

Méthodes d'optimisation pour l'aide à la décision

1. *Data Science & Machine Learning*

Solen Quiniou

`solen.quiniou@univ-nantes.fr`

IUT de Nantes

Année 2023-2024 – Info 3 (Semestre 5)

[Mise à jour du 6 septembre 2023]



Plan du cours

- 1 Introduction à la science des données
- 2 Introduction à l'apprentissage automatique
- 3 Apprentissage supervisé (classification)
- 4 Références

Plan du cours

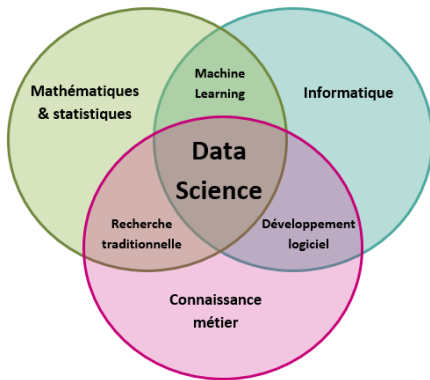
- 1 Introduction à la science des données
 - Introduction
 - Métiers liés à la science des données
- 2 Introduction à l'apprentissage automatique
 - Introduction
 - Préparation des données
 - Choix du modèle d'apprentissage
- 3 Apprentissage supervisé (classification)
 - Introduction
 - Évaluation des classifieurs
 - k plus proches voisins
- 4 Références

Plan du cours

- 1 Introduction à la science des données
 - Introduction
 - Métiers liés à la science des données
- 2 Introduction à l'apprentissage automatique
 - Introduction
 - Préparation des données
 - Choix du modèle d'apprentissage
- 3 Apprentissage supervisé (classification)
 - Introduction
 - Évaluation des classifieurs
 - k plus proches voisins
- 4 Références

Introduction à la science des données (1)

- **Science des données** (*Data Science*) : domaine interdisciplinaire (informatique, statistiques, *machine learning*, analyse des données, mathématiques décisionnelles) dont le but est d'**analyser des quantités importantes de données** afin d'en extraire des connaissances



Définition et image tirées de

<https://www.eurodecision.com/algorithmes/data-science>

Introduction à la science des données (2)

● Pourquoi la science des données ?

- ▶ Augmentation des volumes de données stockés par les entreprises
 - ▶ Disponibilité de données publiques en grande quantités
 - ▶ Possibilité technique de traiter efficacement ces données avec des langages de programmation
- Discipline très récente et en plein développement depuis ces dernières années

● Exemples d'applications

- ▶ **Industrie** : maintenance préventive
- ▶ **Banques et assurances** : automatisation de processus, connaissance client, réduction du taux d'attrition
- ▶ **Santé** : épidémiologie, toxicologie, recherches
- ▶ **Retail** : prévision des ventes, marketing prédictif
- ▶ **Transport et villes** : villes intelligentes, optimisation des transports en fonction des flux de voyageurs
- ▶ ...

Introduction à la science des données (3)

La Data Science par DataScientest

C'est quoi ?

Expansion de la Data

2013 : 4 ZB*



2020 : 44 ZB*



*1 ZB = 1 000 000 000 000 Gigabytes

Objets Connectés



Pourquoi ?

Réseaux Sociaux



À quoi ça sert ?

Le Pétrole du XXIème



Moteurs de Recherche

Comment faire ?



Création & Innovation



Comment fonctionne la Data Science ?

5 étapes

1. Collecte de Données



2. Stockage de Données



3. Data Mining



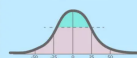
4. Analyse



5. Data Viz

Outils & Technique

Les Statistiques



L'Intelligence Artificielle



Les Mathématiques



Concrètement ça donne quoi ?

Cas d'usage et Applications

Cyber Sécurité



Reconnaissance Faciale



Voitures Autonomes



Médecine Intelligente



Optimisation d'itinéraire



Ah ok merci !

