



Chapitre 2

Classification et évaluation



Plan

- Rappel
 - Apprentissage
 - Apprentissage automatique
 - Types d'apprentissage automatique
- Classification
 - Définition
 - Ensemble d'apprentissage / Ensemble test
 - Approche paramétrique / Approche non paramétrique
 - Classification supervisée / Classification non supervisée
 - Réduction de multi-classes à deux classes
- Evaluation
 - Pourcentage de classification correcte
 - Matrice de confusion
 - Validation Croisée / Leave one out / Le bootstrap
 - Rappel / Précision
 - F_mesure
 - Sensibilité / Spécificité
 - Courbe ROC
 - Intelligibilité



Rappel



Apprentissage



Apprentissage

- Acquérir de nouvelles connaissances.
- Contracter de nouvelles habitudes.
- Avoir une connaissance extraite à partir d'un ensemble d'exemples.



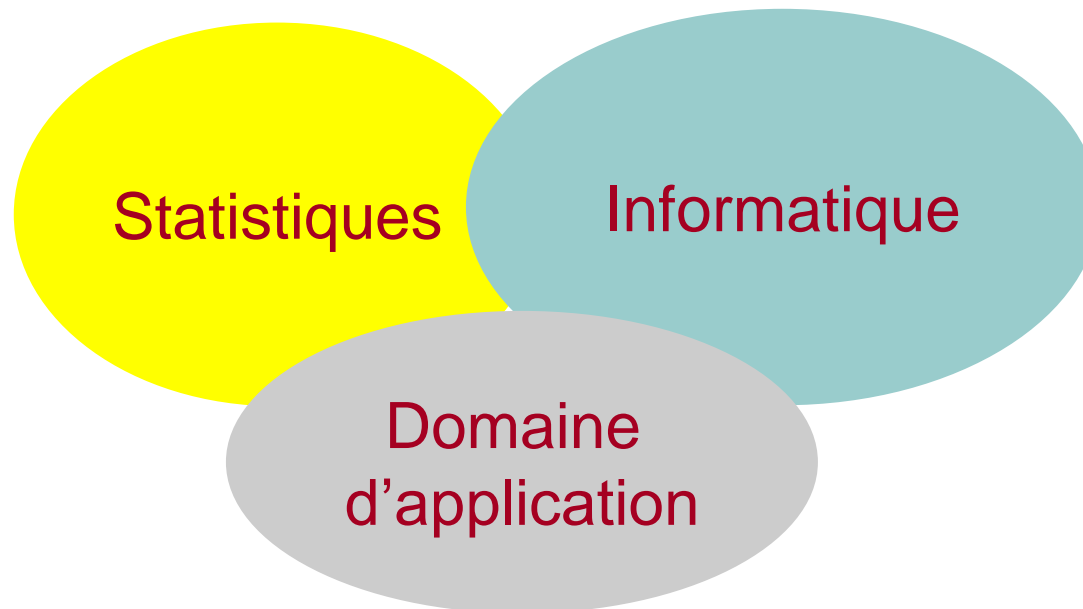
C'est la capacité d'améliorer
l'accomplissement d'une tâche en
interagissant avec un environnement.



Apprentissage automatique

Apprentissage automatique = Machine learning

- Simuler la cognition humaine.
- Doter la machine d'un mécanisme d'apprentissage.
- **Machine learning** = Intersection de l'informatique, statistiques et domaines particuliers





Domaines d'Application

Datamining

Exploiter les données historiques pour améliorer les décisions (Médecine, Banque,...)

Applications non classiques

(Reconnaissance de la parole, Conduite automatique,...)

Programmes auto-adaptatifs

(Lecture des journaux selon le thème,...)

Apprentissage supervisé (1)

- On dispose d'un ensemble de paires d'entrée(s)/sortie(s) de la forme (x_i, y_i) ,
 - x_i : entrée(s) possible(s) \Rightarrow Descriptions ou situations
 - y_i : sortie(s) associée(s) à x_i \Rightarrow Actions ou prédictions
- Les paires d'entrée(s)/sortie(s) sont appelées **les exemples** qui proviennent d'**une fonction inconnue**.
- Il s'agit de trouver une bonne approximation d'une fonction f dont on connaît le résultat que pour un certain nombre d'exemples.

