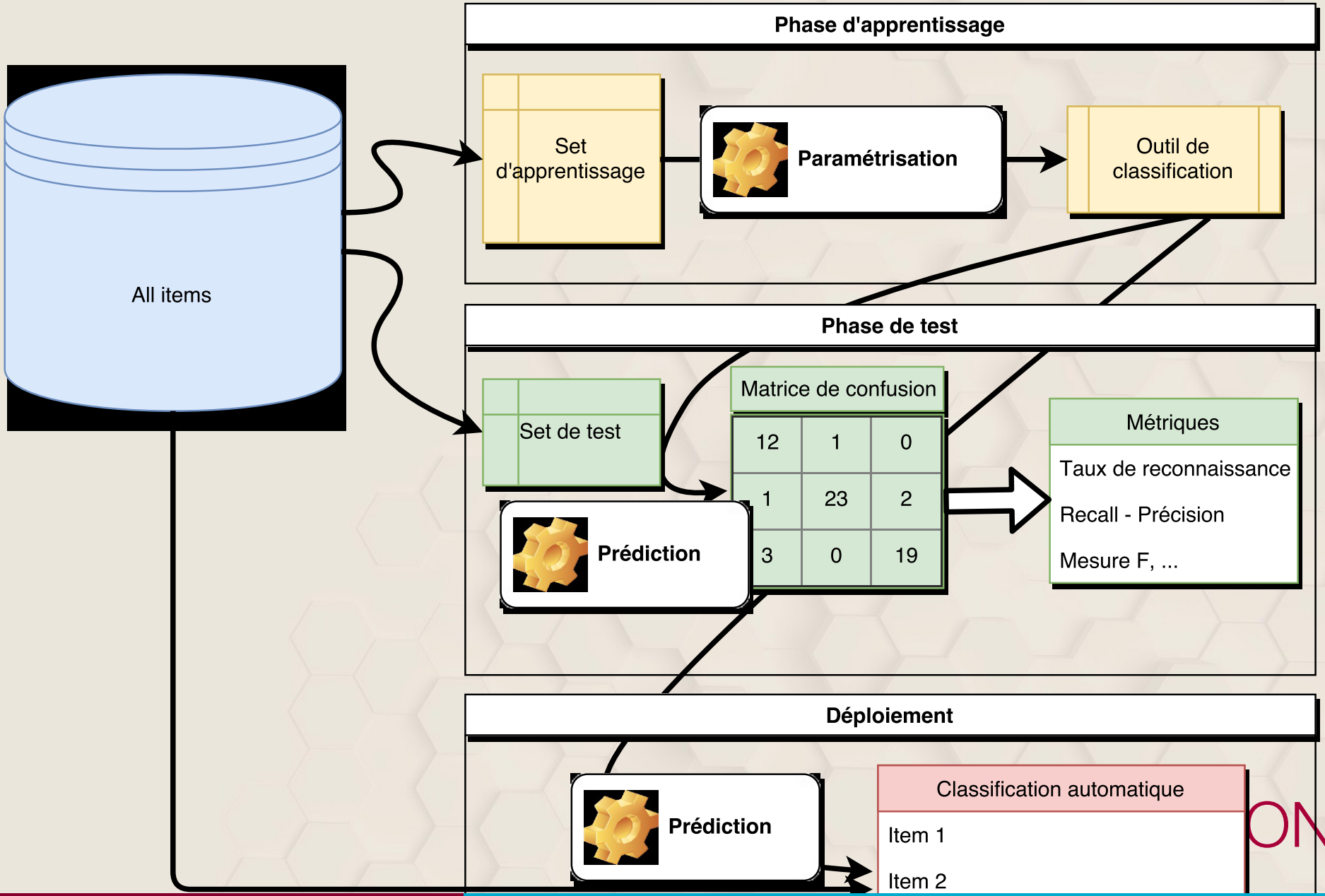
Un sous-ensemble représentatif d'items est **classé manuellement**. Ce sous-ensemble est ensuite divisé aléatoirement en **set d'apprentissage** et **set de test**.

Le travail se réalise en trois phases:

1. **Apprentissage** : un algorithme est "entrainé" (paramétré) pour classer les items sur base du **set d'apprentissage**.
2. **Test** : les performances de l'outil de classification sont vérifiées à l'aide du **set de test**.
3. **Déploiement** : si les perfomances sont satisfaisantes, l'outil de classification est ensuite utilisé pour classer **automatiquement** tous les autres items; une classification manuelle n'est plus nécessaire.

