XAI
eXplainable Artificial Intelligence
IA explicable

Cours 4 - mardi 10 octobre 2023

Marie-Jeanne Lesot
Christophe Marsala
Jean-Noël Vittaut
Gauvain Bourgne

LIP6, Sorbonne Université

Au programme du jour

- Pour commencer : un petit complément sur LIME
  - tel qu'il est implémenté dans le gitlab

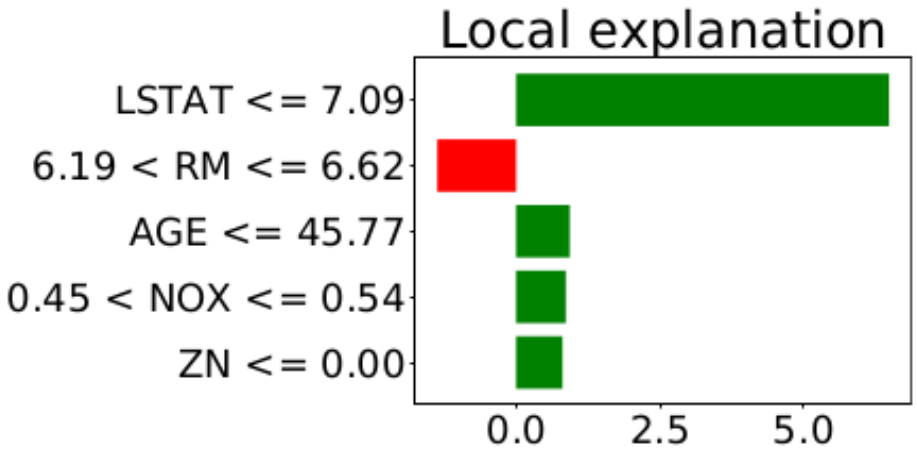
(Looking deeper into tabular LIME, D. Garreau et U. von Luxburg, 2020)

(Theoretical analysis of LIME, Damien Garreau, 2023)

- La théorie des sous-ensembles flous
- Logique floue

Construction d'attributs interprétables

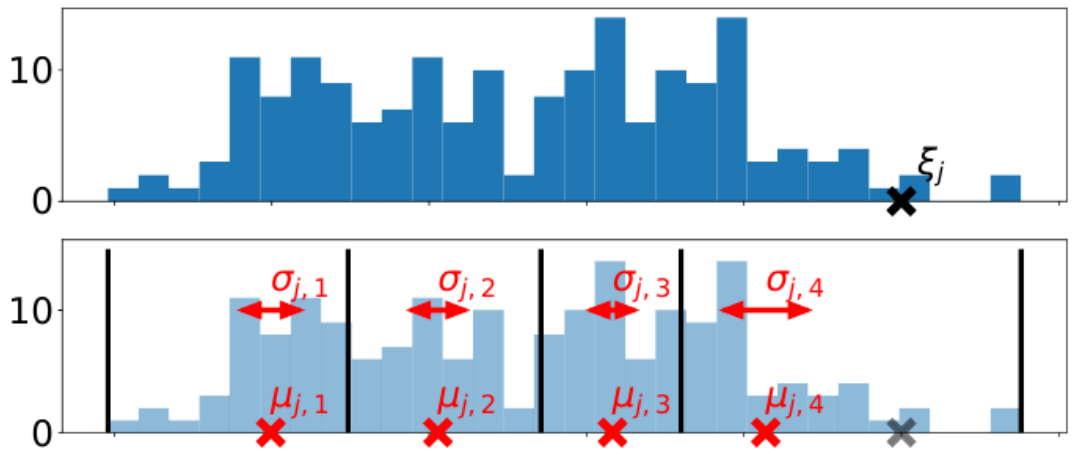
- Exemple de résultat
  - Boston housing



- Importance de **plages de valeurs** des attributs
  - plutôt que les attributs eux-mêmes