On demande au système de généraliser

Exemples

- Une fonction h aussi proche que possible de f où $f(x_i) = y_i$

0	→	0
1	→	1
4	→	64
5	→	125



$$h(x) = x^3$$

- Une distribution de probabilité $P(x_i, y_i)$

Quelle est la probabilité qu'un client avec tel profil achète tel produit ?

- Dans un jeu de cartes:

- les cartes gagnantes sont: 9♥, Roi ♥ et 7♦.
- les cartes perdantes sont: 3♠, 4♣ et 6♣.



Les cartes rouges sont gagnantes et les cartes numériques noires sont perdantes

Apprentissage supervisé (2)

- Apprentissage supervisé avec **variable réponse continue**.



Régression, Estimation de densité

Prédire une variable **Y**
par les variables
explicatives **X1,**
X2,...Xn



Exemple de régression :

Prédire la valeur de la bourse demain,
étant données les valeurs des jours et mois passés

- Apprentissage supervisé avec **variable réponse discrète**.




Classification ou Analyse discriminante

- Apprentissage supervisé avec **variable réponse booléenne**.



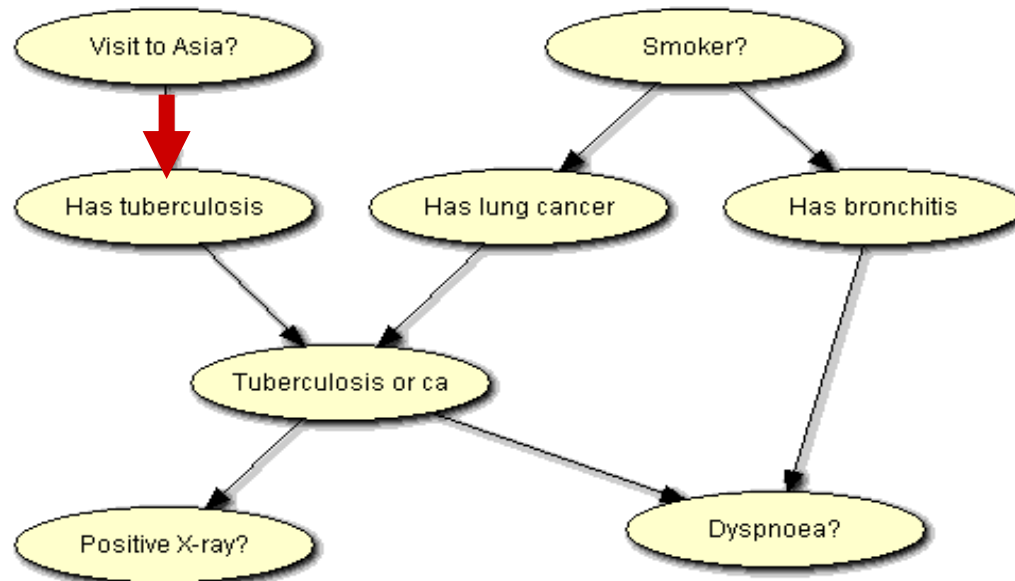
Apprentissage de concept

● ● ● | Apprentissage non supervisé

- On ne dispose pas d'un ensemble de paires d'entrée(s)/sortie(s).
 On dispose uniquement d'un ensemble d'entrées.
- Regrouper les entrées en un nombre fixe de groupes (**clusters**):
 - Les entrées de chaque groupe sont **proches** les uns des autres.
 - On utilise une certaine **métrique** dans l'espace des entrées.
- Découvrir de **nouvelles relations** dans les données (ex: **Réseaux Bayésiens**).

Exemples

- Segmentation du marché
 - Quelles sont les catégories principales des clients typiques dans le domaine vestimentaire?
 - Enfants, jeunes, adultes, etc.
 - Classique, habillé, sport, etc.
- Domaine médical: Découverte de nouveaux liens
 - Il y a un lien entre visiter l'Asie et attraper la tuberculose.



● ● ● | Apprentissage semi-supervisé

- L'apprentissage semi-supervisé utilise un ensemble de données étiquetées et non-étiquetés.



Apprentissage semi-supervisé peut améliorer les performances en combinant les données avec labels et sans labels.



Ne pas en laisser de côté des objets et utiliser toute l'information

● ● ● | Apprentissage par renforcement (1)

- L'algorithme d'apprentissage doit trouver une stratégie d'actions pour obtenir éventuellement une récompense (ou pénalité).

