Raisonner avec l'incertain : Les réseaux bayésiens

Matthieu Amiguet

2008 – 2009

haute école arc ingénierie

Probabilité

- La probabilité $P(a)$ d'un évènement a est un nombre dans l'intervalle $[0,1]$
- $P(a) = 1$ si l'évènement a est certain,
- $P(a) = 0$ s'il est certain que a ne se produit pas
- Si a et b sont des évènements mutuellement exclusifs et couvrent tous les évènements possibles, on a $P(a) + P(b) = 1$
- Si la probabilité décrit les résultats d'un test pouvant être répété un grand nombre de fois, l'interprétation/la détermination de ces nombres est claire
- Dans des cas ne pouvant être répétés, l'interprétation/la détermination de ces nombres est plus subjective.

Probabilité conditionnelle

- On note $P(a|b)$ la probabilité de a sachant b
  - "Étant donné l'évènement b, la probabilité de a est $P(a|b)$"
  - Attention ! ne signifie pas "chaque fois que b est vrai, la probabilité de a est $P(a|b)$" !
  - signifie plutôt "chaque fois que b est vrai *et que toutes nos autres connaissances ne sont pas pertinentes pour* a, la probabilité de a est $P(a|b)$"
- Remarque : dans un certain sens, toute probabilité est toujours "un peu" conditionnelle
  - $P(Dé = 6) = \frac{1}{6}$
  - Oui mais... à condition que le dé ne soit pas pipé !
  - Et que… et que… et que…

Une question piège – 1

- 1% des femmes de 40 ans participant à un contrôle de routine ont un cancer du sein.
- Lors d'une mammographie
  - 80% des femmes ayant un cancer du sein obtiennent un résultat positif
  - 9.6% des femmes n'ayant pas de cancer du sein obtiennent aussi un résultat positif ("faux positif")
- Une femme de 40 ans a obtenu une mammographie positive
- Quelle est la probabilité qu'elle ait un cancer du sein ?

Une question piège – 2

- La majorité des médecins répond que la probabilité est entre 70 et 80%
- Seuls 15% des médecins fournissent la réponse correcte : 7.8%
- Le nombre de réponses correctes monte à 46% avec cette présentation du problème :
  - 100 femmes de 40 ans sur 10'000 participant à un contrôle de routine ont un cancer du sein.
  - Lors d'une mammographie, 80 des 100 femmes ayant un cancer du sein obtiennent un résultat positif. . .
  - . . . et 950 des 9'900 femmes sans cancer obtiennent aussi un résultat positif.

Le théorème de Bayes – 1

- Le raisonnement ci-dessus peut être systématisé. Il est connu sous le nom de

Théorème de Bayes

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}$$

- Malheureusement, il arrive souvent que $P(a)$ ne soit pas connu.

Le théorème de Bayes – 2

- En utilisant l'égalité

$$P(a) = P(a|b) * P(b) + P(a|\neg b) * (1 - P(b)),$$

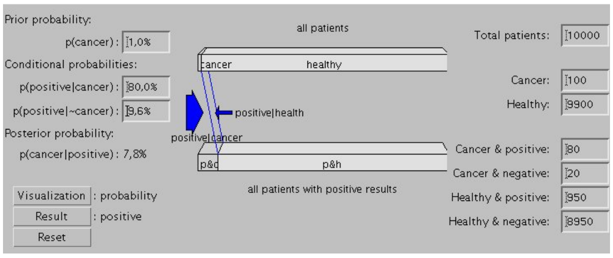
on peut aussi écrire

$$P(b|a) = \frac{P(a|b)P(b)}{P(a|b) * P(b) + P(a|\neg b) * (1 - P(b))}$$

- Donc la probabilité de b sachant a dépend
  - de la probabilité *a priori* de b (ici : P(cancer))
  - de la probabilité de a sachant b (ici : vrais positifs)
  - de la probabilité de a sachant $\neg b$ (ici : faux positifs).