Plan du cours

- 1 Introduction à la science des données
 - Introduction
 - **Métiers liés à la science des données**
- 2 Introduction à l'apprentissage automatique
 - Introduction
 - Préparation des données
 - Choix du modèle d'apprentissage
- 3 Apprentissage supervisé (classification)
 - Introduction
 - Évaluation des classifieurs
 - k plus proches voisins
- 4 Références

Métiers liés à la science des données



Data Engineer	Data Analyst	Data Scientist	Machine Learning Scientist
Store and maintain data	Visualize and describe data	Gain insights from data	Predict with data
SQL + Java/Scala/Python	SQL + BI Tools + Spreadsheets	Python/R	Python/R

Qu'est-ce qu'un *Data Scientist*?

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative



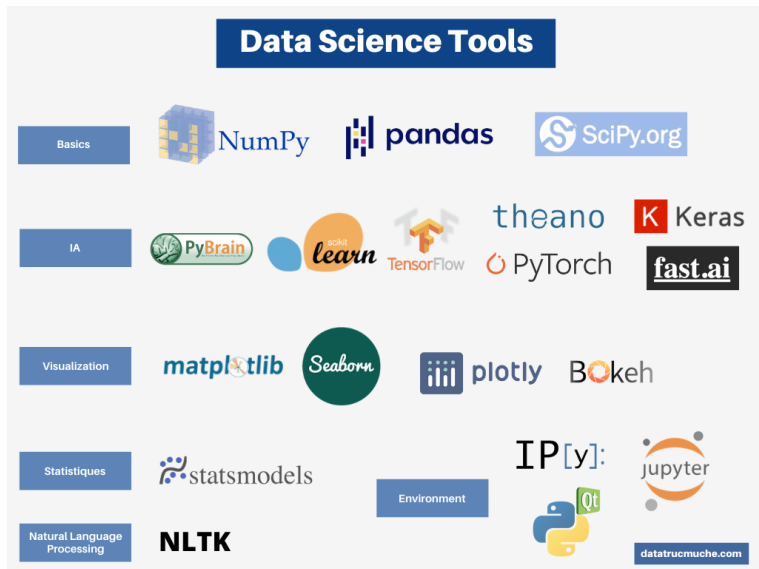
PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Outils pour la science des données



Plan du cours

- 1 Introduction à la science des données
 - Introduction
 - Métiers liés à la science des données
- 2 Introduction à l'apprentissage automatique
 - Introduction
 - Préparation des données
 - Choix du modèle d'apprentissage
- 3 Apprentissage supervisé (classification)
 - Introduction
 - Évaluation des classifieurs
 - k plus proches voisins
- 4 Références

Plan du cours

- 1 Introduction à la science des données
 - Introduction
 - Métiers liés à la science des données
- 2 Introduction à l'apprentissage automatique
 - **Introduction**
 - Préparation des données
 - Choix du modèle d'apprentissage
- 3 Apprentissage supervisé (classification)
 - Introduction
 - Évaluation des classifieurs
 - k plus proches voisins
- 4 Références

Introduction à l'apprentissage automatique

- **Apprentissage automatique** (*Machine Learning*) : processus d'apprentissage d'un **modèle** (ou **classifieur**), à partir de données (étiquetées ou non)

→ Le classifieur permettra ensuite de classer de nouvelles données

- **Exemples de tâches**

- ▶ Classer des images en différentes catégories (différents animaux, fruits...)
- ▶ Affecter des textes à des catégories (ou classes) prédéfinies
- ▶ Aider à décider si un patient est atteint d'une maladie
- ▶ Faire de la reconnaissance de caractères manuscrits
- ▶ Faire de la traduction automatique
- ▶ ...

Étapes d'un problème d'apprentissage automatique

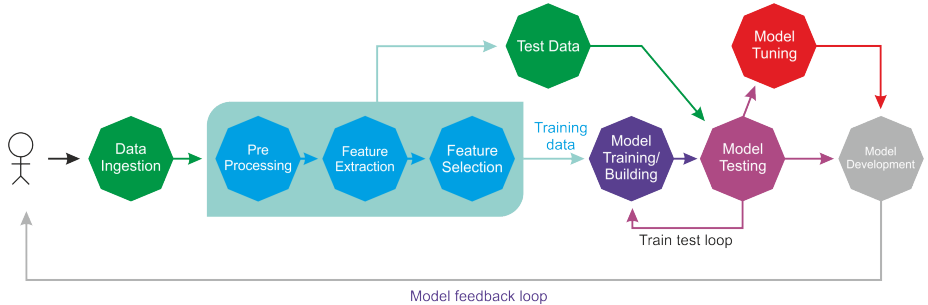


Image tirée de https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781788479042/1/ch01lvl1sec10/machine-learning-and-learning-workflow

