Apprentissage Supervisé Classification

Contenu

Méthodologie d'un projet de Machine Learning

Un projet de machine learning suit généralement une démarche structurée qui permet de passer des données brutes à un modèle prédictif opérationnel. Cette démarche comprend plusieurs étapes clés, allant de la compréhension du problème à la mise en production du modèle.

Méthodologie d'un projet de Machine Learning

Un projet de machine learning suit généralement une démarche structurée qui permet de passer des données brutes à un modèle prédictif opérationnel. Cette démarche comprend plusieurs étapes clés, allant de la compréhension du problème à la mise en production du modèle.

Les Étapes fondamentales d'un projet de Machine Learning

Méthodologie d'un projet de Machine Learning

Un projet de machine learning suit généralement une démarche structurée qui permet de passer des données brutes à un modèle prédictif opérationnel. Cette démarche comprend plusieurs étapes clés, allant de la compréhension du problème à la mise en production du modèle.

Les Étapes fondamentales d'un projet de Machine Learning

1. **Définir les Objectifs du Projet** : Clarifier les buts et les résultats attendus.

Méthodologie d'un projet de Machine Learning

Un projet de machine learning suit généralement une démarche structurée qui permet de passer des données brutes à un modèle prédictif opérationnel. Cette démarche comprend plusieurs étapes clés, allant de la compréhension du problème à la mise en production du modèle.

Les Étapes fondamentales d'un projet de Machine Learning

1. **Définir les Objectifs du Projet** : Clarifier les buts et les résultats attendus.
2. **Collecte des Données** : Identifier et acquérir les sources de données pertinentes.

Méthodologie d'un projet de Machine Learning

Un projet de machine learning suit généralement une démarche structurée qui permet de passer des données brutes à un modèle prédictif opérationnel. Cette démarche comprend plusieurs étapes clés, allant de la compréhension du problème à la mise en production du modèle.

Les Étapes fondamentales d'un projet de Machine Learning

1. **Définir les Objectifs du Projet** : Clarifier les buts et les résultats attendus.
2. **Collecte des Données** : Identifier et acquérir les sources de données pertinentes.
3. **Prétraitement des Données**
   - **Nettoyage des Données** : les valeurs manquantes, les erreurs et les incohérences.
   - **Transformation des Données** : Normaliser, standardiser et encoder.

Méthodologie d'un projet de Machine Learning

Un projet de machine learning suit généralement une démarche structurée qui permet de passer des données brutes à un modèle prédictif opérationnel. Cette démarche comprend plusieurs étapes clés, allant de la compréhension du problème à la mise en production du modèle.

Les Étapes fondamentales d'un projet de Machine Learning

1. **Définir les Objectifs du Projet** : Clarifier les buts et les résultats attendus.
2. **Collecte des Données** : Identifier et acquérir les sources de données pertinentes.
3. **Prétraitement des Données**
   - **Nettoyage des Données** : les valeurs manquantes, les erreurs et les incohérences.
   - **Transformation des Données** : Normaliser, standardiser et encoder.
4. **Sélection des Algorithmes**
   - **Entraînement et Validation du Modèle**
   - **Évaluation des Performances** : Mesurer la précision, la robustesse.

Algorithme de machine learning

Une fois les données prêtes à être injectées dans un algorithme commence la phase à proprement parler de Machine Learning (Choix du/des modèles /algorithmes).
Mais avant de sélectionner les algorithmes les plus adaptés à la nature des données et aux objectifs visés.
Quelles sont les différents types d'algorithmes de machine learning ?

Algorithme de machine learning

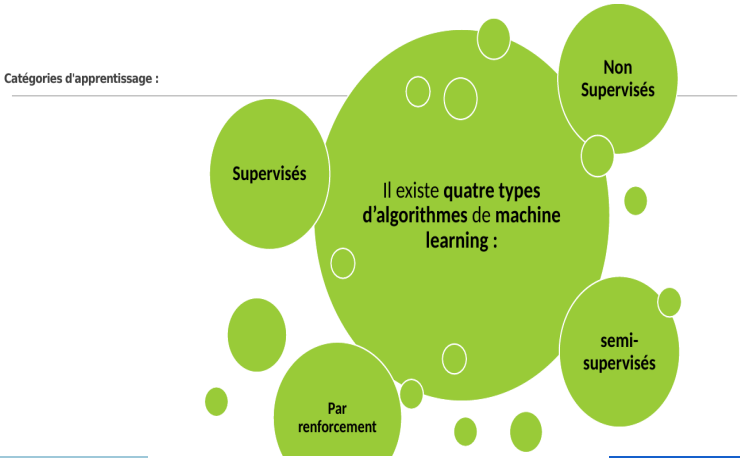
Définition: un algorithme de machine learning

Un algorithme de machine learning est un processus qui permet à une machine d'apprendre à partir de données, en identifiant des motifs ou en prédisant des valeurs à partir de variables d'entrée.

Algorithme de machine learning

Définition: un algorithme de machine learning

Un algorithme de machine learning est un processus qui permet à une machine d'apprendre à partir de données, en identifiant des motifs ou en prédisant des valeurs à partir de variables d'entrée.



Apprentissage Supervisé

- L'apprentissage supervisé, aussi appelé Machine Learning supervisé, fait partie des méthodes d'apprentissage automatique.

Apprentissage Supervisé

- L'apprentissage supervisé, aussi appelé Machine Learning supervisé, fait partie des méthodes d'apprentissage automatique.
- Il consiste à entraîner un algorithme à l'aide de données d'entrée et de sortie connues et étiquetées.

Apprentissage Supervisé

- L'apprentissage supervisé, aussi appelé Machine Learning supervisé, fait partie des méthodes d'apprentissage automatique.
- Il consiste à entraîner un algorithme à l'aide de données d'entrée et de sortie connues et étiquetées.
- L'objectif est d'utiliser ces données d'entraînement pour construire un modèle capable de prédire avec précision la sortie pour de nouvelles données non vues.

Apprentissage Supervisé

- L'apprentissage supervisé, aussi appelé Machine Learning supervisé, fait partie des méthodes d'apprentissage automatique.
- Il consiste à entraîner un algorithme à l'aide de données d'entrée et de sortie connues et étiquetées.
- L'objectif est d'utiliser ces données d'entraînement pour construire un modèle capable de prédire avec précision la sortie pour de nouvelles données non vues.
- L'algorithme mesure sa précision par le biais de la fonction de perte, en s'ajustant jusqu'à ce que l'erreur soit suffisamment réduite.