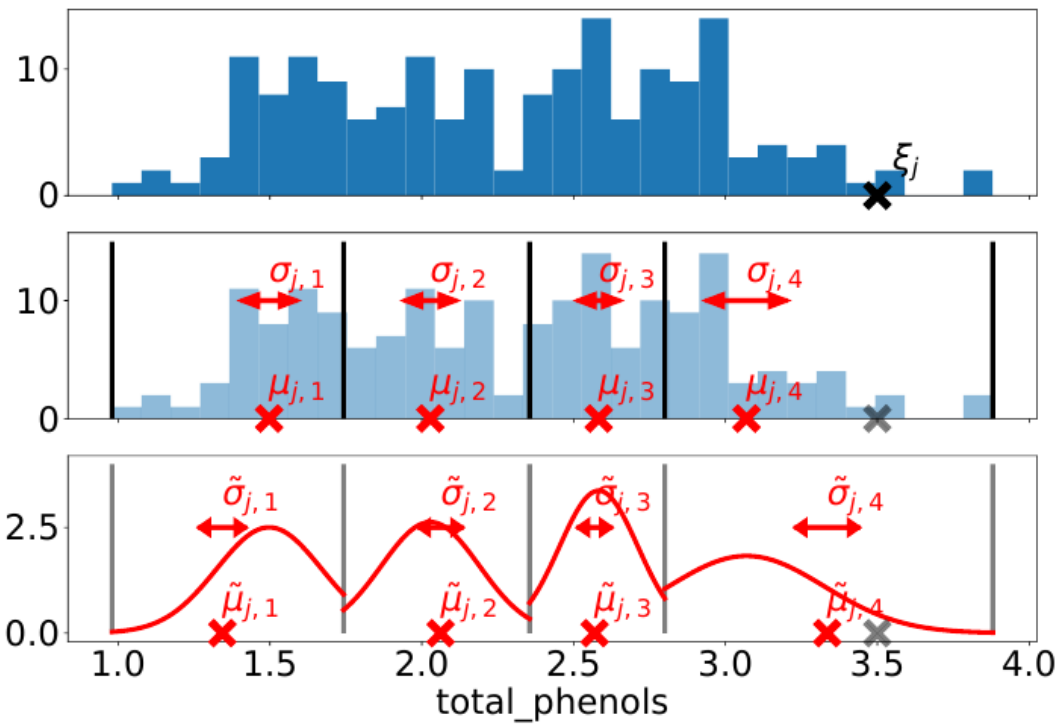
Construction d'attributs interprétables

- Discrétisation attribut par attribut
  - en $p = 4$ régions
  - de même fréquence : autant de données



Construction des données d'apprentissage

- Construction de $\mathcal{Z}$: attribut par attribut
  - sélection aléatoire d'une région
  - selon une gaussienne tronquée sur chaque région



Calcul des poids

- Distance par région
  - dépend des régions d'appartenance de x

$$\pi_x(z) = \exp\left(\frac{-1}{2\sigma^2}\sum_{j=1}^{d} 1_{b_{z,j} \neq b_{x,j}}\right)$$

- Paramètre à choisir : $\sigma = \sqrt{0.75d}$

Propriétés

(Garreau et von Luxburg, 20)

- Traitement indépendant des attributs
  - problème possible en cas d'attributs corrélés
  - problème possible de données non réalistes
- Dépendance aux régions
  - même explication pour des données de la même région
  - peut être vue comme un manque de fidélité au classifieur
- Dépendance au choix de σ
  - choix par défaut satisfaisant

Au programme du jour

- Pour commencer : un petit complément sur LIME
- **La théorie des sous-ensembles flous**
- Logique floue

Théorie des sous-ensembles flous et logique floue

- Formalisme pour représenter et manipuler des informations
  - de façon plus **intuitive, naturelle et interprétable**
  - que les formalismes classiques
    - théorie des ensembles
    - logique binaire

- Limitation des approches classiques
  - bornes qui semblent arbitraires
    - dates de péremption : heure de péremption ?
    - pourquoi ne pas proposer l'hôtel qui est à 16min de la plage ???
    - si je gagnais 38,13 euros de plus par mois ???
  - manque de robustesse