➡ La récompense ou la pénalité arrive (généralement) suite à un ensemble d'actions.



Maximiser le gain (ou inversement) à long terme
(apprentissage de réflexes, apprentissage de planification,...)

Exemples

- Jeu d'échecs

- On joue contre un adversaire.
- il y a une stratégie d'actions.
- C'est en fin de partie, qu'on va avoir le résultat de nos actions:
 - Victoire
 - Nul
 - Défaite

- Contrôle de robots (Aibo Taibo)

- Aibo réagit à son environnement, il peut s'adapter en vue d'obtenir plus de renforcements positifs.



Robot chien-jouet japonais



Apprentissage par renforcement (2)

- L'apprentissage par renforcement se distingue des autres approches d'apprentissage par plusieurs aspects :
 - L'apprentissage se fait sans supervision.
 - Il repose sur le principe d'essai/erreur.



N'oubliez pas

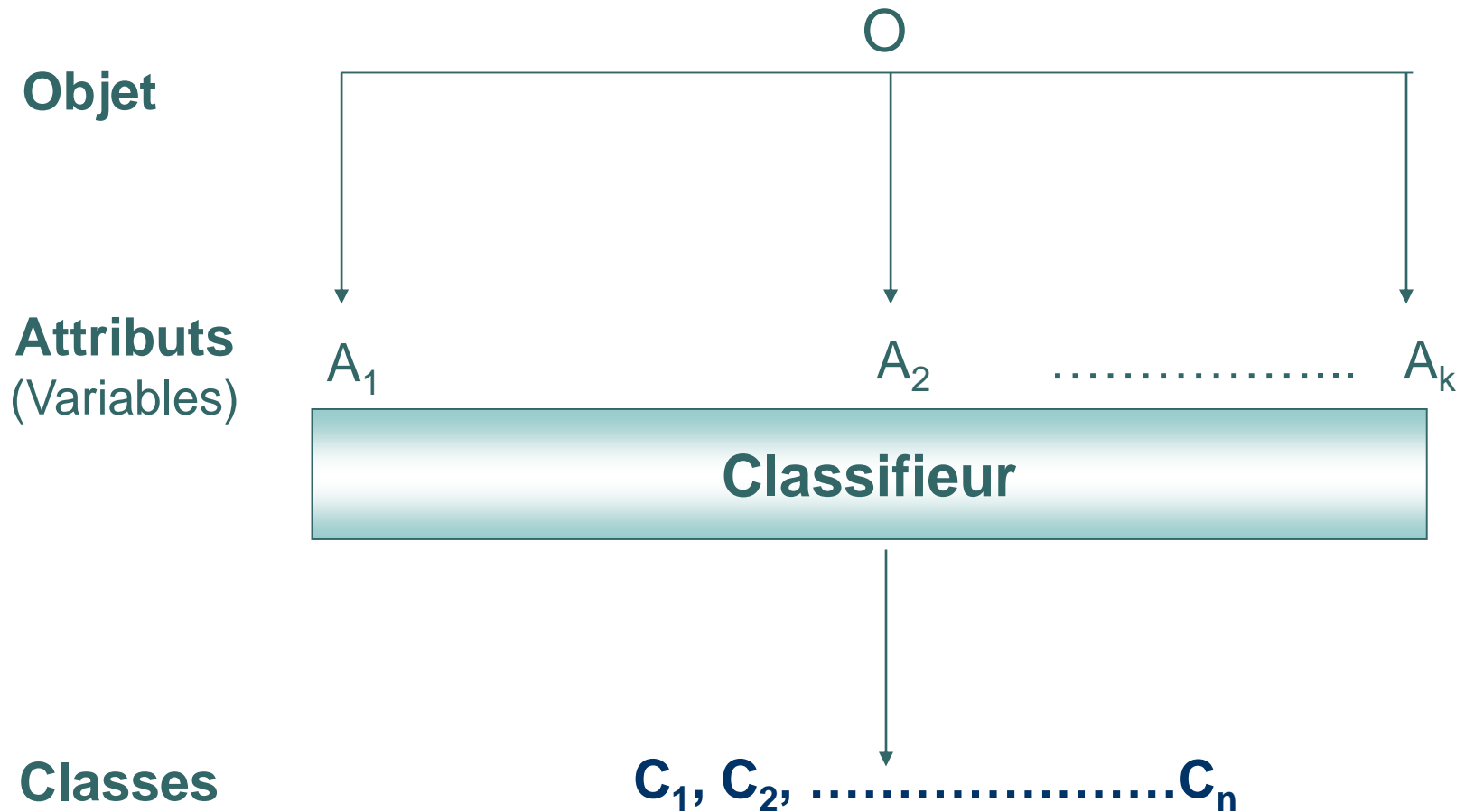


Parmi les tâches de l'apprentissage:
La classification

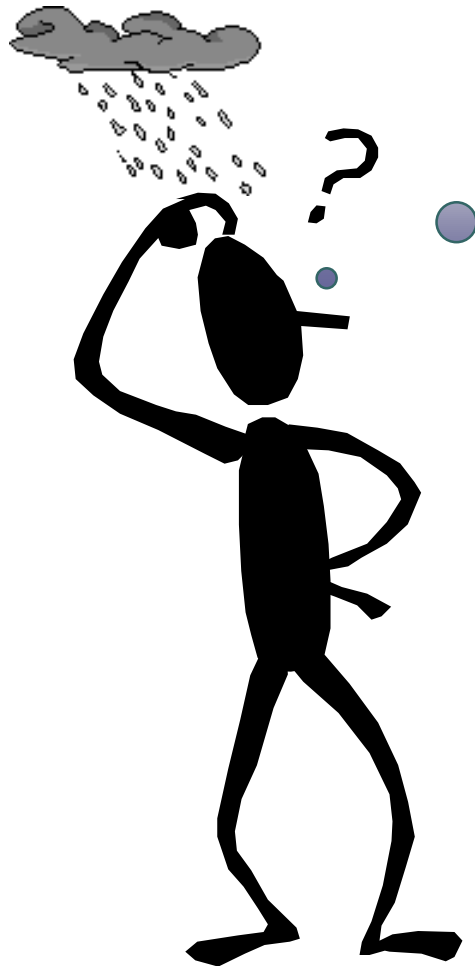


Classification

Classification

