Présentation du logiciel

Orange Data Mining

Pierre **LEJEAIL**

Paul-Nam **LUCIARDI**

Romain **RIGAL**

David **VAUCLIN**

Plan

- Introduction
- Installation, bibliothèques nécessaires
- Méthodes statistiques proposées
- Points forts et points faibles
- Positionnement par rapport aux autres outils existants
- Conclusion

Introduction

- Open source et gratuit
- Data Mining / Data visualisation
- Elaboré pour tout type d’utilisateur
- Disponible sous Windows, Mac OS et Linux
- Créé en 1997 par Université de Ljubljana, Slovénie
- Version actuelle : 3.3

Installation, bibliothèques nécessaires

- Installation intuitive
- Nécessite Python (inclus dans le “package” d’installation)
- Ajout et création d’add-ons
- Plus de 100 widgets disponibles

Méthodes statistiques proposées

Catégorie	Méthodes
Données	Importation, filtrage, échantillonnage, statistiques descriptives, traitement des valeurs manquantes ...
Visualisation	Graphiques classiques : histogramme, nuages de points, boxplot Visualisation multivariée : mosaïque, diagramme de Venn, arbre de Pythagore ...
Classification supervisée	Classifieur Bayesien, régression logistique, arbre de classification, K-Plus proches voisins, random forest, SVM, AdaBoost, algorithme CN2 (induction de règles prédictives)
Régression	Régression linéaire, régression par les k-NN, régression avec random forest, régression SVM, arbre de régression, gradient stochastique
Apprentissage non-supervisé	ACP, AFC, calcul de distance, CAH, K-Means, multidimensional scaling
Evaluation	Courbe de gain, matrice de confusion, courbe ROC

Fonctionnalités additionnelles

- Add-ons:
  - **Associate**: widgets pour les règles d'association
  - **Bioinformatics**: analyse de données génétiques, enrichissement
  - **Data fusion**: concaténation de plusieurs tableau de données, factorisation de matrices
  - **Educational**: widgets pour l'enseignement des méthodes de machine learning (k-means, régression polynomiale, gradient stochastique …)
  - **Image analytics**: analyse d'images
  - **Network**: analyse des réseaux sociaux
  - **Text mining**: widgets pour le text-mining, traitement automatique du langage naturel
  - **Time series**: modélisation et analyse des séries temporelles