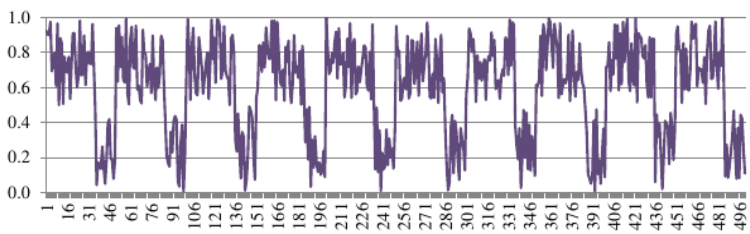
Théorie des sous-ensembles flous et logique floue

- Motivation : ***computing with words***
  - naturel pour un utilisateur
  - possibilité de personnaliser
  - cadre formel

Lotfi Zadeh

Applications et mises en œuvre

- **Apprentissage automatique flou**
  - clustering flou, arbres de décision flous, ...
  - résumés linguistiques flous : QRX sont P
    - Q quantifier, R qualifier, P summariser
    - exemples : quelques patients âgés ont un rythme cardiaque élevé la plupart du temps
    - pour les données relationnelles : **extraction en moins d'une seconde pour des données contenant des millions de tuples** (Smits et al. 19)
    - autre forme : résumé de période



- Approximately every 30 minutes, the data take high values

Applications et mises en œuvre

- **Bases de données avec requêtes floues**
  - augmenter l'expressivité des langages d'interrogation
  - faciliter l'interprétation des résultats de la requête
  - question d'exécution efficace : stratégies de couplage fort, faible ou intermédiaire

PostgreSQLf :

```
SELECT *,  get_mu() as mu
FROM cars
WHERE most(year ~= 'very recent', km ~= 'low, brand = 'VW')
ORBER BY mu LIMIT 10;
```

FUDGE :

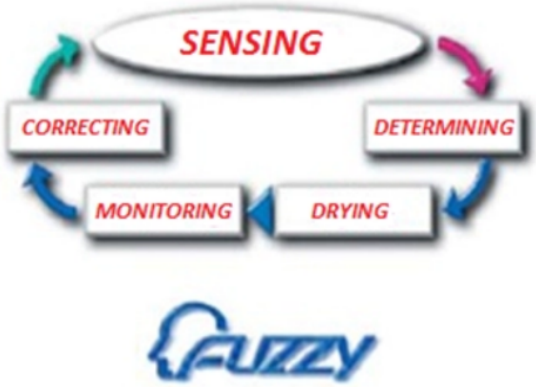
```
MATCH (x:author)-[authorof|ST IS strong]->(p:paper),
     (p:paper)-[:published]->(j:journal)-[:impactfactor]->(i:impactfactor),
     (j:journal)-[:domain]->(d) WHERE p.year IS recent
WITH x HAVING most(p) ARE (i.value IS high AND d.name="database")
RETURN x
```

Applications et mises en œuvre

- **Commande floue**
  - exemple, site web de Samsung https://www.samsung.com/in/support/home-appliances/what-is-fuzzy-logic-in-a-washing-machine/

What is Fuzzy Logic in a Washing Machine?

Last Update date : Oct 12. 2020



• **Fuzzy logic** washing machines are gaining popularity. These machines offer the advantages of **performance**, **productivity**, **simplicity**, **productivity**, and **less cost**. Sensors continually monitor varying conditions inside the machine and accordingly adjust operations for the best wash results. As there is no standard for fuzzy logic, different machines perform in different manners.

• Typically, fuzzy logic controls the washing process, **water intake**, **water temperature**, **wash time**, **rinse performance**, and **spin speed**. This optimises the life span of the washing machine. More sophisticated machines weigh the load (so you can't overload the washing machine), advise on the required amount of detergent, assess cloth material type and water hardness, and check whether the detergent is in powder or liquid form. Some machines even learn from **past experience**, **memorising programs** and adjusting them to **minimise running costs**.

• Most fuzzy logic machines feature '**onetouch control**.' Equipped with energy saving features, these consume less power and are worth paying extra for if you wash full loads more then three times a week. Inbuilt sensors monitor the washing process and make corrections to produce the best washing results.