Catégories d'apprentissage supervisé

L'apprentissage supervisé peut être divisé en deux sous-catégories : la régression et la classification.

Catégories d'apprentissage supervisé

L'apprentissage supervisé peut être divisé en deux sous-catégories : la régression et la classification.

Régression

- La régression permet de comprendre la relation entre les variables dépendantes et indépendantes.
- Elle est couramment utilisée pour établir des projections, telles que le chiffre d'affaires d'une entreprise donnée.

Catégories d'apprentissage supervisé

L'apprentissage supervisé peut être divisé en deux sous-catégories : la régression et la classification.

Régression

- La régression permet de comprendre la relation entre les variables dépendantes et indépendantes.
- Elle est couramment utilisée pour établir des projections, telles que le chiffre d'affaires d'une entreprise donnée.

Classification

- La classification utilise un algorithme pour attribuer avec précision des données de test à des catégories particulières.
- Attribuer une catégorie ou une classe à chaque observation d'un ensemble de données, en fonction de ses caractéristiques.

Classification

Définition

La classification permet de prédire si un élément est membre d'un groupe ou d'une catégorie donnée.

Classes:

- Identification de groupes avec des profils particuliers.
- Possibilité de décider de l'appartenance d'une entité à une classe.

Caractéristiques de classification:

- Apprentissage supervisé: classes connues à l'avance.
- Qualité de la classification (taux d'erreur).

Exemples de problème de classification

Exemples de problème de classification

- La détection de spams:
  - Après avoir scanné le texte d'un mail,
  - Tagguer certains mots et phrases.
  - La signature du message peut être injectée dans un algorithme de classification.

Déterminer s'il s'agit d'un spam ou non.

Exemples de problème de classification

- La détection de spams:
  - Après avoir scanné le texte d'un mail,
  - Tagguer certains mots et phrases.
  - La signature du message peut être injectée dans un algorithme de classification.

  Déterminer s'il s'agit d'un spam ou non.
- L'analyse du risque dans le domaine de la santé:
  - Les statistiques vitales d'un patient.
  - L'historique de santé.
  - Les niveaux d'activités.
  - Les données démographiques

  Ces données peuvent être croisées pour attribuer une note (un niveau de risque) et évaluer la probabilité d'une maladie.

Processus de classification

Le processus de classification se fait en deux étapes:

Processus de classification

Le processus de classification se fait en deux étapes:

Etape 1

Construction du modèle à partir de l'ensemble d'apprentissage (training set).

Processus de classification

Le processus de classification se fait en deux étapes:

Etape 1

Construction du modèle à partir de l'ensemble d'apprentissage (training set).

Etape 2

Utilisation du modèle: tester la précision du modèle (test set) et l'utiliser dans la classification de futur donnée(nouvelles données) .

Etape 1: Construction de modèle

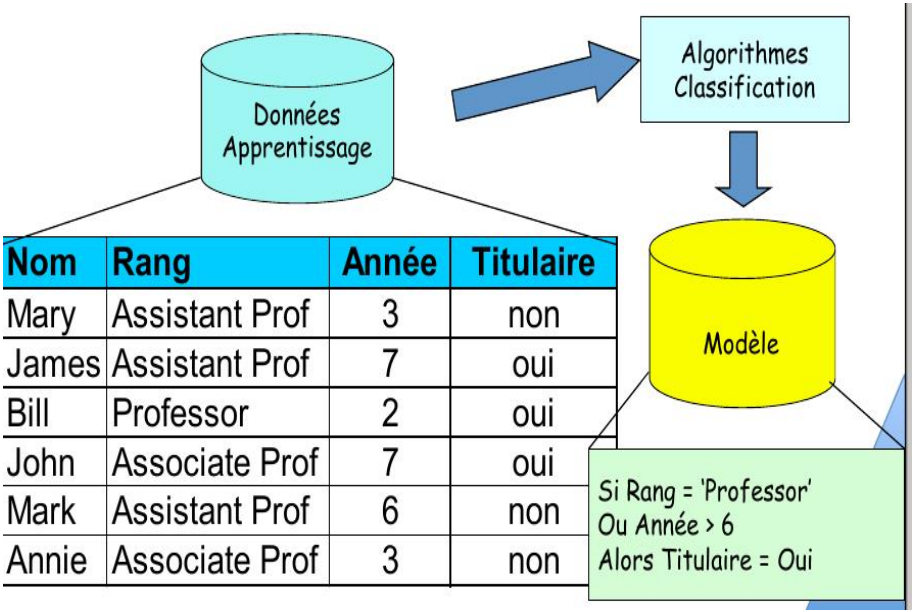
Chaque donnée est affectée à une classe selon ces valeurs.

1. La classe d'une donnée est déterminée par l'attribut classe.
2. L'ensemble des données d'apprentissage (train set) est utilisé dans la construction du modèle (entrainement).
3. Le modèle est représenté par des règles de classification (Algorithme d'apprentisage)

Etape 2: Utilisation du modèle

- Classification de nouvelles donnée ou donnée inconnues
- Estimer le taux d'erreur du modèle
  - La classe connue d'une donnée test est comparée avec le résultat du modèle.
  - Taux d'erreur = pourcentage de tests incorrectement classés par le modèle.

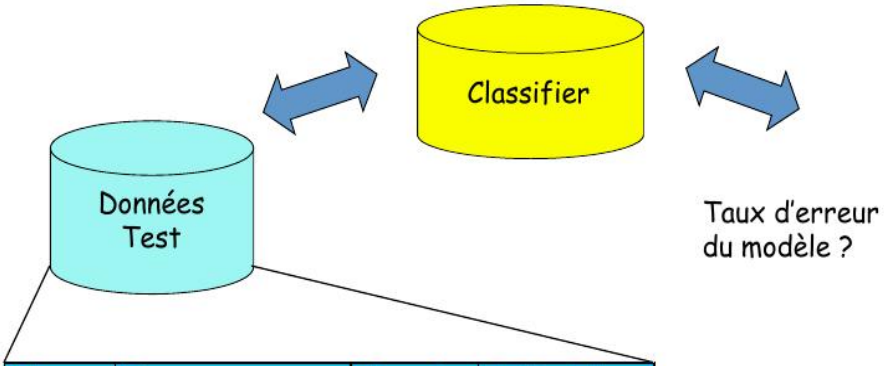
Exemple:Construction de modèle



Nom	Rang	Année	Titulaire
Mary	Assistant Prof	3	non
James	Assistant Prof	7	oui
Bill	Professor	2	oui
John	Associate Prof	7	oui
Mark	Assistant Prof	6	non
Annie	Associate Prof	3	non