Interprétation intuitive

- Intuitivement : les probabilités conditionnelles “poussent” les probabilités *a priori* dans le sens indiqué
  - Les applets de `http://yudkowsky.net/bayes/bayes.html` permettent de visualiser ce phénomène



Indépendance conditionnelle

Indépendance conditionnelle

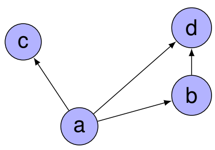
Les évènements A et C sont dits indépendants étant donné l’évènement B si la condition suivante est valable :

$$P(A|B) = P(A|B, C)$$

- En particulier, A et C sont dits indépendants si $P(A) = P(A|C)$
- La définition a l’air asymétrique… cependant, à l’aide du théorème de Bayes, on peut montrer qu’elle est symétrique (ie $P(A|B) = P(A|B, C) \Rightarrow P(C|B) = P(C|B, A)$).

Graphes dirigés

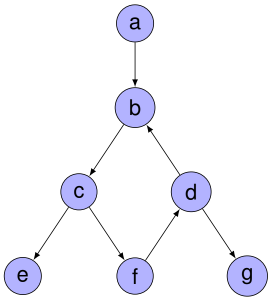
- Un *graphe dirigé* est constitué d’un ensembles de *noeuds* N et d’un ensemble d’arcs $A \in N \times N$



- Un graphe dirigé est dit *acyclique* s’il n’existe aucun chemin (dirigé) du type $N_1 \rightarrow N_2 \rightarrow \ldots \rightarrow N_1$
  - On dit parfois aussi DAG, pour *directed acyclic graph*

Contre-exemple

Un graphe dirigé cyclique



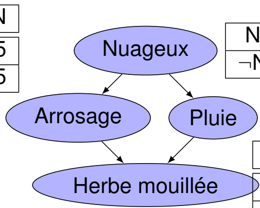
Réseaux bayésiens

Réseau bayésien

- Un réseau bayésien est composé d'un ensemble de *variables* et d'un ensemble d'arcs entre ces variables, tels que
  - Les variables et les arcs forment un graphe dirigé acyclique
  - Chaque variable possède un ensemble fini d'états mutuellement exclusifs
  - À chaque variable A ayant pour parents $B_1, \ldots, B_n$ est attachée une table de probabilité $P(A|B_1, \ldots, B_n)$

Exemple

	N	¬N
A	0.1	0.5
¬A	0.9	0.5



N	0.5
¬N	0.5

	N	¬N
P	0.8	0.2
¬P	0.2	0.8

	A		¬A	
	P	¬P	P	¬P
H	0.99	0.9	0.9	0.0
¬H	0.01	0.1	0.1	1.0

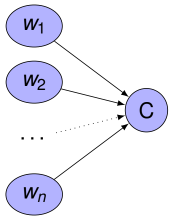
Remarques

- Chaque colonne des tables de probabilité doit avoir une somme de 1
  - Les états des parents sont mutuellement exclusifs
- Les flèches sont à interpréter comme des relations de *causalité*
  - Ceci n'est pas requis dans la définition, mais c'est ce qui donne les meilleurs résultats par la suite

Propagation de croyances

- Les tables de probabilité données permettent de “descendre” dans le réseau : si on observe les causes, on peut déduire la probabilité des effets
- En utilisant le théorème de Bayes, on pourra aussi “remonter” : calculer la probabilité des causes à partir de l’observation des effets
- On peut donc utiliser les réseaux bayésiens pour adapter nos degrés de croyances en fonction des observations
  - On peut propager aussi bien vers le “haut” que vers le “bas”
  - Par construction, les données seront cohérentes !

Cas particulier : le classifieur bayésien



- *n* “causes” *indépendantes* (ou presque !)
- 1 “conséquence” (la classification)
- Exemple : filtre à spam
- cf. par exemple `http://divmod.org/trac/wiki/DivmodReverend`

Quelques applications possibles

- Systèmes experts : médecine, ...
  - Recherche de la cause la plus probable
- Filtres à pourriel
  - Sachant qu’un pourriel a une probabilité donnée de contenir certains mots, comment déduire la probabilité qu’un mail donné soit un pourriel ?
- Traitement du langage
  - Désambiguïser le sens des mots en fonction du contexte

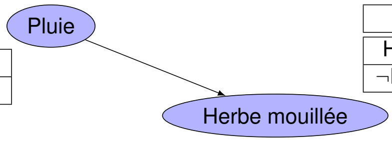
Quelques applications possibles (suite)

- Aide à l’utilisateur (à la “Clippy”)
  - Quelle est la probabilité que l’utilisateur ait besoin d’aide, en fonction de ses actions ?
- Commerce en ligne
  - Cibler des offres ayant de fortes chances d’intéresser un client donné...
- Robotique
  - Construction de la représentation du monde la plus probable en fonction des observations
- ...