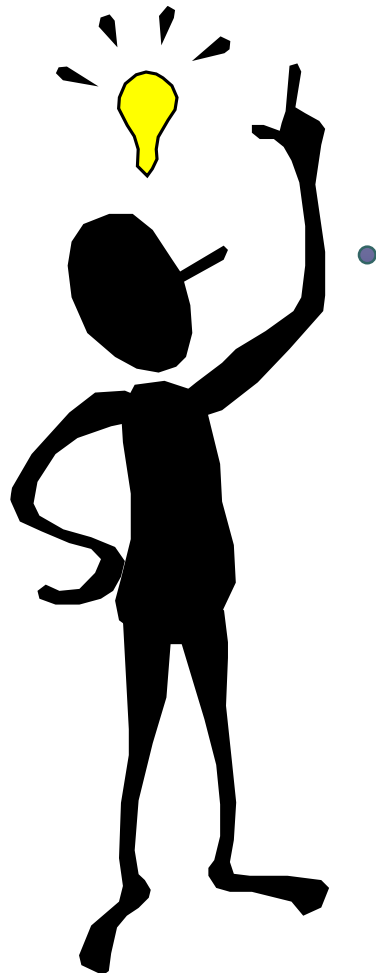


Question ?



Comment savoir si je dois
attribuer un crédit
bancaire à un client ?

Réponse



Mais bien sûr : apprendre
à partir d'exemples que la
banque connaît déjà !



Apprentissage par exemples

- On dispose d'un grand ensemble d'exemples (objets).
- On cherche à trouver une structure relative à ces exemples pour obtenir un modèle.
- Ce modèle permet de:
 - Extraire une procédure de **classification** à partir **d'exemples**.
 - **Classer** un nouvel exemple.
 -
 -
 - Prévoir une valeur numérique.
 - Comprendre la structure des exemples.

Ensemble d'apprentissage

Attributs

Valeurs des attributs

Revenu	Propriété	Crédit non remboursé	Classes
Elevé	Supérieur	Non	C_1
Elevé	Supérieur	Oui	C_2
Elevé	Supérieur	Non	C_1
Elevé	Inférieur	Oui	C_2
Moyen	Supérieur	Non	C_1
Moyen	Supérieur	Oui	C_2
Moyen	Inférieur	Non	C_2
Moyen	Inférieur	Oui	C_2
Faible	Inférieur	Non	C_3
Faible	Inférieur	Oui	C_3

C_1 : Attribuer tout le crédit.