Figure: Construction de modèle

Exemple:Validation de modèle



Nom	Rang	Année	Titulaire
Tom	Assistant Prof	2	non
Lisa	Associate Prof	7	non
Jack	Professor	5	oui
Ann	Assistant Prof	7	oui

Figure: Construction de modèle

Exemple:Utilisation de modèle

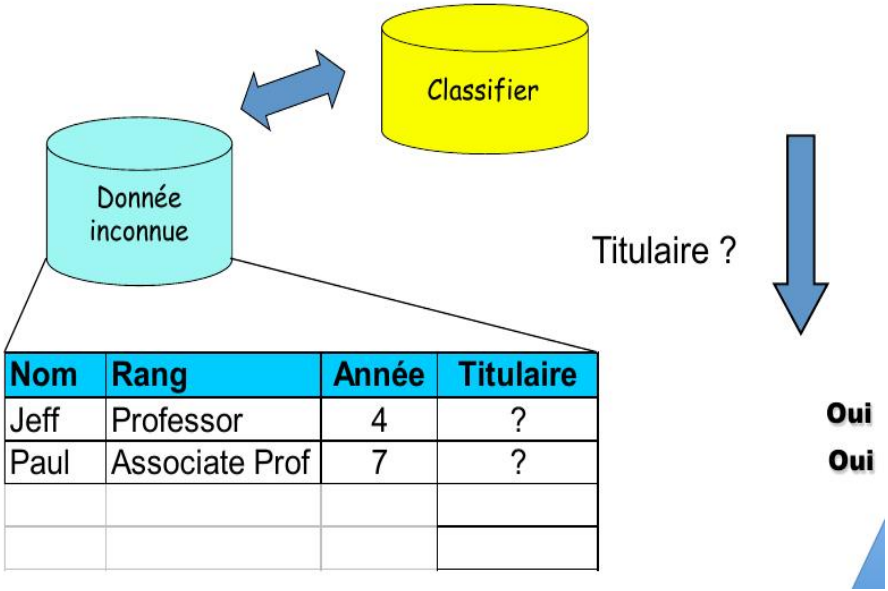


Figure: Construction de modèle

Evaluation des méthodes de classification

Évaluer les performances d'un modèle de classification est primordial:

Evaluation des méthodes de classification

Évaluer les performances d'un modèle de classification est primordial:

- Pour savoir si le modèle est globalement significatif: Mon modèle traduit-il vraiment une causalité ?

Evaluation des méthodes de classification

Évaluer les performances d'un modèle de classification est primordial:

- Pour savoir si le modèle est globalement significatif: Mon modèle traduit-il vraiment une causalité ?
- Pour se donner une idée des performances en déploiement : Quelle sera la fiabilité (les coûts associés) lorsque j'utiliserai mon modèle ?

Evaluation des méthodes de classification

Évaluer les performances d'un modèle de classification est primordial:

- Pour savoir si le modèle est globalement significatif: Mon modèle traduit-il vraiment une causalité ?
- Pour se donner une idée des performances en déploiement : Quelle sera la fiabilité (les coûts associés) lorsque j'utiliserai mon modèle ?
- Pour comparer plusieurs modèles candidats: Lequel parmi plusieurs modèles sera le plus performant compte tenu de mes objectifs ?

Evaluation des méthodes de classification

Évaluer les performances d'un modèle de classification est primordial:

- Pour savoir si le modèle est globalement significatif: Mon modèle traduit-il vraiment une causalité ?
- Pour se donner une idée des performances en déploiement : Quelle sera la fiabilité (les coûts associés) lorsque j'utiliserai mon modèle ?
- Pour comparer plusieurs modèles candidats: Lequel parmi plusieurs modèles sera le plus performant compte tenu de mes objectifs ?

Remarque

La mesure et l'évaluation de la performance d'un modèle de classification se fait toujours sur l'échantillon de test: Il faut tester la performance de modèle sur des données qui n'ont pas été utilisées pour construire le modèle de classification.

Evaluation des méthodes de classification

- Plusieurs indicateurs permettent de mesurer la performance des modèles de classification.
- Chaque indicateur a ses spécificités.
- il faut bien souvent en utiliser plusieurs pour avoir une vision complète de la performance de votre modèle.

Pour évaluer la performance d'un modèle de classification nous présentons quatre indicateurs qui sont calculés à partir de la matrice de confusion:

- L'accuracy.
- Le recall.
- La precision.
- F1 score.

Exemple de classification

score de churn

Nous avons une base de données client qui ont été abonnés à un service.

- Des clients qui sont encore abonnés.
- Des clients qui ont résilié le service.

Client	Âge	Durée abonnement	Tarif	Statut
1	25	24	3,99	Abonné
2	32	12	3,99	Abonné
3	28	24	7,99	Abonné
...		...		
10000	53	36	2,99	Résilié

Base de données client – score de churn

Exemple de classification

Nous construisons un score de churn : pour chaque client, on prédit s'il va résilier ou conserver son abonnement le mois suivant.

Client	Âge	Durée abonnement	Tarif	Statut	Prédiction
1	25	24	3,99	Abonné	Résilié
2	32	12	3,99	Abonné	Abonné
3	28	24	7,99	Abonné	Abonné
...			...		
10000	53	36	2,99	Résilié	Résilié

Base de données client – score de churn

Exemple de classification

Nous construisons un score de churn : pour chaque client, on prédit s'il va résilier ou conserver son abonnement le mois suivant.

Client	Âge	Durée abonnement	Tarif	Statut	Prédiction
1	25	24	3,99	Abonné	Résilié
2	32	12	3,99	Abonné	Abonné
3	28	24	7,99	Abonné	Abonné
...			...		
10000	53	36	2,99	Résilié	Résilié

Base de données client – score de churn

- Quelle est la performance de ce score ?

Exemple de classification

Nous construisons un score de churn : pour chaque client, on prédit s'il va résilier ou conserver son abonnement le mois suivant.