Dépendance conditionnelle

- Considérons le réseau suivant :

P	0.8
¬P	0.2



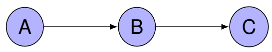
	P	¬P
H	0.99	0.1
¬H	0.01	0.9

- La table de droite nous dit que le fait d'observer qu'il pleut influence sur la probabilité que l'herbe soit mouillée
- Bayes nous dit que l'inverse est vrai aussi :
  - $P(P) = 0.8$
  - $P(P|H) = \frac{P(H|P)*P(P)}{P(H|P)*P(P)+P(H|\neg P)*(1-P(P))} = \frac{0.99*0.8}{0.99*0.8+0.1*0.2} =$ 0.975

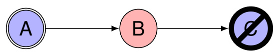
Dépendance conditionnelle

Le cas linéaire

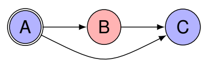
- De manière plus générale, dans une portion de graphe linéaire, tous les noeuds sont dépendants



- Mais si on *observe* l'état d'un noeud intermédiaire



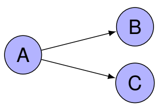
les noeuds extrêmes (A et C) deviennent indépendants !
  - Le fait d'observer A ne nous apprendra rien de nouveau sur C
  - Si ce n'est pas le cas, on a oublié un arc dans notre réseau !



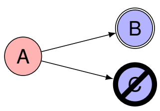
Dépendance conditionnelle

Le cas divergent

- Dans le cas d'une connexion divergente, les noeuds sont dépendants. . .



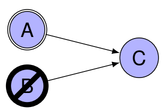
- Mais si on observe le parent, alors les enfants deviennent indépendants !



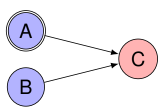
Dépendance conditionnelle

Le cas convergent

- Dans le cas d'une connexion convergente, les parents sont indépendants. . .



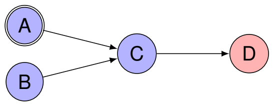
- Mais si on observe l'enfant commun, alors les parents deviennent dépendants !



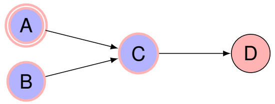
Dépendance conditionnelle

Le cas convergent avec descendant

- Dans le cas convergent, si un descendant de l'enfant commun est observé, cela suffit à rendre les parents dépendants !



- Pour mettre en évidence ce phénomène, on peut marquer spécialement les noeuds dont un descendant a été observé



Condition de dépendance conditionnelle

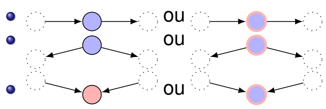
Dépendance conditionnelle

- Deux variables A et B d'un réseau bayésien sont dites *conditionnellement dépendantes* (ou *d-liées*) s'il existe un chemin (non-orienté) de A à B tel que, pour tout noeud C de ce chemin
  - Si C est linéaire dans ce chemin, C n'a pas été observé
  - Si C est divergent dans ce chemin, C n'a pas été observé
  - Si C est convergent dans ce chemin, ou bien C ou bien un descendant de C a été observé
- Si A et B ne sont pas conditionnellement dépendantes, elles sont dites *conditionnellement indépendantes* (ou *d-séparées*)

Condition de dépendance conditionnelle

La vision "Lego"

- Commencer par entourer en rouge tout les ancêtres d'un noeud observé
- A et B sont d-liés si on peut les relier en combinant les éléments suivants (avec une "fenêtre glissante" de trois noeuds) :



Attention

La question "A et B sont-ils d-liés ?" n'a de sens que si ni A ni B ne sont observés !

Couverture de Markov

Couverture de Markov

La couverture de Markov d'une variable A est constituée

- des parents de A
- des enfants de A
- des variables partageant un enfant avec A

Résultat

Si toutes les variables de la couverture de Markov de A sont observées, alors A est d-séparée du reste du réseau

Calculer les probabilités

- Si on a un ensemble de variables $U = \{A_1, \ldots, A_n\}$, on peut s'intéresser à toutes les combinaisons $P(A_i|A_{j_1}, \ldots, A_{j_k})$
  - Autrement dit : comment adapter nos croyances en fonction de l'information disponible ?
- On peut montrer qu'il "suffit" de connaître la table complète $P(A_1, \ldots, A_n)$ pour calculer toutes ces probabilités
- Oui mais... la taille de cette table augmente exponentiellement avec le nombre de variables (et de leurs valeurs)
  - Lourdeur des calculs
  - Quantité de données ingérable !