• The **fuzzy logic** checks for the extent of dirt and grease, the amount of soap and water to add, direction of spin, and so on. The machine rebalances washing load to ensure correct spinning. Else, it reduces spinning speed if an imbalance is detected. Even distribution of washing load reduces spinning noise. **Neuro fuzzy logic** incorporates optical sensors to sense the dirt in water and a **fabric sensor** to detect the type of fabric and accordingly adjust wash cycle.

Applications et mises en œuvre

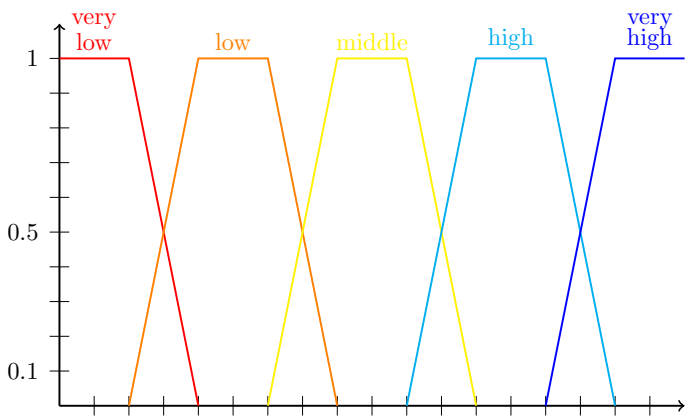
- Commande floue
  - véhicules : métro autonome (Hitachi), transmission automatique (Nissan), arrimage de la navette spatiale (NASA), contrôle de l'inclinaison des trains dans les virages (Mycom), hélicoptère sans pilote (Life)
  - électroménager : air conditionné (Mitsubishi), machine à laver, réfrigérateur, aspirateur, contrôle de température d'une douche (Matsushita)
  - photo : appareils photos, contrôle de sublimation pour impression couleur, caméra suivant un objet mobile, lecteur de codes-barres endommagés ou mal imprimés (LXE)
  - robotique : convoyeur en vitesse, bras robot pour la préhension d'objets de consistance variable (Omron)
  - usines : mélangeur en sidérurgie (Fujitsu), contrôle de production de papier (Yokogawa), cimenterie (Fuji)

Au programme du jour

- 1. La théorie des sous-ensembles flous
  - fonction d'appartenance et éléments caractéristiques
  - opérations ensemblistes
  - principe d'extension
- 2. Logique floue
  - valeurs de vérité
  - conjonction, disjonction, implication
  - raisonnement : Modus Ponens Généralisé

Fonction d'appartenance

- Généralisation de la fonction caractéristique d'un ensemble
  - pour modéliser une appartenance graduelle
- Univers de référence $\mathcal{U}$, sous-ensemble A
  - cas classique : $\chi_A : \mathcal{U} \longrightarrow \{0, 1\}$
  - cas flou : $f_A : \mathcal{U} \longrightarrow [0, 1]$
- Exemples
  - cas continu : $\mathcal{U} = \mathbb{R}^+$
  - cas discret : $\mathcal{U} = \{\text{auteurs du 19ème}\}$



auteurs célèbres =
{Balzac | 1 + Dickens | 1
+ Barbey d'Aurevilly | 0.4
+ ...}

- Sémantique : tout $x \in \mathcal{U}$ appartient **plus ou moins** à A

Sémantique

- Tout $x \in \mathcal{U}$ appartient **plus ou moins** à A
  - exemple : définition de chauve, riche, vieux, ...
  - transition progressive, gradualité
- Différence de sens avec les probabilités
  - probabilité $\sim$ incertitude

  $\neq$

  - sous-ensembles flous $\sim$ imprécision

  - autre exemple : perdu dans le désert
    - flacon A : probabilité(poison) = 0.6
    - flacon B : degré d'appartenance(poison) = 0.6