Client	Âge	Durée abonnement	Tarif	Statut	Prédiction
1	25	24	3,99	Abonné	Résilié
2	32	12	3,99	Abonné	Abonné
3	28	24	7,99	Abonné	Abonné
...			...		
10000	53	36	2,99	Résilié	Résilié

Base de données client – score de churn

- Quelle est la performance de ce score ?
- A quel point je peux lui faire confiance pour prédire les résiliations futures?

Evaluation des méthodes de classification

Evaluation des méthodes de classification

Matrice de confusion

Une matrice de confusion sert à évaluer la qualité d'une classification. Elle est obtenue en comparant les données classées avec des données de référence (test set) qui doivent être différentes de celles ayant servi à réaliser la classification (train set).

Evaluation des méthodes de classification

Matrice de confusion

Une matrice de confusion sert à évaluer la qualité d'une classification. Elle est obtenue en comparant les données classées avec des données de référence (test set) qui doivent être différentes de celles ayant servi à réaliser la classification (train set).

Classification supervisée binaire, $y \in \{0, 1\}$, où la modalité de la variable à prédire correpond à la classe postive et l'autre à la classe négative, on nomme les coefficients de la matrice de confusion:

		Données prédites par l'algorithme	
		Résilié prédit	Abonné prédit
Données réelles	Résilié	Vrai positif	Faux négatif
	Abonné	Faux positif	Vrai négatif

Matrice de confusion

Matrice de confusion

- Les fausses prédictions:
  - Nombre de faux négatifs (FN): les clients qui ont résilié mais pour lesquels le score a prédit à tort qu'ils allaient rester abonnés.
  - Nombre de faux positifs (FP): les clients qui sont restés abonnés alors que le score a prédit à tort qu'ils allaient résilier.

Matrice de confusion

- Les fausses prédictions:
  - Nombre de faux négatifs (FN): les clients qui ont résilié mais pour lesquels le score a prédit à tort qu'ils allaient rester abonnés.
  - Nombre de faux positifs (FP): les clients qui sont restés abonnés alors que le score a prédit à tort qu'ils allaient résilier.
- Les bonnes prédictions:
  - Nombre de vrais positifs (VP): les clients qui ont résilié pour lesquels le score a bien prédit qu'ils allaient résilier.
  - Nombre de vrais négatifs(VN): les clients qui sont toujours abonnés et pour lesquels l'algorithme a bien prédit qu'ils resteraient abonnés.

Evaluation des méthodes de classification

Accuracy

Il indique le pourcentage de bonnes prédictions.

$$\text{Accuracy} = \frac{\text{vrais positifs} + \text{vrais négatifs}}{\text{total}}$$

Evaluation des méthodes de classification

Accuracy

Il indique le pourcentage de bonnes prédictions.

$$\text{Accuracy} = \frac{\text{vrais positifs} + \text{vrais négatifs}}{\text{total}}$$

Parfois, l'accuracy ne suffit pas:

- Considérons un problème de 2-classes:
  - Nombre de Classes 0 égal à 9990
  - Numbre de Classes 1 égal à 10.
  - La base de données n'est pas équilibrée.

Evaluation des méthodes de classification

Accuracy

Il indique le pourcentage de bonnes prédictions.

$$\text{Accuracy} = \frac{\text{vrais positifs} + \text{vrais négatifs}}{\text{total}}$$

Parfois, l'accuracy ne suffit pas:

- Considérons un problème de 2-classes:
  - Nombre de Classes 0 égal à 9990
  - Numbre de Classes 1 égal à 10.
  - La base de données n'est pas équilibrée.
- Si le modèle prédit que tout est de classe 0, la précision est de $9990/10000 = 99,9\%$. La précision est trompeuse car le modèle ne détecte aucun exemple de classe 1.

Recall

Le recall (rappel) permet de répondre à la question suivante :

Quelle proportion de résultats positifs réels a été identifiée correctement ?

Recall

Le recall (rappel) permet de répondre à la question suivante :

Quelle proportion de résultats positifs réels a été identifiée correctement ?

Recall

Il donne une indication sur la part de faux négatifs.

$$\text{Recall} = \frac{\text{vrais positifs}}{\text{Vrais positif} + \text{faux négatifs}}$$

Recall

Le recall (rappel) permet de répondre à la question suivante :

Quelle proportion de résultats positifs réels a été identifiée correctement ?

Recall

Il donne une indication sur la part de faux négatifs.

$$\text{Recall} = \frac{\text{vrais positifs}}{\text{Vrais positif} + \text{faux négatifs}}$$

Un modèle ne produisant aucun faux négatif a un rappel de 1.

Précision

La précision permet de répondre à la question suivante:

Quelle proportion d'identifications positives était effectivement correcte ?

Précision

La précision permet de répondre à la question suivante:

Quelle proportion d'identifications positives était effectivement correcte ?

Précision

Il donne une indication sur les faux positifs.

$$\text{Precision} = \frac{\text{vrais positifs}}{\text{Vrais positifs} + \text{faux positifs}}$$

Précision

La précision permet de répondre à la question suivante:

Quelle proportion d'identifications positives était effectivement correcte ?

Précision

Il donne une indication sur les faux positifs.

$$\text{Precision} = \frac{\text{vrais positifs}}{\text{Vrais positifs} + \text{faux positifs}}$$

Un modèle de classification ne produisant aucun faux positif a une précision de 1.

Précision et Recall

- Pour évaluer les performances d'un modèle de façon complète: Il faut analyser à la fois la précision et le rappel.
- La précision et rappel sont fréquemment en tension: l'amélioration de la précision se fait généralement au détriment du rappel et réciproquement.
- Si on veut comparer les performances de deux classificateurs et on a:
  - Le classificateur A a un recall plus élevé et le classificateur B a une précision plus élevée.

  Alors on ne peut pas comparer les classificateurs A et B

Différents outils ont été créés pour évaluer simultanément la précision et le rappel. La F1 score en fait partie.

F1 score

- Le F1 score combine la précision et le recall d'un classificateur en une seule métrique en prenant leur moyenne harmonique.
- Le F1 score est utilisé pour comparer les performances de deux classificateurs dans le cas suivant:
  - Supposons que le classificateur A a un recall plus élevé et le classificateur B a une précision plus élevée.
  - Dans ce cas, les F1 score des deux classificateurs peuvent être utilisés pour déterminer celui qui produit les meilleurs résultats.