Plan du cours

- 1 Introduction à la science des données
 - Introduction
 - Métiers liés à la science des données
- 2 Introduction à l'apprentissage automatique
 - Introduction
 - **Préparation des données**
 - Choix du modèle d'apprentissage
- 3 Apprentissage supervisé (classification)
 - Introduction
 - Évaluation des classifieurs
 - k plus proches voisins
- 4 Références

Représentation des données

- Les algorithmes d'apprentissage ne peuvent pas directement traiter les données telles quelles
- Les données sont généralement représentées par des **vecteurs de caractéristiques** (*feature vectors*)
- Les caractéristiques peuvent être **binaires** : la valeur de la caractéristique x_i , dans le vecteur représentant la donnée, sera égale à 1 ou 0 selon que la caractéristique est présente ou non dans la donnée
- Les caractéristiques peuvent également être **réelles**
- Il peut parfois être nécessaire d'effectuer une **sélection de caractéristiques**, pour ne conserver que les plus pertinentes

Découpage des données en sous-ensembles

- L'ensemble total des données est décomposé en deux ou trois sous-ensembles :
 - ▶ L'ensemble d'apprentissage (ou d'entraînement) contient les exemples utilisés pour apprendre le modèle
 - ▶ L'ensemble de test contient les données utilisées pour évaluer les performances du modèle
 - Si on a appris différents modèle sur le même ensemble d'apprentissage, on pourra comparer leurs performances sur l'ensemble de test
 - ▶ L'ensemble de validation est utilisé s'il y a des paramètres à optimiser dans le modèle
 - L'ensemble de validation sert à tester le modèle appris, à différentes itérations de l'apprentissage, pour éviter le sur-apprentissage (*overfitting*), c'est-à-dire un apprentissage « par cœur » des données d'apprentissage

Validation croisée : cas du nombre limité de données

- Si le nombre d'exemples est trop faible, on utilise la **validation croisée**
 - 1 Une partie des exemples sert d'ensemble d'apprentissage et de validation
 - ★ Les exemples sont séparés en k sous-ensembles
 - ★ À chaque itération d'apprentissage, $k - 1$ sous-ensembles sont utilisés pour l'apprentissage et le sous-ensemble restant est utilisé pour la validation
 - 2 Le reste des exemples sert d'ensemble de test, à l'issue de l'apprentissage

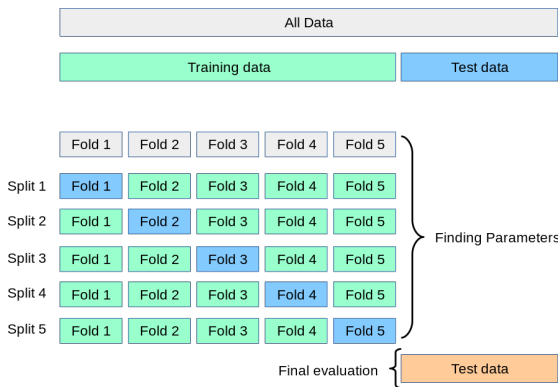


Image tirée de <https://scikit-learn.org>

Plan du cours

- 1 Introduction à la science des données
 - Introduction
 - Métiers liés à la science des données
- 2 Introduction à l'apprentissage automatique
 - Introduction
 - Préparation des données
 - Choix du modèle d'apprentissage
- 3 Apprentissage supervisé (classification)
 - Introduction
 - Évaluation des classifieurs
 - k plus proches voisins
- 4 Références