UMONS

Processus


Phase d'apprentissage
All items
Set d'apprentissage
Paramétrisation
Outil de classification
Phase de test
Set de test
Matrice de confusion
12 1 0
1 23 2
3 0 19
Métriques
Taux de reconnaissance
Recall - Précision
Mesure F, ...
Prédiction
Déploiement
Prédiction
Classification automatique
Item 1
Item 2

Conditions d'application

- Tous les groupes sont connus et disjoints
- La classification manuelle est réalisée sans erreur
- Les mesures utilisées sont suffisamment discriminantes
- Toute la variabilité est représentée dans le set d'apprentissage
- Le système est statique: pas de changement des caractéristiques des items à classer

UMONS

Evaluation des performances de classification

- L'évaluation doit **toujours** se faire sur un échantillon indépendant du set d'apprentissage = **set de test**. Sinon les résultats sont biaisés
- L'outil de base est une **tableau de contingence à double entrée** croisant la classification manuelle et la classification automatique, appelé aussi **matrice de confusion**

	Espèce 1	**Espèce 2**	**Espèce 3**	**Espèce 4**
Espèce 1	correct	erreur	erreur	erreur
Espèce 2	erreur	correct	erreur	erreur
Espèce 3	erreur	erreur	correct	erreur
Espèce 4	erreur	erreur	erreur	correct

UMONS

Exemple de matrice de confusion

```
manuel      <- c("A", "B", "C", "A", "A", "C", "B", "C", "B", "B")
automatique <- c("B", "B", "C", "C", "A", "C", "B", "C", "A", "B")
table(manuel, automatique)
```

```
##       automatique
## manuel A B C
##      A 1 1 1
##      B 1 3 0
##      C 0 0 3
```

UMONS

Métriques calculées sur la matrice de confusion

(voir document en annexe)

Métriques les plus importantes:

- Taux de reconnaissance global *versus* erreur globale
- Recall + précision ou spécificité (par classe)
- Mesure F, balanced accuracy

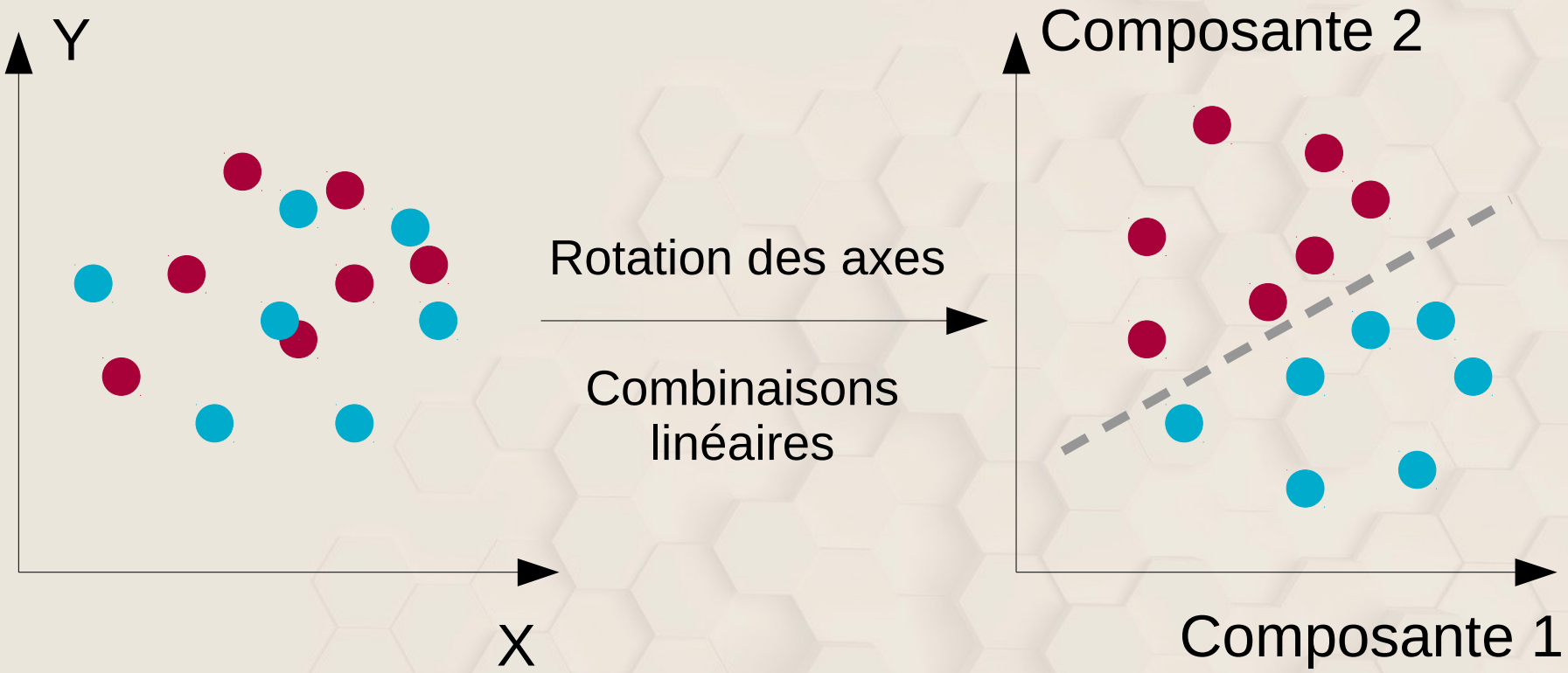
Calcul à la main et dans R

Sur base des formules, calculons ces différentes métriques sur base de la matrice de confusion d'exemple (voir plus haut)

UMONS

Analyse Discriminante Linéaire (ADL)

Même principe que l'ACP : rotation du système d'axes des attributs, mais avec un objectif différent, de séparer les différentes classes au mieux.



Application sur `iris`

Calcul complet pour séparer les fleurs d'iris grâce à l'ADL dans R