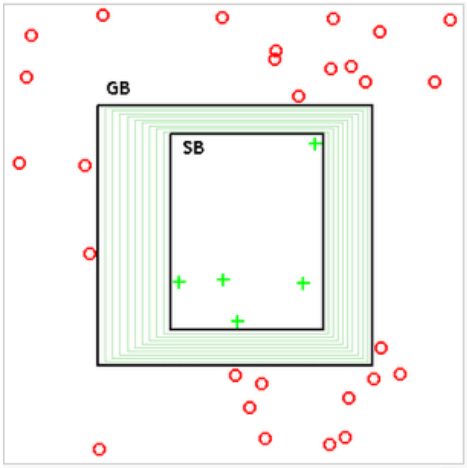
Cours d'apprentissage automatique – M1 WIC – 2011

Espace des versions

On a des exemples positifs et des exemples négatifs (contre-exemples), décrits par des descripteurs. L'idée est de trouver l'ensemble des hypothèses qui acceptent les exemples et rejettent les contre-exemples. Les hypothèses sont uniquement des conjonctions de descripteurs. Elles se situent entre la borne des hypothèses les plus générales et la borne des hypothèses les plus spécifiques.
Processus incrémental : on ajoute progressivement des exemples ou des contre-exemples.
On gère en permanence deux ensembles :
G : Ensemble des généralisations *les plus générales* qui soient complètes et cohérentes avec les instances présentées
S : Ensemble des généralisations *les plus spécifiques* qui soient complètes et cohérentes avec les instances présentées



Représentation des hypothèses de rectangles à partir d'exemples exemples positifs (les croix vertes, qui doivent être à l'intérieur du rectangle) et négatifs (les ronds rouges, qui doivent être à l'extérieur du rectangle). Le rectangle GB est l'hypothèses la plus **générale** (en généralisant plus on couvrirait des exemples négatifs), et SB est la plus **spécifique** (en spécialisant plus on ne couvrirait plus certains exemples positifs). Les rectangles verts représentent d'autres hypothèses de l'espace de versions.

Algorithme

Soit un nouvel exemple EX

- Généraliser les hypothèses de S qui rejettent EX en effectuant les modifications minimales (plus petit “pas” de généralisation)
    - Chaque hypothèses doit être une spécialisation d’une hypothèse de G
    - Aucune hypothèse ne doit être une généralisation d’une autre de S
- Supprimer de G les modèles ne couvrant pas EX

Soit un nouveau contre-exemple CE

- Spécialiser les hypothèses de G qui couvrent CE en effectuant les modifications minimales (plus petit “pas” de généralisation)
    - Chaque hypothèse doit être une généralisation d’une hypothèse de S
    - Aucune hypothèse ne doit être une spécialisation d’une autre de G
- Supprimer de S les hypothèses qui couvrent CE

Si S = G Alors SUCCES (convergence des deux bornes)
Si S ou G sont vides ECHEC (le concept n’est pas apprenable dans le langage actuel)

Exemple

Origin	Manufacturer	Color	Decade	Type	Example Type
Japan	Honda	Blue	1980	Economy	Positive
Japan	Toyota	Green	1970	Sports	Negative
Japan	Toyota	Blue	1990	Economy	Positive
USA	Chrysler	Red	1980	Economy	Negative
Japan	Honda	White	1980	Economy	Positive
Japan	Toyota	Green	1980	Economy	Positive
Japan	Honda	Red	1990	Economy	Negative

EX= [Japan, Honda, Bleu, 1980, Economy]
G = { (?, ?, ?, ?, ?) }
S = { (Japan, Honda, Blue, 1980, Economy) }

CEX= [Japan, Toyota, Green, 1970, Sports]
G = { (?, Honda, ?, ?, ?), (?, ?, Blue, ?, ?), (?, ?, ?, 1980, ?), (?, ?, ?, ?, Economy) }

S = { (Japan, Honda, Blue, 1980, Economy) }
EX = [Japan, Toyota, Blue, 1990, Economy]
G = { (?, ?, Blue, ?, ?), (?, ?, ?, ?, Economy) }
S = { (Japan, ?, Blue, ?, Economy) }

CEX = [USA, Chrysler, Red, 1980, Economy]
G = { (?, ?, Blue, ?, ?),(Japan, ?, ?, ?, Economy) }
S = { (Japan, ?, Blue, ?, Economy) }

EX= [Japan, Honda, Bleu, 1980, Economy]
G = { (Japan, ?, ?, ?, Economy) }
S = { (Japan, ?, ?, ?, Economy) }

EX= [Japan, Toyota, Green, 1980, Economy]
G = { (Japan, ?, ?, ?, Economy) }
S = { (Japan, ?, ?, ?, Economy) }

CEX= [Japan, Honda, Red, 1990, Economy]
S = {}
G = {}

Exercice 1

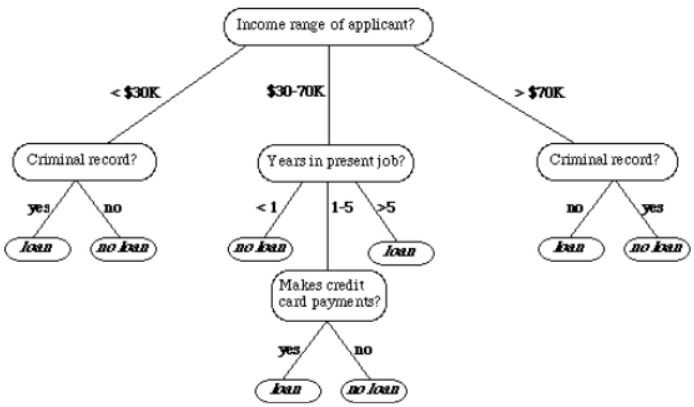
Voici une série d'exemples et contre-exemples décrivant les conditions météorologiques observées lors de la pratique d'un sport de plein air associées au jugement des pratiquants sur la qualité des conditions.