Différentes approches d'apprentissage automatique

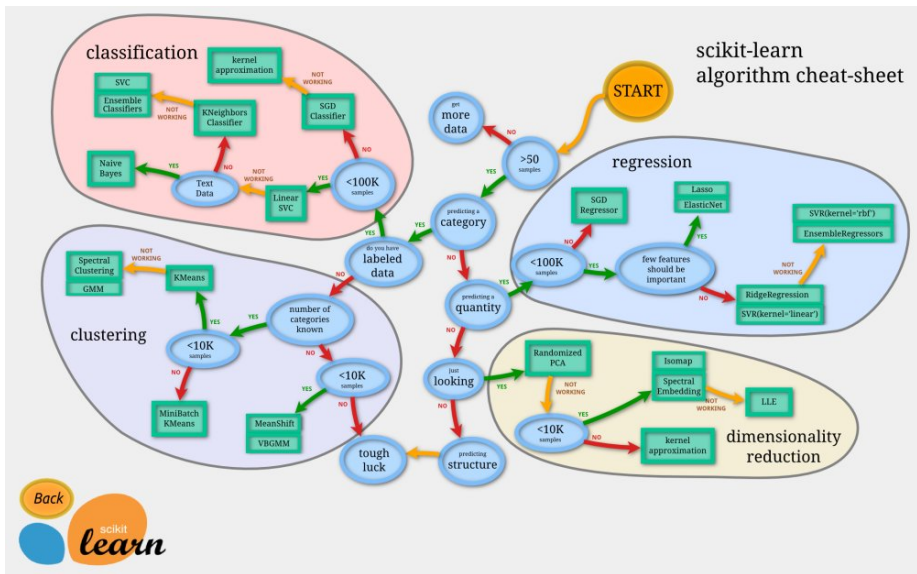


Image tirée de <https://scikit-learn.org>

Plan du cours

- 1 Introduction à la science des données
 - Introduction
 - Métiers liés à la science des données
- 2 Introduction à l'apprentissage automatique
 - Introduction
 - Préparation des données
 - Choix du modèle d'apprentissage
- 3 Apprentissage supervisé (classification)
 - Introduction
 - Évaluation des classifieurs
 - k plus proches voisins
- 4 Références

Plan du cours

- 1 Introduction à la science des données
 - Introduction
 - Métiers liés à la science des données
- 2 Introduction à l'apprentissage automatique
 - Introduction
 - Préparation des données
 - Choix du modèle d'apprentissage
- 3 Apprentissage supervisé (classification)
 - **Introduction**
 - Évaluation des classifieurs
 - k plus proches voisins
- 4 Références

Principe de l'apprentissage supervisé

- 2 étapes

- 1 Apprentissage d'un **classifieur** à partir d'exemples **étiquetés** (*training*)
- 2 Utilisation du classifieur pour trouver la **classe** de nouvelles données (*prediction*)

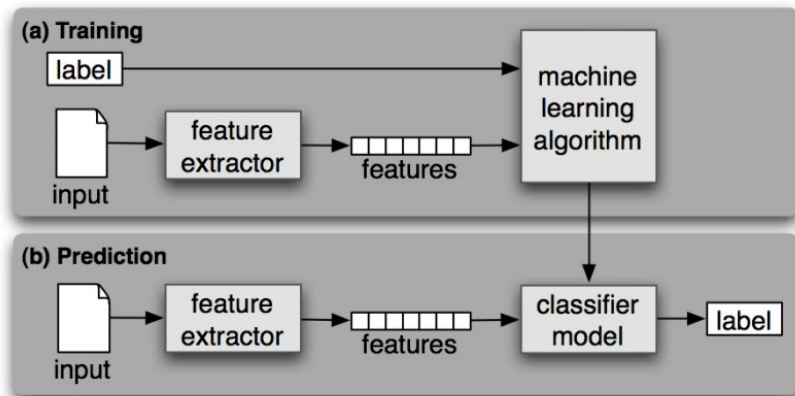


Image tirée de <https://github.com/mmmayo13/scikit-learn-beginners-tutorials>

Formalisation de la classification de données

- Étant donné un ensemble d'exemples X et un ensemble de classes C , la **classification de données** vise à apprendre une fonction $F : X \mapsto C$
 - ▶ Le classifieur appris doit donner une approximation la plus proche possible de la fonction F
 - ▶ Chaque exemple de l'ensemble X est étiqueté avec une ou plusieurs classes de C
- Une fois le classifieur appris, celui-ci peut être utilisé pour **classer de nouvelles données** en déterminant la ou les classes de ces données

Nombre de classes : deux ou plusieurs ?