Une représentation compacte

- La propriété fondamentale des réseaux bayésiens est la suivante :

Probabilité jointe

$P(A_1, \ldots, A_n) = \prod_i P(A_i|\text{parents}(A_i))$

- Sur la base des tables spécifiées dans un réseau bayésien, on peut donc calculer la table de probabilité jointe $P(A_1, \ldots, A_n)$ et par conséquent toutes les probabilités conditionnelles
- Un réseau bayésien peut donc être vu comme une *représentation compacte* de la table de probabilité complète.

Indépendance conditionnelle

- Tout l'intérêt de la d-séparation étudiée ci-dessus repose dans le résultat suivant :

Réseau bayésien et indépendance

Les variables A et C sont d-séparées étant données les observations O

$$\Updownarrow$$

$$P(A|C,O) = P(A|O)$$

- En d'autres termes : deux variables sont d-séparées si et seulement si elles sont conditionnellement indépendantes au sens de la théorie des probabilités.

Le problème de la modélisation 33

- Les réseaux bayésiens fournissent un langage puissant pour modéliser des situations présentant un aspect incertain
- Cependant, même pour des réseaux simples, l'adaptation des probabilités en fonction des observations représente un calcul fastidieux
- Nous ne détaillerons pas dans ce cours la manière d'effectuer ces calculs
  - Nous nous contenterons d'utiliser un logiciel qui les effectue pour nous
- Il reste cependant une question centrale : comment construire un réseau représentant une situation donnée ?

Modélisation et *d*-séparation 34

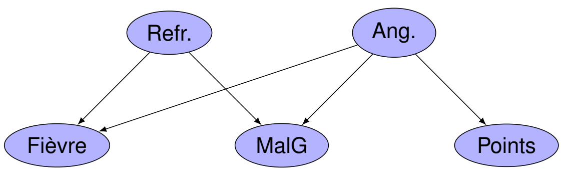
- Si on cherche à modéliser une situation par un réseau bayésien, de nombreux réseaux peuvent sembler faire l'affaire
- Pour choisir le plus adapté, on suivra avec profit les deux règles suivantes :
  1. Les flèches du réseau représentent la causalité directe et sont orientées de la cause à l'effet
  2. Les propriétés de *d*-séparation du réseau doivent correspondre aux propriétés d'indépendance conditionnelle du domaine modélisé

Situation 35

- Je me réveille ce matin avec un mal de gorge
- Cela pourrait résulter d'un début de refroidissement ou d'une angine
- Un refroidissement peut causer de la fièvre et des douleurs dans la gorge
- Une angine peut causer ces deux symptômes, et en plus des points jaunâtres dans la gorge

Le réseau 36

- Cette modélisation ne pose pas de difficulté majeure : il suffit de mettre les flèches dans le bon sens...



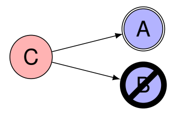
- À noter : dans ce réseau, le fait de savoir que l'on a une angine sépare le symptôme “points” des autres symptômes
  - Cette propriété est à vérifier auprès d'un spécialiste

Causalité

- Dans l'exemple ci-dessus, les liens de causalité sont assez évidents. . .
- . . . mais ce n'est pas forcément toujours le cas !
  - Il est parfois très difficile de distinguer une corrélation d'une causalité !
- Dans certains cas, “l'expérience de pensée” suivante peut aider :
  - Soient A et B deux variables corrélées mais dont on ignore si l'une cause l'autre ou inversement
  - Imaginons que quelqu'un d'extérieur a la possibilité de fixer la valeur de A ; si ceci n'a pas pour effet de changer notre croyance de B, alors A ne cause pas B

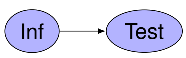
Cause commune

- Si A et B sont corrélés mais aucun de cause l'autre, il se peut qu'on ait oublié une variable : la cause commune de A et B
- Si on trouve un candidat C pour la cause commune, on peut vérifier que A et B deviennent indépendant étant donné C :