C_2 : Attribuer une partie crédit.

C_3 : Ne pas attribuer le crédit.

Ensemble test

Revenu	Propriété	Crédit non payé	Classes
Elevé	Supérieur	Oui	?
Moyen	Inférieur	Non	?
Elevé	Supérieur	Oui	?
Moyen	Supérieur	Oui	?
Faible	Inférieur	Oui	?
Nul	Inférieur	Oui	?
Elevé	Supérieur	Non	?
Moyen	Inférieur	Oui	?

**On cache les
vraies classes**



Domaines d'application (1)

- Banque: attribution de crédits
 - Utiliser un historique de crédits accordés et non accordés avec la situation personnelle du client.
- Médecine: aide au diagnostic
 - Caractériser les symptômes des anciens patients et de leurs maladies.
- Marketing: établir un profil client
 - Faire une segmentation automatique des clients pour le marketing direct.



Domaines d'application (2)

- Analyse financière: prévision d'évolution des marchés
- Assurance: analyse de risques
- Télécoms: détection de fraudes
- Sécurité: détection d'intrusions

■

■

■

● ● ● | Approche paramétrique

- Proposition d'un modèle dont on estime ses paramètres à partir des exemples (phase d'apprentissage).
- Les hypothèses que l'on fait sur les lois de probabilité font partie d'une famille de distributions.
- Si on sait que P est une distribution normale, il suffit d'estimer ses deux paramètres:
 - Sa moyenne
 - Son écart type



Avoir **une bonne approximation** de la distribution P .



Déterminer une procédure de **classification**.



Approche non paramétrique

- Pas d'hypothèses sur le modèle que suivent les données.
- Les problèmes à résoudre sont plus complexes que ceux traités par les méthodes paramétriques.
- Méthodes statistiques.
- Méthodes issues de l'intelligence artificielle.



Classification supervisée

- Les classes sont définies **a priori**.
- Découverte de règles ou formules pour ranger les données dans des classes prédéfinies.
 - Construction d'un modèle sur les données dont la classe est connue (Ensemble d'apprentissage).
 - Utilisation pour classification des nouveaux objets.



Exemples

- Arbres de décision
- Méthodes K plus proches voisins
- Réseaux de neurones
- Machines à vecteurs supports (SVM)

-
-
-

● ● ● | Classification non supervisée

- Les instances d'apprentissage ne sont pas fournies avec des classes.



L'ensemble d'apprentissage n'est pas étiqueté (on ne connaît pas les classes a priori).

- Intuitivement les objets de même classe sont “proches” les uns des autres.



Mesure de similarité ou de distance

**Regrouper les exemples similaires:
Segmentation et clustering**



Exemples

- Clustering par partitionnement
- Clustering hiérarchique

-
-
-

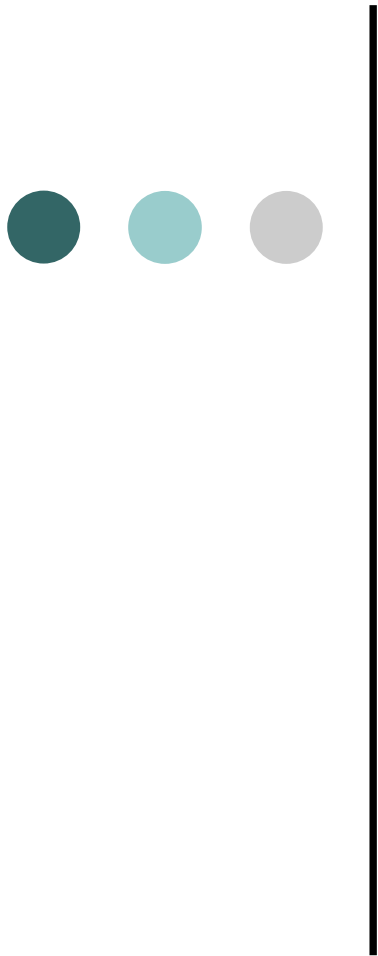
Réduction de multi-classes à deux classes

- Problème de classification à plusieurs classes.



Plusieurs problèmes de classification binaire.