- L'ensemble des classes C peut être constitué soit de **deux classes**, soit de **plus de deux classes**
- **Cas avec deux classes : classification binaire**
 - ▶ Il existe deux classes : par exemple, *messages indésirables* et *messages désirables*
 - Une donnée ne peut appartenir qu'à une seule classe
- **Cas avec plusieurs classes** : différentes stratégies pour l'apprentissage des classifieurs
 - ▶ Apprentissage « un contre un » (*one versus one*)
 - Un classifieur binaire est appris pour discriminer chaque couple de classes (autant de classifieurs que de couples de classes)
 - ▶ Apprentissage « un contre tous » (*one versus all*)
 - Un classifieur binaire est appris pour discriminer une classe par rapport au reste des classes (autant de classifieurs que de classes)

Type de classification : stricte ou floue ?

- Une donnée peut soit appartenir à une seule classe, soit appartenir à plusieurs classes
- **Classification stricte** (*hard categorization*) : une donnée ne peut appartenir qu'à une seule classe
 - La fonction F peut se réécrire en : $F : X \mapsto c$ où $c \in C$
- **Classification floue** (*soft clustering*) : une donnée peut appartenir à plusieurs classes
 - Par exemple, un document peut appartenir à la catégorie *sport* et à la catégorie *finance*
 - Le nombre de classes auxquelles une donnée appartient peut être choisi selon un seuil (selon le type de classifieur utilisé)

Plan du cours

- 1 Introduction à la science des données
 - Introduction
 - Métiers liés à la science des données
- 2 Introduction à l'apprentissage automatique
 - Introduction
 - Préparation des données
 - Choix du modèle d'apprentissage
- 3 Apprentissage supervisé (classification)
 - Introduction
 - **Évaluation des classifieurs**
 - k plus proches voisins
- 4 Références

Matrice de confusion et mesures d'évaluation binaires

		Classe réelle	
		Classe +	Classe -
Classe prédite	Classe +	TP : vrais positifs	FP : faux positifs
	Classe -	FN : faux négatifs	TN : vrais négatifs

- L'**exactitude** (*accuracy*) correspond au taux de décisions correctes :

$$Acc = \frac{TP + TN}{TP + FN + TN + FP}$$

- La **précision** (*precision*) correspond à la proportion de prédictions correctes, parmi toutes les prédictions positives : $P = \frac{TP}{TP+FP}$
- Le **rappel** (*recall*) correspond à la proportion d'exemples positifs qui ont été correctement prédits : $R = \frac{TP}{TP+FN}$
- La **F-mesure** est la moyenne harmonique du rappel et de la précision :

$$F = \frac{2 \times P \times R}{P + R}$$

Exemple : évaluation d'un dépistage

	Personnes malades	Personnes en bonne santé
Personnes considérées malades	190	210
Personnes considérées non malades	10	3590

Calcul de différentes métriques

- Calcul de l'exactitude :
- Calcul de la précision :
- Calcul du rappel :
- Calcul de la F-mesure :
- Calcul de la spécificité :

Exemple : évaluation d'un dépistage

	Personnes malades	Personnes en bonne santé
Personnes considérées malades	190	210
Personnes considérées non malades	10	3590

Calcul de différentes métriques

- Calcul de l'exactitude : $Acc = \frac{190+3590}{190+10+3590+210} = \frac{3780}{4000} = 0,945$
- Calcul de la précision :
- Calcul du rappel :
- Calcul de la F-mesure :
- Calcul de la spécificité :

Exemple : évaluation d'un dépistage

	Personnes malades	Personnes en bonne santé
Personnes considérées malades	190	210
Personnes considérées non malades	10	3590

Calcul de différentes métriques

- Calcul de l'exactitude : $Acc = \frac{190+3590}{190+10+3590+210} = \frac{3780}{4000} = 0,945$
- Calcul de la précision : $P = \frac{190}{190+210} = \frac{190}{400} = 0,475$
- Calcul du rappel :
- Calcul de la F-mesure :
- Calcul de la spécificité :

Exemple : évaluation d'un dépistage

	Personnes malades	Personnes en bonne santé
Personnes considérées malades	190	210
Personnes considérées non malades	10	3590