F1 score

- Le F1 score combine la précision et le recall d'un classificateur en une seule métrique en prenant leur moyenne harmonique.
- Le F1 score est utilisé pour comparer les performances de deux classificateurs dans le cas suivant:
  - Supposons que le classificateur A a un recall plus élevé et le classificateur B a une précision plus élevée.
  - Dans ce cas, les F1 score des deux classificateurs peuvent être utilisés pour déterminer celui qui produit les meilleurs résultats.

F1 score

Il est la moyenne pondérée de la précision et du recall. Par conséquent, ce score prend en compte à la fois les faux positifs et les faux négatifs.

$$\text{F1 score} = \frac{2 * (\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}}$$

KNN : K voisins les plus proches (K-Nearest Neighbors)

- K plus proches voisins (KNN): Méthode a pour but de classifier des points cibles (classe inconnue) en fonction de leurs distances par rapport à des points constituant un échantillon d'apprentissage (c'est-à-dire dont la classe est connue a priori).

KNN : K voisins les plus proches (K-Nearest Neighbors)

- K plus proches voisins (KNN): Méthode a pour but de classifier des points cibles (classe inconnue) en fonction de leurs distances par rapport à des points constituant un échantillon d'apprentissage (c'est-à-dire dont la classe est connue a priori).
- L'algorithme des k plus proches voisins est un algorithme d'apprentissage supervisé, il est nécessaire d'avoir des données labellisées.

KNN : K voisins les plus proches (K-Nearest Neighbors)

- K plus proches voisins (KNN): Méthode a pour but de classifier des points cibles (classe inconnue) en fonction de leurs distances par rapport à des points constituant un échantillon d'apprentissage (c'est-à-dire dont la classe est connue a priori).
- L'algorithme des k plus proches voisins est un algorithme d'apprentissage supervisé, il est nécessaire d'avoir des données labellisées.
- À partir d'un ensemble E de données labellisées, il sera possible de classer (déterminer le label) une nouvelle donnée (n'appartenant pas à E).

KNN : K voisins les plus proches (K-Nearest Neighbors)

- K plus proches voisins (KNN): Méthode a pour but de classifier des points cibles (classe inconnue) en fonction de leurs distances par rapport à des points constituant un échantillon d'apprentissage (c'est-à-dire dont la classe est connue a priori).
- L'algorithme des k plus proches voisins est un algorithme d'apprentissage supervisé, il est nécessaire d'avoir des données labellisées.
- À partir d'un ensemble E de données labellisées, il sera possible de classer (déterminer le label) une nouvelle donnée (n'appartenant pas à E).
- Il est aussi possible d'utiliser l'algorithme des k plus proches voisins pour la régression (on cherche à déterminer une valeur à la place d'une classe).

Généralités

- Méthode de raisonnement à partir de cas: prendre des décisions en recherchant un ou des cas similaires déjà résolus.

Généralités

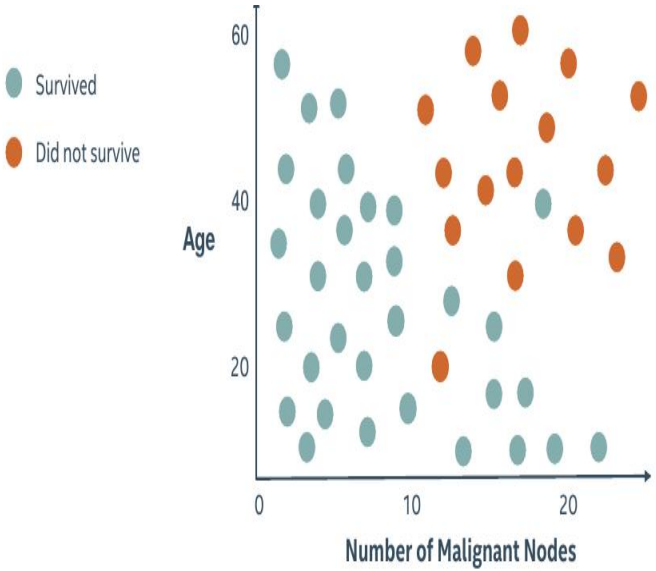
- Méthode de raisonnement à partir de cas: prendre des décisions en recherchant un ou des cas similaires déjà résolus.
- Pas d'étape d'apprentissage: construction d'un modèle à partir d'un échantillon d'apprentissage.

Généralités

- Méthode de raisonnement à partir de cas: prendre des décisions en recherchant un ou des cas similaires déjà résolus.
- Pas d'étape d'apprentissage: construction d'un modèle à partir d'un échantillon d'apprentissage.
- Modèle= échantillon d'apprentissage + fonction de distance + fonction de choix de la classe en fonction des classes des voisins les plus proches.

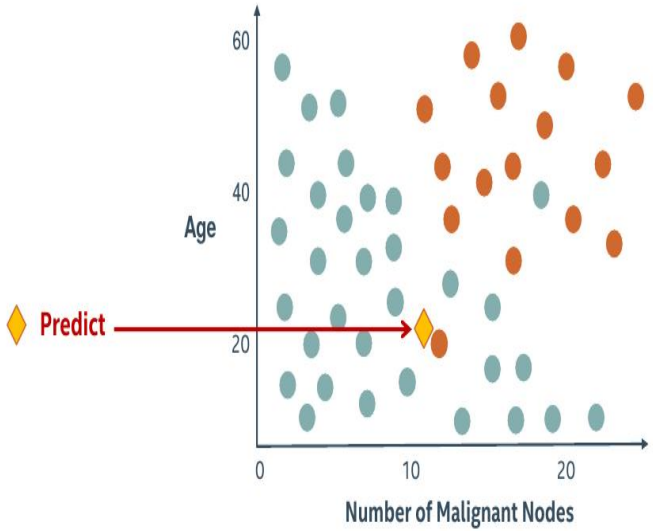
Principe

Soit la base de donnée:



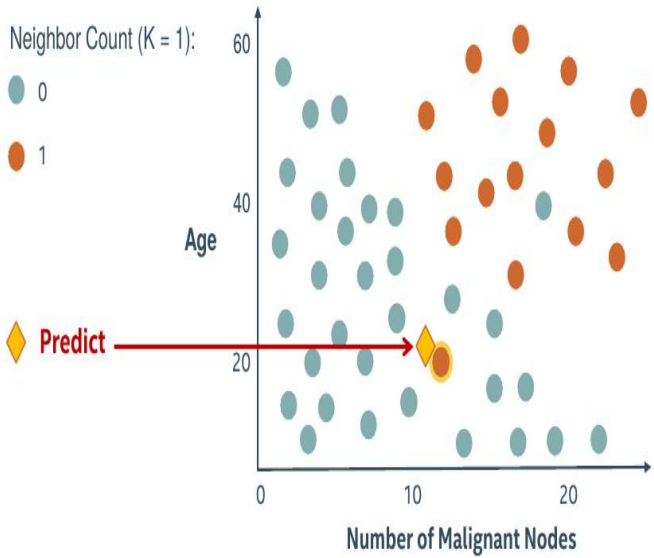
Principe

On veut prédire à quelle classe appartient la nouvelle donnée:



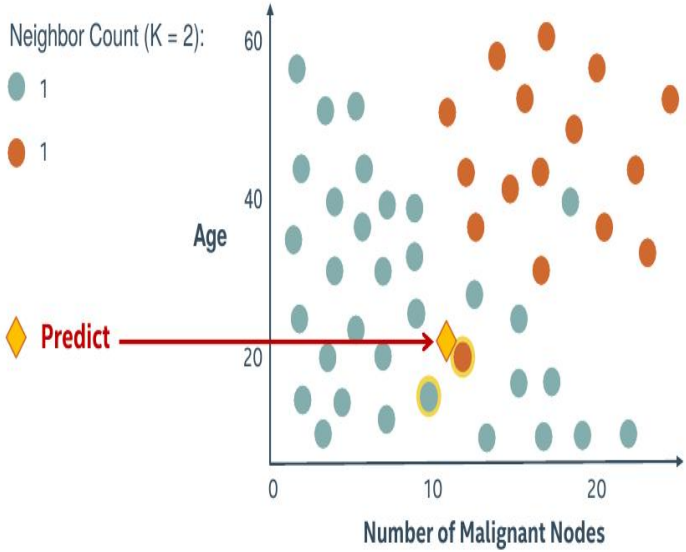
Principe

Si on prend un seul voisin:



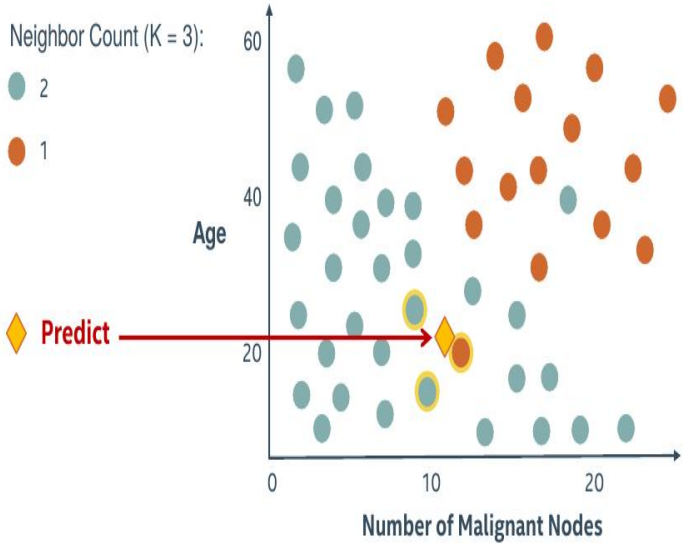
Principe

Si on considère deux voisins:



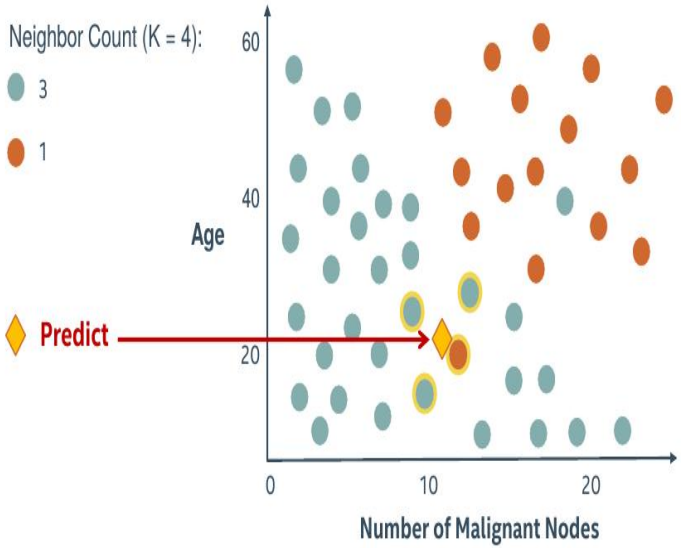
Principe

Si on considère trois voisins:



Principe

Si on considère quatres voisins:



Algorithme de KNN

Début Algorithme

Algorithme de KNN

Début Algorithme

Input :

- un ensemble de données D .
- une fonction de définition distance d.
- Un nombre entier k
- une nouvelle observation X

Algorithme de KNN

Début Algorithme

Input :

- un ensemble de données D .
- une fonction de définition distance d.
- Un nombre entier k
- une nouvelle observation X

Output:

- Prédire la variable de sortie y de X:

Algorithme de KNN

Début Algorithme

Input :

- un ensemble de données D .
- une fonction de définition distance d.
- Un nombre entier k
- une nouvelle observation X

Output:

- Prédire la variable de sortie y de X:

Faire :

1. Calculer toutes les distances de cette observation X avec les autres observations du jeu de données D.
2. Retenir les k observations du jeu de données D les proches de X en utilisation le fonction de calcul de distance d
3. Prendre les valeurs de y des k observations retenues : et calculer le mode de y retenues.
4. Retourner la valeur calculée dans l'étape 3 comme étant la valeur qui a été prédite par KNN pour l'observation X.

Distance

L'algorithme, K-NN a besoin d'une fonction de calcul de distance entre deux observations.

Distance

L'algorithme, K-NN a besoin d'une fonction de calcul de distance entre deux observations.

Définition distance

on appelle distance sur un ensemble E de $\mathbb{R}^n$, une application définie de $E \times E$ à valeurs dans $\mathbb{R}_+$ notée d qui à tout couple $(x, y) \in E \times E$ fait correspondre un réel positif ou nul $d(x, y)$ vérifiant:

Distance

L'algorithme, K-NN a besoin d'une fonction de calcul de distance entre deux observations.