N°	Ciel	Tp-air	Humidité	Vent	Tp-eau	Prévision	Conditions
1	clair	chaude	normale	fort	chaude	stable	bonnes
2	clair	chaude	élevée	fort	chaude	stable	bonnes
3	pluvieux	fraîche	élevée	fort	chaude	variable	mauvaises
4	clair	chaude	élevée	fort	fraîche	variable	bonnes

1) En utilisant l'algorithme de l'Espace des Versions essayez de caractériser les situations que les personnes trouvent satisfaisantes pour pratiquer un sport ...
2) Une fois l'algorithme terminé donnez l'ensemble des généralisations possibles qui sont comprises entre la borne supérieure (G) et inférieure (S) du treillis des généralisations.

Induction d'arbres de décisions (ID3)

Exemple



Objectif : obtenir l'arbre le plus compact possible pour minimiser le nombre de questions permettant de classer un nouvel exemple.

Problème : comment choisir à chaque étape la variable qui permettra de séparer au mieux les données ? Un critère : minimiser l'entropie.

Méthode ID3

Entropie = Mesure du désordre = Σ -pi $\log_2$(pi)
(1,0,0,0) → 0
(½, ¼, ¼, 0) → 1,5
(¼, ¼, ¼, ¼) → 2

On explore toutes variables possibles. Pour chacune, on calcule l'entropie pour chacune de ses valeurs et on pondère par les probabilités de ces valeurs (estimées par les fréquence).
On choisit la variable qui a le gain d'entropie maximal.

Exemple :

```
size:   small medium large
colour: red blue green
shape:  brick wedge sphere pillar

%% yes

medium  blue    brick
small   red     sphere
large   green   pillar
large   green   sphere

%% no

small   red     wedge
large   red     wedge
large   red     pillar
```

Entropie initiale : -4/7 $\log_2$(4/7) – 3/7 $\log_2$(3/7) = 0,99 (grand désordre)

Si on coupe sur size :
- size=large : entropie = -2/4 $\log_2$(2/4) -2/4 $\log_2$(2/4) = 1
- size=small : entropie = -1/2 $\log_2$(1/2) – 1/2 $\log_2$(1/2) = 1
- size=medium : entropie = 0
- information attendue = 1*4/7 + 1*2/7 + 0*1/7 = 6/7 = 0,86
- Gain d'entropie = 0,99 – 0,86 = 0,13

Si on coupe sur color :
- Gain d'entropie = 0,52

Si on coupe sur shape :
- Gain d'entropie = 0,70

Donc on coupe sur shape

Et on recommence.

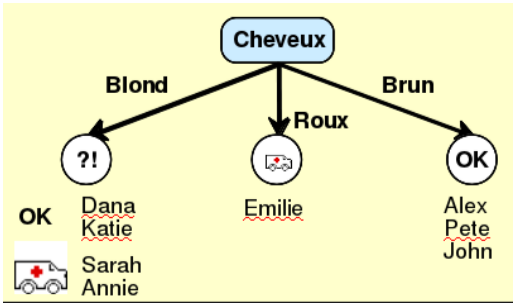
Exercice 2

Construire un arbre de décision sur le concept de coup de soleil.

	Taille	Poids	Cheveux	Lotion	Concept
Sarah	moyenne	léger	blonds	non	coup de soleil
Dana	grande	moyen	blonds	oui	non coup de soleil
Alex	petite	moyen	bruns	oui	non coup de soleil
Annie	petite	moyen	blonds	non	coup de soleil
Emily	moyenne	lourd	roux	non	coup de soleil
Pete	grande	lourd	bruns	non	non coup de soleil
John	moyenne	lourd	bruns	non	non coup de soleil
Katie	petite	léger	blonds	oui	non coup de soleil

Entropie initiale = $-3/8 \log_2(3/8) – 5/8 \log_2(5/8) = 0,95$
Si on coupe sur Taille :
gain d'entropie = $0,95 – (2/8*0 + 3/8*(-1/3\log_2(1/3)-2/3\log_2(2/3)) + 3/8*(-1/3\log_2(1/3)-2/3\log_2(2/3)) = 0,26$
Si on coupe sur Poids :
gain d'entropie = $0,95 – (2/8 *1 + 3/8*(-1/3\log_2(1/3)-2/3\log_2(2/3)) + 3/8 *(-1/3\log_2(1/3)-2/3\log_2(2/3)) = 0,01$
Si on coupe sur Cheveux :
gain d'entropie = $0,95 – (4/8 * 1 + 3/8 *0 + 1/8 *0) = 0,45$
si on coupe sur Lotion :
gain d'entropie = $0,95 – (5/8 * (-3/5*\log_2(3/5) – 2/5\log_2(2/5)) + 3/8 * 0) = 0,34$

on coupe donc sur Cheveux



Finalement on trouve :