Calcul de différentes métriques

- Calcul de l'exactitude : $Acc = \frac{190+3590}{190+10+3590+210} = \frac{3780}{4000} = 0,945$
- Calcul de la précision : $P = \frac{190}{190+210} = \frac{190}{400} = 0,475$
- Calcul du rappel : $R = \frac{190}{190+10} = \frac{190}{200} = 0,95$
- Calcul de la F-mesure :
- Calcul de la spécificité :

Exemple : évaluation d'un dépistage

	Personnes malades	Personnes en bonne santé
Personnes considérées malades	190	210
Personnes considérées non malades	10	3590

Calcul de différentes métriques

- Calcul de l'exactitude : $Acc = \frac{190+3590}{190+10+3590+210} = \frac{3780}{4000} = 0,945$
- Calcul de la précision : $P = \frac{190}{190+210} = \frac{190}{400} = 0,475$
- Calcul du rappel : $R = \frac{190}{190+10} = \frac{190}{200} = 0,95$
- Calcul de la F-mesure : $F = \frac{2 \times 0,475 \times 0,95}{0,475 + 0,95} = 0,633$
- Calcul de la spécificité :

Exemple : évaluation d'un dépistage

	Personnes malades	Personnes en bonne santé
Personnes considérées malades	190	210
Personnes considérées non malades	10	3590

Calcul de différentes métriques

- Calcul de l'exactitude : $Acc = \frac{190+3590}{190+10+3590+210} = \frac{3780}{4000} = 0,945$
- Calcul de la précision : $P = \frac{190}{190+210} = \frac{190}{400} = 0,475$
- Calcul du rappel : $R = \frac{190}{190+10} = \frac{190}{200} = 0,95$
- Calcul de la F-mesure : $F = \frac{2 \times 0,475 \times 0,95}{0,475 + 0,95} = 0,633$
- Calcul de la spécificité : $Spec = \frac{3590}{210+3590} = \frac{3590}{3800} = 0,945$

Évaluation de la classification multiclasse

- Les mesures précédentes permettent d'évaluer la performance d'une seule classe
- On veut pouvoir mesurer les performances sur l'ensemble des classes

● Macro-moyenne

- ▶ Dans la **macro-moyenne** (*macro average*), chaque classe c_i a le même poids
- ▶ Elle est définie par (pour la précision, par exemple) :

$$MacP = \frac{\sum_{i=1}^M \frac{TP_i}{TP_i + FP_i}}{M}$$

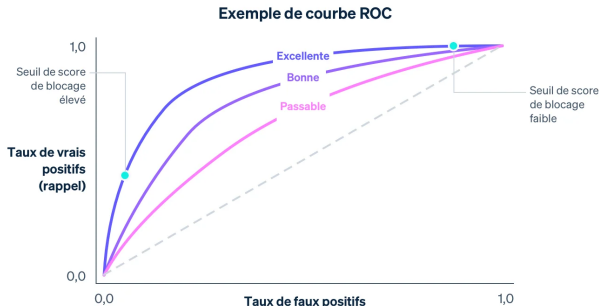
● Micro-moyenne

- ▶ Dans la **micro-moyenne** (*micro average*), chaque classe a un poids proportionnel au nombre d'exemples qu'elle contient
- Une classe avec beaucoup d'exemples aura donc un poids plus important
- ▶ Elle est définie par (pour la précision, par exemple) :

$$MicP = \frac{\sum_{i=1}^M TP_i}{\sum_{i=1}^M (TP_i + FP_i)}$$

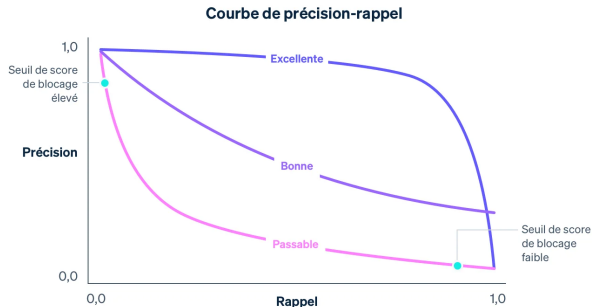
Courbe ROC

- Une **courbe ROC** (*Receiver Operating Characteristics*) affiche le taux de vrais positifs (ou sensibilité, *sensitivity*) par rapport au taux de faux positifs (ou 1-spécificité)
 - ▶ La spécificité (*specificity*) est définie par : $Spec = \frac{TN}{FP+TN}$
- La courbe ROC se concentre sur la capacité d'un modèle à distinguer les exemples positifs des exemples négatifs, en minimisant les faux positifs
- Chaque point de la courbe correspond à un classifieur avec des réglages de paramètres différents
- Le meilleur classifieur se trouve dans le coin supérieur gauche



