

Science des données III



Machine learning

Guyliann Engels & Philippe Grosjean

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



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

Matrice de confusion

La matrice de confusion est un outil informant sur la qualité d'un algorithme de classification supervisée. Il s'agit d'un tableau à double entrée comparant les prédictions de l'algorithme par rapport à une classification manuelle.

	Orange	Mandora	Mandarine
Orange	25	0	0
Mandora	0	16	8
Mandarine	1	9	18

sur cette matrice de confusion simplifiée, calculez : **Le nombre de vrai positif, vrai négatif, faux positif, faux négatif, le taux de reconnaissance global (accuracy), taux de vrai positif (recall), spécificité (specificity), précision, le score F (F-measure) et taux de reconnaissance balancé (BalAcc)** en vous ciblant sur les Mandora (On retrouve en ligne la classification manuelle)

Les formules sont disponibles dans le document annexe.

Les organismes planctoniques de la mer du Nord

Regroupement des organismes animaux et végétaux dont la capacité de mouvements horizontaux est inférieure à celles des mouvements de l'eau. Ils ne peuvent pas nager à contre courant, par exemple.



- Jeu de données : **plankton_set.csv**

Les organismes planctoniques de la baie de Calvi

Le jeu de données mis à votre disposition est le jeu de données :

- **plancton_calvi.csv**

Il est plus important que le premier et comprend plus de 10000 observations