Définition distance

on appelle distance sur un ensemble E de $\mathbb{R}^n$, une application définie de $E \times E$ à valeurs dans $\mathbb{R}_+$ notée d qui à tout couple $(x, y) \in E \times E$ fait correspondre un réel positif ou nul $d(x, y)$ vérifiant:

1. $d(x, y) = 0$ ssi $x = y$.

Distance

L'algorithme, K-NN a besoin d'une fonction de calcul de distance entre deux observations.

Définition distance

on appelle distance sur un ensemble E de $\mathbb{R}^n$, une application définie de $E \times E$ à valeurs dans $\mathbb{R}_+$ notée d qui à tout couple $(x, y) \in E \times E$ fait correspondre un réel positif ou nul $d(x, y)$ vérifiant:

1. $d(x, y) = 0$ **ssi** $x = y$.
2. $d(x, y) = d(y, x)$, $\forall (x, y) \in E^2$.

Distance

L'algorithme, K-NN a besoin d'une fonction de calcul de distance entre deux observations.

Définition distance

on appelle distance sur un ensemble E de $\mathbb{R}^n$, une application définie de $E \times E$ à valeurs dans $\mathbb{R}_+$ notée d qui à tout couple $(x, y) \in E \times E$ fait correspondre un réel positif ou nul $d(x, y)$ vérifiant:

1. $d(x, y) = 0$ **ssi** $x = y$.
2. $d(x, y) = d(y, x)$, $\forall (x, y) \in E^2$.
3. $d(x, y) \leq d(x, z) + d(z, y)$, $\forall (x, y, z) \in E^3$.

Type des distance

Il existe plusieurs fonctions de calcul de distance:

- La distance euclidienne.
- la distance de Manhattan.
- la distance de Minkowski
- la distance de Jaccard.
- la distance de Hamming.

Type des distance

Il existe plusieurs fonctions de calcul de distance:

- La distance euclidienne.
- la distance de Manhattan.
- la distance de Minkowski
- la distance de Jaccard.
- la distance de Hamming.

Le choix de la fonction de distance en fonction des types de données qu'on manipule.

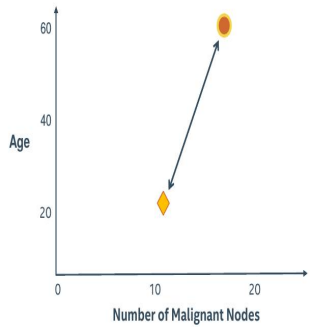
- Ainsi pour les données quantitatives (exemple : poids, salaires, taille, montant de panier éléctronique etc...) et du même type: la distance euclidienne est un bon candidat.
- La distance de Manhattan est une bonne mesure à utiliser quand les données (input variables) ne sont pas du même type (exemple :age, sexe, longueur, poids etc...).

Distance euclidienne

Définition

C'est la distance qui calcule la racine carrée de la somme des différences carrées entre les coordonnées de deux points:
Soient $X = (x_1, x_2, ..., x_n)$ et $Y = (y_1, y_2, ..., y_n)$ la distance euclidienne entre X et Y est:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}.$$

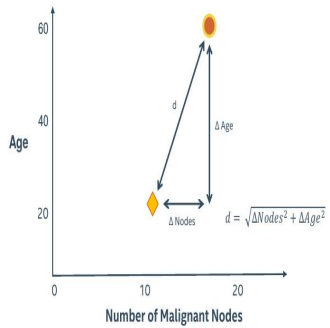


Distance euclidienne

Définition

C'est la distance qui calcule la racine carrée de la somme des différences carrées entre les coordonnées de deux points:
Soient $X = (x_1, x_2, ..., x_n)$ et $Y = (y_1, y_2, ..., y_n)$ la distance euclidienne entre X et Y est:

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

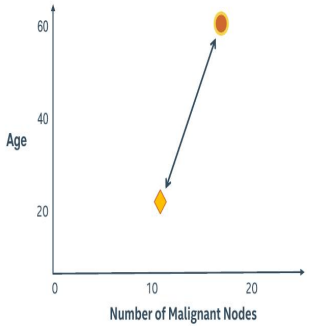


Distance Manhattan

Définition

C'est la distance qui calcule la somme des valeurs absolues des différences entre les coordonnées de deux points:
Soient $X = (x_1, x_2, ..., x_n)$ et $Y = (y_1, y_2, ..., y_n)$ la distance Manhattan entre X et Y est:

$$d_m(X, Y) = \sum_{i=1}^{n} |x_i - y_i|.$$

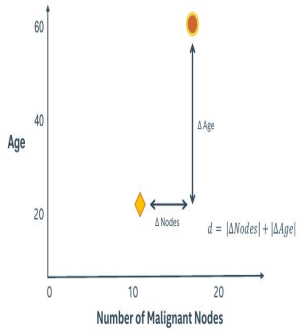


Distance Manhattan

Définition

C'est la distance qui calcule la somme des valeurs absolues des différences entre les coordonnées de deux points:
Soient $X = (x_1, x_2, ..., x_n)$ et $Y = (y_1, y_2, ..., y_n)$ la distance Manhattan entre X et Y est:

$$d_m(X, Y) = \sum_{i=1}^{n} |x_i - y_i|.$$



Distance de Minkowski

Définition

La distance de Minkowski ou métrique de Minkowski est une généralisation à la fois de la distance euclidienne et de la distance de Manhattan: Soient $X = (x_1, x_2, ..., x_n)$ et $Y = (y_1, y_2, ..., y_n)$ la distance Minkowski d'ordre p entre X et Y est:

$$d_M(X,Y) = \sqrt[p]{\sum_{i=1}^{n}(x_i - y_i)^p}.$$

Remarque

Si $p \to +\infty$ alors la distance de Minkowski nous obtenons la distance de Chebyshev: Soient $X = (x_1, x_2, ..., x_n)$ et $Y = (y_1, y_2, ..., y_n)$ la distance Minkowski d'ordre p entre X et Y est:

$$d_T(X,Y) = \lim_{p \to +\infty} \sqrt[p]{\sum_{i=1}^{n}(x_i - y_i)^p} = \max_i |x_i - y_i|.$$

Comment choisir la valeur K

- Le paramètre K dans l'algorithme KNN doit être choisi en fonction du jeu de données utilisé.