Courbe précision-rappel

- Une **courbe précision-rappel** (*Precision-Recall*) affiche la précision par rapport au rappel
- La courbe précision-rappel se concentre sur la capacité d'un modèle à fournir des résultats précis, pour les exemples positifs
- Chaque point de la courbe correspond à un classifieur avec des réglages de paramètres différents
- **Le meilleur classifieur se trouve dans le coin supérieur droit**



Plan du cours

- 1 Introduction à la science des données
 - Introduction
 - Métiers liés à la science des données
- 2 Introduction à l'apprentissage automatique
 - Introduction
 - Préparation des données
 - Choix du modèle d'apprentissage
- 3 Apprentissage supervisé (classification)
 - Introduction
 - Évaluation des classifieurs
 - **k plus proches voisins**
- 4 Références

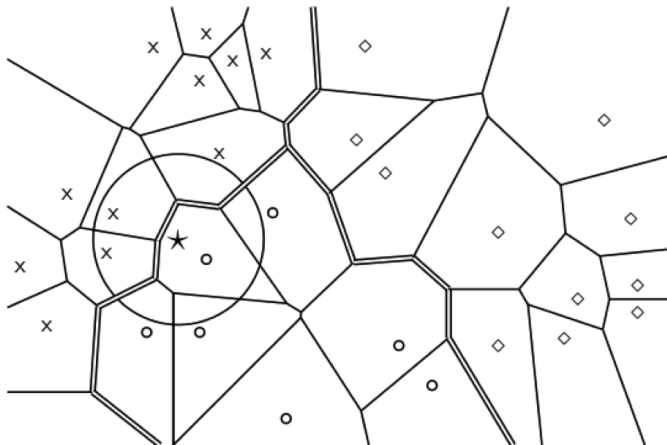
k plus proches voisins (1)

- Dans la classification par les **k plus proches voisins** (*k nearest neighbors*), il n'y a **pas d'étape d'apprentissage** à proprement parler
- Les exemples du corpus d'apprentissage, avec leur classe associée, constituent le classifieur
- Lors de l'étape de classification d'une nouvelle donnée, on mesure la similarité de cette donnée avec chacun des exemples du corpus d'apprentissage
- La classe c de la donnée d dépend de celles de ses k plus proches voisins :

$$c = \arg \max_i \sum_{j=1}^k \text{sim}(d_j, d) \times \delta(c_{d_j}, i)$$

- ▶ $\text{sim}(d_j, d)$ est la similarité entre la donnée d et l'exemple d_j
- ▶ $\delta(c_{d_j}, i)$ est la fonction de Krœnecker qui vaut 1 si la classe i considérée est égale à la classe de l'exemple d_j et 0 sinon

k plus proches voisins (2)



- Avec $k = 1$, l'étoile est classée dans la classe « cercle »
- Avec $k = 3$, l'étoile est classée dans la classe « croix »

k plus proches voisins (3) : exemple

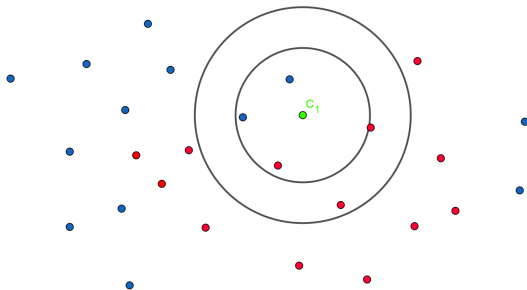


Image tirée de <https://www.numerique-sciences-informatiques.fr/coursAlgorithmiquekPlusProches.php>

Classification par k plus proches voisins

- Avec $k = 1$, quelle est la classe de la nouvelle donnée c_1 ?
- Avec $k = 3$, quelle est la classe de la nouvelle donnée c_1 ?
- Avec $k = 5$, quelle est la classe de la nouvelle donnée c_1 ?