- **OvA** (One vs. All) : C'est la méthode un contre tous mais qui n'est pas forcément la meilleure (classes souvent très déséquilibrées).
- **AvA** (All vs. All) : Pour chaque paire de classes (C1;C2), construire un classifieur pour discriminer entre C1 et C2 (Classes équilibrées mais on a $\approx n^2/2$ classifieurs).



Evaluation



Données d'apprentissage

- **Ensemble d'apprentissage** : ensemble des objets (exemples) utilisés pour générer le modèle d'apprentissage.
- **Ensemble test** : ensemble des objets (exemples) sur lequel sera appliqué le modèle d'apprentissage (pour tester et corriger l'algorithme).
- **Ensemble de validation** : peut être utilisé lors de l'apprentissage (comme sous population de l'ensemble d'apprentissage) afin de valider (intégrer) le modèle et d'éviter le sur-apprentissage.

Evaluation

- Utilisation d'un ensemble test.
- Pourcentage de classification correcte (PCC).



- Taux d'erreur de classification.
- Utilisation de la validation croisée, leave one out, etc.

Pourcentage de Classification Correcte (PCC) (1)

$$\text{PCC} = \frac{\text{Nombre d'objets correctement classés}}{\text{Nombre total des objets tests}}$$

Ensemble test

Revenu	Propriété	Crédit non payé	Classes prédites	Vraies classes
Elevé	Supérieur	Oui	C ₁	C ₁
Moyen	Inférieur	Non	C ₂	C ₂
Elevé	Supérieur	Oui	C ₁	C ₁
Moyen	Supérieur	Oui	C ₃	C ₂
Faible	Inférieur	Oui	C ₁	C ₃
Nul	Inférieur	Oui	C ₂	C ₃
Elevé	Supérieur	Non	C ₁	C ₁
Moyen	Inférieur	Oui	C ₃	C ₂

Pourcentage de Classification Correcte (PCC) (2)

Revenu	Propriété	Crédit non payé	Classes prédites	Vraies classes
Elevé	Supérieur	Oui	C ₁	C ₁
Moyen	Inférieur	Non	C ₂	C ₂
Elevé	Supérieur	Oui	C ₁	C ₁
Moyen	Supérieur	Oui	C ₃	C ₂
Faible	Inférieur	Oui	C ₁	C ₃
Nul	Inférieur	Oui	C ₃	C ₃
Elevé	Supérieur	Non	C ₁	C ₁
Moyen	Inférieur	Oui	C ₂	C ₂

$$PCC = \frac{6}{8} = 75\%$$

Taux d'erreur = 25%

Matrice de confusion

Classifieur			
Prédites \ Vraies	C_1 (4)	C_2 (2)	C_3 (2)
C_1 (3)	3	0	0
C_2 (3)	0	2	1
C_3 (2)	1	0	1

- Bon classifieur: sur les diagonales.
- Identifier les classes mal comprises (appries).
- Comparer les classifieurs selon la classe.
- Fixer des pénalités.



Validation croisée

- Partition de l'ensemble d'apprentissage T en k ensembles disjoints (T_1, T_2, \dots, T_k) de même taille $|T_i|$.
- Pour chaque $i = 1, 2, \dots, k$
 - 1- On fait l'apprentissage sur $T - \{T_i\}$
 - 2- On teste sur T_i
 - 3- On calcule le PCC sur T_i
- On fait la moyenne des PCC.



Leave one out

- Cette méthode est dérivée de la méthode de validation croisée, en prenant $k=n$, n étant le nombre d'exemples.
- A chaque itération, on va donc *faire l'apprentissage sur tous les exemples moins un*, et *tester sur un seul exemple*, afin de vérifier s'il est prédit correctement.



Le bootstrap

- Le *bootstrap* diffère des techniques précédentes qu'il utilise des tirages avec remise pour l'ensemble des exemples: on tire aléatoirement un exemple, pour le placer dans un ensemble appelé bootstrap, le procédé est répété n fois.
- Le *bootstrap* est très souvent utilisé dans le cadre de jeux de données contenant peu d'exemples.

Matrice de confusion

Classifieur		
Vraies Prédites	Oui	Non
	Oui	Non
Oui	VP	FN
Non	FP	VN

VP (Vrai Positif) = Nombre d'individus prédits positifs par le test et qui le sont effectivement.

FP (Faux Positif) = Nombre d'individus prédits positifs par le test mais qui sont en réalité négatifs.