Eléments caractéristiques

- A : sous-ensemble flou de $\mathcal{U}$, défini par f_A
- Sous-ensembles classiques issus de A :
  - **noyau** : noy(A)= $\{x \in \mathcal{U} \mid f_A(x) = 1\}$
  - **support** : supp(A)= $\{x \in \mathcal{U} \mid f_A(x) \neq 0\}$
- Cardinal $|A|$: $|A| = \sum_{x \in \mathcal{U}} f_A(x)$
- Hauteur : $h(A) = \sup_{x \in \mathcal{U}} f_A(x)$
  - A est **normalisé** si $h(A) = 1$
- Deux sous-ensembles flous particuliers
  - $\mathcal{U}$ lui-même
  - l'ensemble vide $\emptyset$

Egalité et inclusion

- A et B : deux sef de $\mathcal{U}$

- **Egalité** : $A = B$ ssi ?
  - $\forall x \in \mathcal{U},\ f_A(x) = f_B(x)$

- **Inclusion** : $A \subseteq B$ ssi ?
  - $\forall x \in \mathcal{U},\ f_A(x) \leq f_B(x)$

Complémentaire, intersection, union

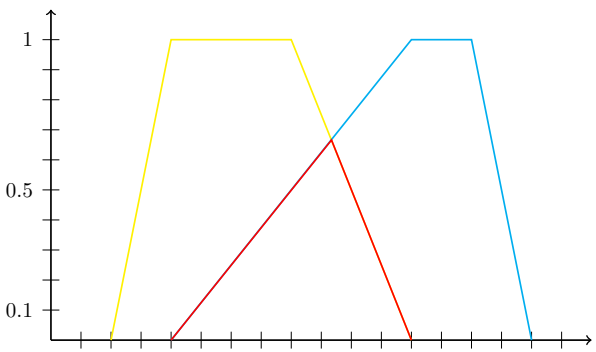
- Contrainte :
  - le cas classique doit être un cas particulier
  - si A et B sont crisp, on veut retrouver les mêmes résultats qu'en théorie des sous-ensembles classiques
- Complémentaire : $\forall x \in \mathcal{U},\ f_{A^c}(x) = 1 - f_A(x)$
- Intersection
  - Alphonse veut un hôtel autour de 75 euros la nuit
  - Berthe veut un hôtel pas trop cher, entre 60 et 70 euros
  - à quel point un hôtel à 72 euros leur convient à tous les deux ?
  - formellement : $\forall x \in \mathcal{U},\ f_{A\cap B}(x) = \min(f_A(x), f_B(x))$
    - intersection de Zadeh

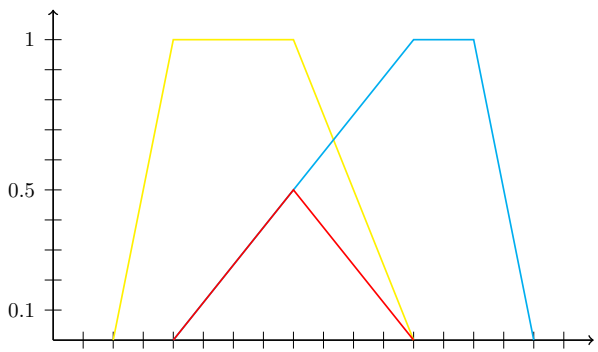
Propriétés souhaitées pour l'intersection

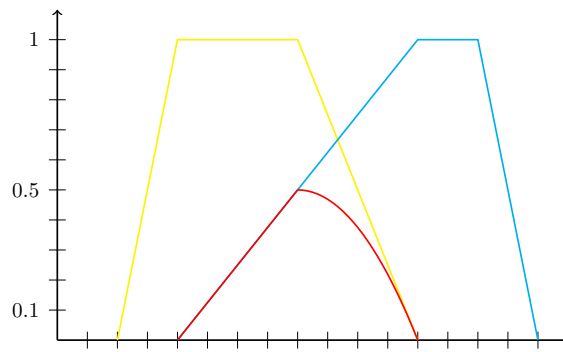
- Préserver les propriétés classiques :
  - commutativité : $A \cap B = B \cap A$
  - associativité : $A \cap (B \cap C) = (A \cap B) \cap C$
  - monotonie : si $A \subseteq B$ alors $(A \cap C) \subseteq (B \cap C)$
  - élément neutre : $A \cap \mathcal{U} = A$
- Conséquences
  - $A \cap B \subseteq A$ et $A \cap B \subseteq B$