Comment choisir la valeur K

- Le paramètre K dans l'algorithme KNN doit être choisi en fonction du jeu de données utilisé.
- En règle générale, moins on utilisera de voisins (un nombre K petit) plus on sera sujette au sous apprentissage (underfitting).

Comment choisir la valeur K

- Le paramètre K dans l'algorithme KNN doit être choisi en fonction du jeu de données utilisé.
- En règle générale, moins on utilisera de voisins (un nombre K petit) plus on sera sujette au sous apprentissage (underfitting).
- Par ailleurs, plus on utilise de voisins (un nombre K grand) plus, sera fiable dans notre classification.
- Toutefois, si on utilise K nombre de voisins avec $K = N$ et N étant le nombre d'observations, on risque d'avoir du overfitting.

Limitations de KNN

- KNN est un algorithme assez simple à appréhender.
- Principalement, grâce au fait qu'il n'a pas besoin de modèle pour pouvoir effectuer une prédiction.
- Le contre coût est qu'il doit garder en mémoire l'ensemble des observations pour pouvoir effectuer sa prédiction. Ainsi il faut faire attention à la taille du jeu d'entrainement.
- Le choix de la méthode de calcul de la distance ainsi que le nombre de voisins k peut ne pas être évident. Il faut essayer plusieurs combinaisons et faire du tuning de l'algorithme pour avoir un résultat satisfaisant.

Application : Classification à l'aide de l'algorithme KNN

- Nous avons un ensemble d'objets avec les caractéristiques suivantes :Poids (en grammes) et couleurs.
- Notre objectif est de classer un nouvel objet en fonction de ses voisins les plus proches pour déterminer s'il s'agit d'un fruit ou d'un légume.
- Le nouvel objet a comme caractéristiques : Poids= 92g et Couleur = 1 (Rouge).

Application : Classification à l'aide de l'algorithme KNN

Objets	Poids	Couleurs	Classe	Distance Ecludienne
Cerise	43	1 (rouge)	Fruit	.....
Pomme	150	1 (rouge)	Fruit	.....
Poire	174	2 (jaune)	Fruit	.....
Brocoli	140	3 (vert)	Légume	.....
Laitue	92	3 (vert)	Légume	.....
Carott	165	4 (orangé)	Légume	.....

Application : Classification à l'aide de l'algorithme KNN

- Pour que les valeurs des données se situent dans une plage similaire, nous allons tout d'abord transformer les données en standardisant les poids et les couleurs de notre ensemble d'entrainement avant de calculer les distances euclidiennes.
- Moyenne et écart-type des caractéristiques : Moyenne Poids : 127.33 g, Écart-type Poids : 45.89 g, Moyenne Couleur : 2.33, Écart-type Couleur : 1.11 .

Objets	**Poids**	**Couleurs**	**Classe**	**Distance Ecludienne**
Cerise	-1,84	-1,2	Fruit	.....
Pomme	0,49	-1,2	Fruit	.....
Poire	1,02	-0,3	Fruit	.....
Brocoli	0,28	0,6	Légume	.....
Laitue	-0,77	0,6	Légume	.....
Carott	0,82	1,5	Légume	.....

Application : Classification à l'aide de l'algorithme KNN

Application : Classification à l'aide de l'algorithme KNN

- Nous allons calculer les distances euclidiennes entre le nouvel objet et chaque objet de cet ensemble, puis sélectionner les 3 plus proches voisins.

Application : Classification à l'aide de l'algorithme KNN

- Nous allons calculer les distances euclidiennes entre le nouvel objet et chaque objet de cet ensemble, puis sélectionner les 3 plus proches voisins.
- Le nouvel objet (Poids standardisé = -0.77, Couleur standardisée = -1.2)

Application : Classification à l'aide de l'algorithme KNN

- Nous allons calculer les distances euclidiennes entre le nouvel objet et chaque objet de cet ensemble, puis sélectionner les 3 plus proches voisins.
- Le nouvel objet (Poids standardisé = -0.77, Couleur standardisée = -1.2)

Objets	**Poids**	**Couleurs**	**Classe**	**Distance Ecludienne**
Cerise	-1,84	-1,2	Fruit	1,07
Pomme	0,49	-1,2	Fruit	1,26
Poire	1,02	-0,3	Fruit	2
Brocoli	0,28	0,6	Légume	2,08
Laitue	-0,77	0,6	Légume	1,8
Carott	0,82	1,5	Légume	3,13

Application : Classification à l'aide de l'algorithme KNN

- Nous allons calculer les distances euclidiennes entre le nouvel objet et chaque objet de cet ensemble, puis sélectionner les 3 plus proches voisins.
- Le nouvel objet (Poids standardisé = -0.77, Couleur standardisée = -1.2)

Objets	**Poids**	**Couleurs**	**Classe**	**Distance Ecludienne**
Cerise	-1,84	-1,2	Fruit	1,07
Pomme	0,49	-1,2	Fruit	1,26
Poire	1,02	-0,3	Fruit	2
Brocoli	0,28	0,6	Légume	2,08
Laitue	-0,77	0,6	Légume	1,8
Carott	0,82	1,5	Légume	3,13

- Les 3 plus proches voisins sont : Cerise (Fruit), Pomme (Fruit) et Laitue (Légume).

Application : Classification à l'aide de l'algorithme KNN

- Nous allons calculer les distances euclidiennes entre le nouvel objet et chaque objet de cet ensemble, puis sélectionner les 3 plus proches voisins.
- Le nouvel objet (Poids standardisé = -0.77, Couleur standardisée = -1.2)

Objets	Poids	Couleurs	Classe	Distance Ecludienne
Cerise	-1,84	-1,2	Fruit	1,07
Pomme	0,49	-1,2	Fruit	1,26
Poire	1,02	-0,3	Fruit	2
Brocoli	0,28	0,6	Légume	2,08
Laitue	-0,77	0,6	Légume	1,8
Carott	0,82	1,5	Légume	3,13

- Les 3 plus proches voisins sont : Cerise (Fruit), Pomme (Fruit) et Laitue (Légume).
- Conclusion: Selon l'algorithme KNN (K=3), comme la majorité des voisins (2 sur 3) sont des Fruits, le nouvel objet sera classé comme un Fruit.

Thank you for your attention.