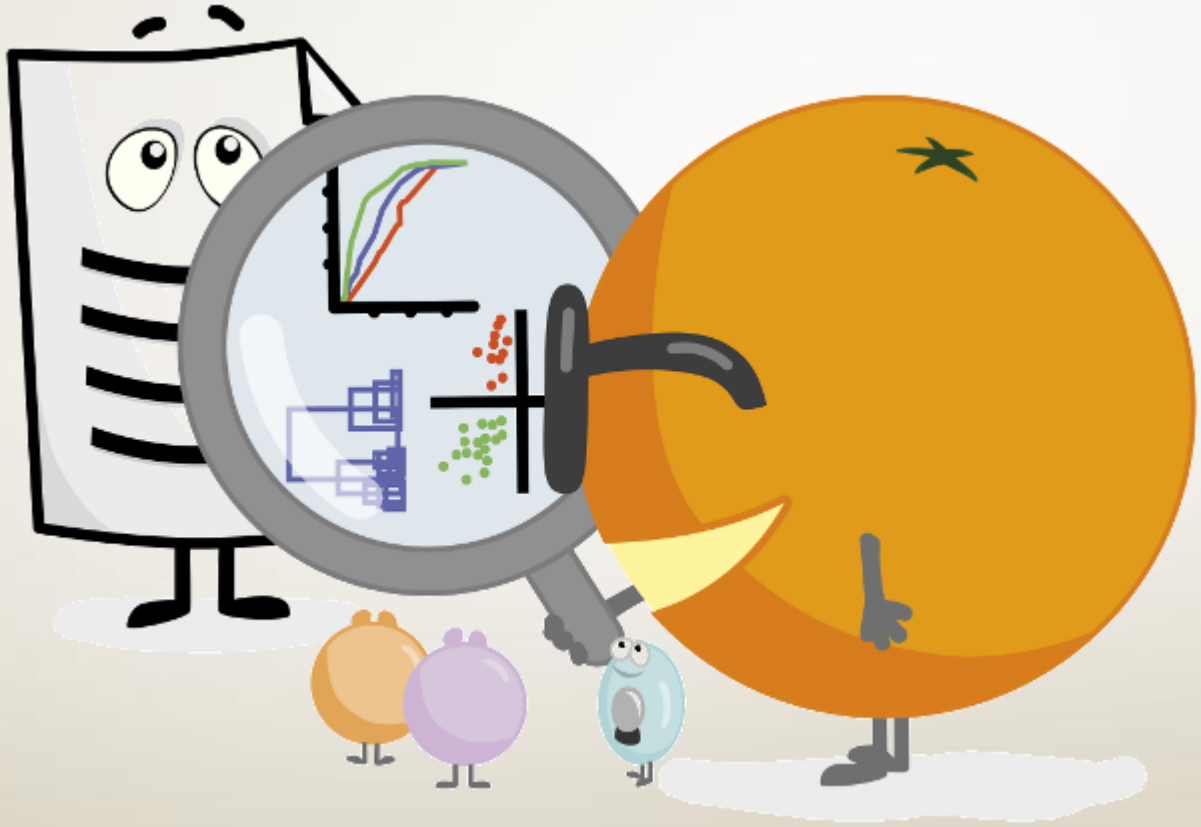
Points forts, points faibles

- Avantages :
  - Totalement gratuit
  - Prise en main intuitive
  - Possibilité d'effectuer ses analyses via des scripts en Python
  - Accessible aussi bien pour les novices que les experts
  - Visualisation interactive des données
- Limites :
  - Fonctionnalités de reporting/statistiques inexistantes
  - Analyses factorielles limitées
  - Obligation de passer par des scripts pour des analyses plus poussées
  - Erreurs inopinées

Positionnement par rapport aux autres outils

Catégories	Méthodes	RapidMiner	R	Weka	Orange	Knime
Import et transformation	textual files (.txt, .csv)	+	+	+	+	+
	specific input format files	+	+	+	+	+
	Excel/ spreadsheet	+	+	-	-	+
	discretization	+	+	+	+	+
	normalization	+	+	+	+	+
	ACP	+	+	+	+ / -	+
Arbres de Décision	C4.5	+	+	+	+	-
	CART	+	+	+	+	+
Regles de Classification	1Rule, PART, RIPPER	+	+	+	-	+
Reseaux Bayesiens	Naive Bayes	+	Naive Bayes	+	Naive Bayes	+
Instance Based Learning	K + proche voisin	+	+	+	+	+
Function Based Learning	Analyse de la régression	log, lin, poly	lin...nonlin	log,lin	log, lin, lasso, PLS, trees, mean	log, lin, poly, trees
Apprentissage	SVM	+	+	+	+	+
	AdaBoost	+	a	+	+	a
	Random Forest	+	a	+	+	+
	Autres méthodes	Rotation forest, LogitBoost, Option tree, Stacking, Bayesian boosting	Rotation forest, Randomize Tree, LogitBoost, Stacking, Boosted Regression	Rotation forest, LogitBoost, Option tree, Stacking, MultiBoost	-	Rotation forest, LogitBoost, Option tree, Stacking, Bayesian boosting
Classification Hierarchique	Linkage Based	+	+	+	+	+
Centroid	K Means	+	+	+	+	+
	Fuzzy C-means	-	a	-	a	+
Distribution basé sur Clustering	EM clustering	+	+	+	-	+
Densité basé sur Clustering	DBScan	+	+	+	-	+
Non Supervisé	Regles d'associations	GSP, Apriori,FP-Growth,Tertius	Apriori, Eclat, Tertius	GSP, Apriori, tertius	Apriori	GSP, Apriori,Fpgrowth, Tertius
Méthodes d'évaluation	Cross validation	+	+	+	+	+
	Matrice de confusion	+	+	+	+/-	+
	Courbe lift et Roc	+	+	+	+	+

Démonstration



En conclusion

- Interface simple et ergonomique
- Très orienté data mining
- Certaines manipulations pas intuitives
- Accessible pour les statisticiens non-informaticiens
- Mais offre aussi des possibilités pour les programmeurs

Merci pour votre attention.
A vous d’extraire le jus d’Orange !