t-normes

- Fonction $\top : [0,1] \times [0,1] \longrightarrow [0,1]$
  - commutative, associative, monotone
  - pour laquelle 1 est élément neutre : $\forall u \in [0,1],\ \top(u,1) = a$
- Opérateur d'intersection : $\forall x \in \mathcal{U},\ f_{A \cap B}(x) = \top(f_A(x), f_B(x))$
- Exemples
  - Zadeh $\min(u,v)$
  - Łukasiewicz $\max(0, u+v-1)$
  - produit $u \cdot v$







t-normes et propriétés de l'intersection

- Les propriétés classiques précédentes sont conservées
- D'autres sont relâchées :
  - pas nécessairement de non-contradiction : $A \cap A^c \neq \emptyset$
  - pas nécessairement d'idempotence : $A \cap A \neq A$

- Degré de sévérité (= exercices !)
  - la t-norme de Zadeh est la plus grande des t-normes
  - c'est la seule qui soit idempotente

Cas de l'union

- Gérée par dualité
  - pour préserver les lois de De Morgan
  - $(A \cup B)^c = A^c \cap B^c$
  - $(A \cap B)^c = A^c \cup B^c$
- **t-conormes** : $\perp \ : [0,1] \times [0,1] \longrightarrow [0,1]$
  - commutative, associative, monotone
  - 0 comme élément neutre : $\forall u \in [0,1],\ \perp(u,0) = u$
- Exemples :
  - t-conorme de Zadeh : $\perp(u,v) = \max(u,v)$
  - t-conorme probabiliste : $\perp(u,v) = u + v - uv$
  - t-conorme de Łukasiewicz : $\perp(u,v) = \min(1, u+v)$

Au programme du jour

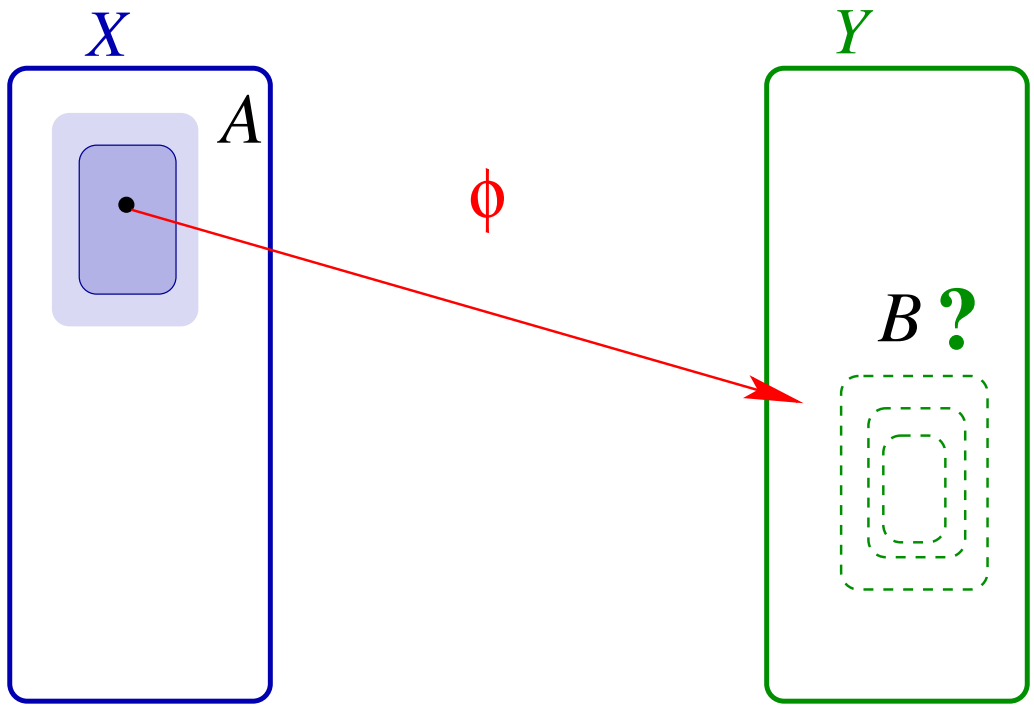
- 1. La théorie des sous-ensembles flous
  - fonction d'appartenance et éléments caractéristiques
  - opérations ensemblistes
  - **principe d'extension**
- 2. Logique floue
  - valeurs de vérité
  - conjonction, disjonction, implication
  - raisonnement : Modus Ponens Généralisé