k plus proches voisins (3) : exemple

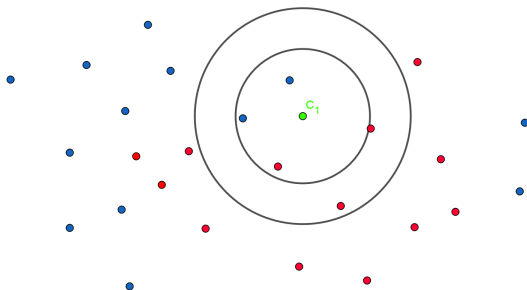


Image tirée de <https://www.numerique-sciences-informatiques.fr/coursAlgorithmiquekPlusProches.php>

Classification par k plus proches voisins

- Avec $k = 1$, quelle est la classe de la nouvelle donnée c_1 ? Bleu
- Avec $k = 3$, quelle est la classe de la nouvelle donnée c_1 ?
- Avec $k = 5$, quelle est la classe de la nouvelle donnée c_1 ?

k plus proches voisins (3) : exemple

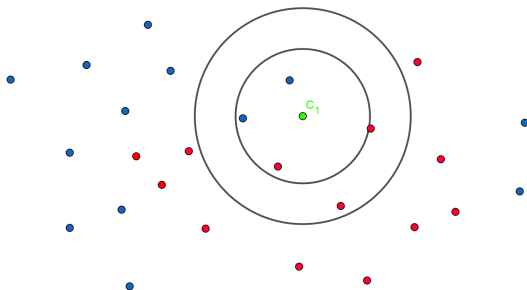


Image tirée de <https://www.numerique-sciences-informatiques.fr/coursAlgorithmiquekPlusProches.php>

Classification par k plus proches voisins

- Avec $k = 1$, quelle est la classe de la nouvelle donnée c_1 ? Bleu
- Avec $k = 3$, quelle est la classe de la nouvelle donnée c_1 ? Bleu
- Avec $k = 5$, quelle est la classe de la nouvelle donnée c_1 ?

k plus proches voisins (3) : exemple

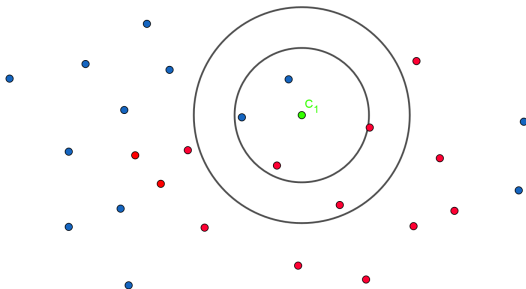


Image tirée de <https://www.numerique-sciences-informatiques.fr/coursAlgorithmiquekPlusProches.php>






Classification par k plus proches voisins

- Avec $k = 1$, quelle est la classe de la nouvelle donnée c_1 ? Bleu
- Avec $k = 3$, quelle est la classe de la nouvelle donnée c_1 ? Bleu
- Avec $k = 5$, quelle est la classe de la nouvelle donnée c_1 ? Rouge

Plan du cours

- 1 Introduction à la science des données
 - Introduction
 - Métiers liés à la science des données
- 2 Introduction à l'apprentissage automatique
 - Introduction
 - Préparation des données
 - Choix du modèle d'apprentissage
- 3 Apprentissage supervisé (classification)
 - Introduction
 - Évaluation des classifieurs
 - k plus proches voisins
- 4 Références

Références I

-  J.A. Hartigan and M.A. Wong, *A k-means clustering algorithm*, Applied Statistics **28** (1979), 100–108.
-  B. King, *Step-wise clustering procedures*, Journal of the American Statistical Association **69** (1967), 86–101.
-  S.P. Lloyd, *Least squares quantization in pcm*, IEEE Transactions on Information Theory **28** (1982), no. 2, 129–136.
-  W.M. Rand, *Objective criteria for the evaluation of clustering methods*, Journal of the American Statistical Association **66** (1971), no. 336, 846–850.
-  P.H.A. Sneath and R.R. Sokal, *Numerical taxonomy : the principles and practice of numerical classification*, W.H. Freeman, 1973.