VN (Vrai Négatif) = Nombre d'individus prédits négatifs par le test et qui le sont effectivement.

FN (Faux Négatif) = nombre d'individus prédits négatifs par le test mais qui sont en réalité positifs.

TP = VP, TN = VN

Rappel / Précision

Classifieur

Prédites \ Vraies	Oui	Non
Oui	VP	FN
Non	FP	VN

$$\text{PCC (Accuracy)} = \frac{VP + VN}{VP + FN + FP + VN}$$

- Parmi toutes les étiquettes positives possibles, combien d'entre elles le modèle a-t-il correctement identifiées.

$$\text{Rappel (Recall)} = \frac{VP}{VP + FN}$$

- Quelle est la fréquence à laquelle le modèle prédit correctement la classe positive.

$$\text{Précision (Precision)} = \frac{VP}{VP + FP}$$



F_mesure (F-measure)

- Moyenne harmonique entre la précision et le rappel

$$\text{F-measure} = \frac{2 * \text{Précision} * \text{Rappel}}{\text{Précision} + \text{Rappel}} = \frac{2VP}{2VP + FP + FN}$$

- Si Précision = Rappel alors F_mesure = Précision = Rappel = 1



Sensibilité / spécificité

- Taux de vrais positifs : proportion d'individus positifs effectivement bien détectés par le test.

$$\text{Sensibilité (Sensitivity)} = \frac{VP}{VP + FN}$$

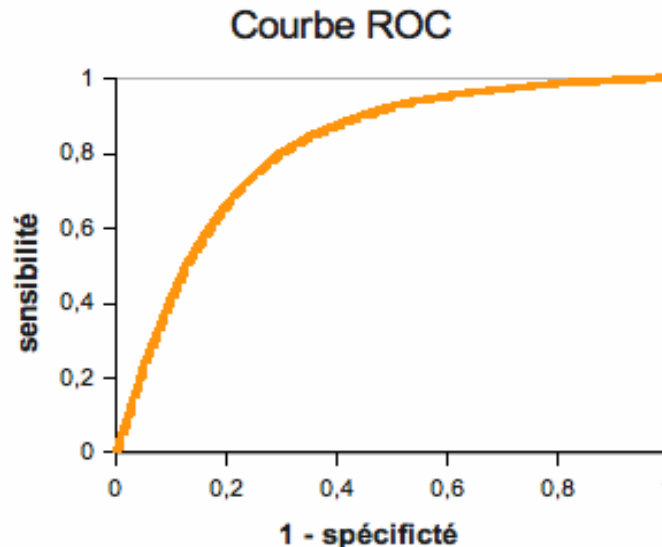
- Taux de vrais négatifs : proportion d'individus négatifs effectivement bien détectés par le test.

$$\text{Spécificité (Specificity)} = \frac{VN}{VN + FP}$$

Courbe ROC (1)

- Receiver Operating Characteristic -

- **Courbe ROC** est une représentation graphique de la relation existante entre la sensibilité et la spécificité d'un test pour chaque valeur seuil considérée.
- L'ordonnée représente la sensibilité et l'abscisse correspond à la quantité $(1 - \text{spécificité})$. Cette représentation fait varier le seuil de la probabilité qu'un exemple soit dans la classe positive.



Courbe ROC (2)

- Receiver Operating Characteristic -

- **Principe de la courbe ROC :**

Soit x_i positif, si $p(y = 1|x_i) > \text{seuil}$, sinon il est négatif ($y = 0$).

$$\hat{y}_i = 1 \iff p(y = 1|x_i) > \text{seuil}$$

- Si on prend $\text{seuil} = 0,5$, alors on aura une matrice de confusion, et donc on peut calculer la sensibilité et $(1 - \text{spécificité})$.
- Si on choisit un autre seuil (par exemple 0,6), on aura une autre matrice de confusion, et par conséquent de nouvelles valeurs de sensibilité et de $(1 - \text{spécificité})$.

➡ L'idée de la courbe ROC est de faire varier les seuils pour obtenir à chaque seuil, la sensibilité et $(1 - \text{spécificité})$.



Autre critère : Intelligibilité

- Améliorer la compréhension des résultats d'apprentissage.



Permettre au modèle de fournir une connaissance claire et compréhensible, au sens **interprétable**.



A suivre...

- Une technique de **classification** en **apprentissage supervisé**.



Arbres de décision