Calculs avec des valeurs floues

- Problèmes considérés :
  - quel est le prix TTC d'un livre qui vaut **autour de 10** euros HT ?
  - quelle est la surface d'une pièce carrée d'**environ 3m** de côté ?
  - la vitesse est d'**environ 60** km/h, quelle distance est parcourue en un **peu moins d'un quart** d'heure ?
- Besoin d'étendre les fonctions classiques
  - pour autoriser des valeurs de paramètres floues
  - et en déduire des valeurs de sorties floues

Graphiquement

- Image **floue** d'un élément **flou** de X



Formellement

- Etant donné
  - deux univers X et Y
  - une fonction $\varphi : X \longrightarrow Y$
  - un sous-ensemble flou A de X
- Objectif : calculer l'**image de A par φ**
  - B, sef de Y : trouver la valeur $f_B(y)$ pour tout $y \in Y$
- Principe : utiliser les **antécédents de y** par φ
  - $\varphi^{-1}(y) = \{x \in X \mid y = \varphi(x)\}$
- Formule

$$\forall y \in Y, f_B(y) = \begin{cases} \sup\limits_{\{x \in X | y = \varphi(x)\}} f_A(x) & \textbf{si } \varphi^{-1}(y) \neq \emptyset \\ 0 & \textbf{si } \varphi^{-1}(y) = \emptyset \end{cases}$$

Au programme du jour

- 1. La théorie des sous-ensembles flous
  - fonction d'appartenance et éléments caractéristiques
  - opérations ensemblistes
  - principe d'extension
- **2. Logique floue**
  - valeurs de vérité
  - conjonction, disjonction, implication
  - raisonnement : Modus Ponens Généralisé

Le raisonnement naturel

- Connaissances imparfaites
  - règles **imprécises**
    - si vitesse **élevée** et obstacle **proche** alors freiner **fort**
    - et non : si $v \geq 42.0$ km/h et $d \leq 31.58$m alors $f = 9.6$N
  - connaissances **incertaines**
    - il est à **peu près sûr** que le métro arrive dans 2 mn
    - et non : la probabilité que le métro arrive dans 2 mn est 0.742
- Faits **ne correspondant pas** tout à fait aux règles
  - si le livre vaut moins de 10 euros, alors l'acheter
  - mais le livre vaut 10.25 euros
- Raisonner avec des **connaissances imprécises et incertaines**
  - la vérité des propositions n'est souvent pas **binaire**
  - $\Longrightarrow$ **plus ou moins vrai**, plus ou moins faux

Principe de la logique floue

- Structure $M = \langle \mathcal{D}, \bullet^M \rangle$, $\mathcal{D} = |M|$
  - **sens d'un prédicat : P^M sous-ensemble flou** de $\mathcal{D} \times \ldots \mathcal{D}$ défini par sa fonction d'appartenance

$$P^M : \mathcal{D} \times \ldots \mathcal{D} \longrightarrow \textbf{[0, 1]}$$

- **Valeur de vérité** de F : $[F]_v^M \in [0, 1]$
  - formule atomique
    si $F = P(t_1, \ldots t_n)$, alors $[F]_v^M = P^M([t_1]_v^M, \ldots, [t_n]_v^M)$
  - formule avec connecteur : utilisation des **opérateurs flous**
    si $F = F_1 \oplus F_2$, alors $[F]_v^M = op_\oplus([F_1]_v^M, [F_2]_v^M)$
  - formule avec quantificateur:
    utilisation du sup pour $\exists$
    du inf pour $\forall$