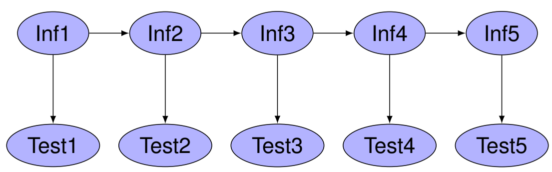
La version “hors temps”

- Une vache malade peut produire du lait infecté
- On dispose d'un test permettant de détecter cette infection dans le lait
  - Le test présente un certain taux de *faux positifs* et de *faux négatifs*
- Le réseau ne pose pas de problème :



Un temps discret

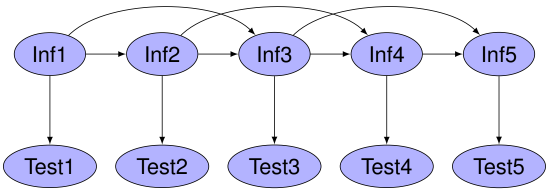
- Si le fermier effectue un test chaque jour, on peut tenir compte de la temporalité dans le réseau :



- Un tel réseau est dit *markovien* : la connaissance de l'état courant détermine entièrement l'avenir
  - $Test_i$ est d-séparé du reste du réseau lorsque Inf_i est observé

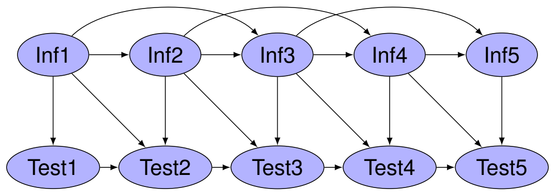
Un peu de mémoire. . .

- Il se peut que cette *d*-séparation ne soit pas satisfaisante :
  - Sachant qu'une vache est malade aujourd'hui, le fait qu'elle l'ait été ou non hier peut influencer sur sa probabilité de l'être demain. . .



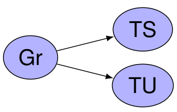
Un peu *plus* de mémoire. . .

- Dans la version précédente, deux tests successifs sont encore *d*-séparés étant donné l'état de *Inf*
  - Autrement dit, la réussite d'un test aujourd'hui ne dépend pas de celle d'hier si je sais que la vache est malade
  - Ceci peut être correct ou non suivant les cas (cf. expert. . .)
- Une solution possible :



Le problème

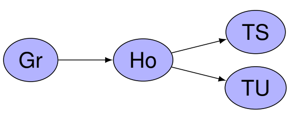
- Six semaines après l'insémination d'une vache, on peut faire deux tests de grossesse : Un test sanguin et un test urinaire



- Dans ce réseau, TS et TU sont séparés étant donné G
  - On pose la question à un expert : "Supposons que l'on sache que la vache est enceinte ; si on obtient un test sanguin négatif, cela va-t-il influencer nos croyances sur le résultat d'un test urinaire ?"
  - Il se trouve que la réponse est oui. . . ces variables ne devraient pas être séparées !

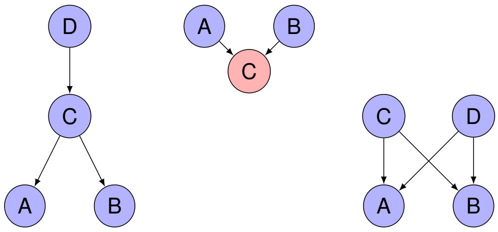
Alors. . . que faire ?

- La première idée qui vient à l'esprit est de relier directement TS et TU
  - Oui mais, dans quelle direction ? la causalité n'a rien d'évident !
- Là encore, la connaissance de l'expert peut nous aider : les deux tests servent en fait à détecter des changements hormonaux
- On va donc introduire une variable supplémentaire, ou *variable intermédiaire* Ho



Autres possibilités

- Quelques exemples de situations où une variable C peut "résoudre" des corrélations non causales entre A et B :



Problématique

- Jusqu'à maintenant, nous avons considéré des cas où
  - un expert (et un cogniticien ?) modélise(nt) le domaine concerné sous forme de réseau bayésien ;
  - on utilise ensuite le réseau pour faire des prévisions ou des calculs de probabilités
- Dans certains cas, on dispose déjà d'une base de données permettant de déduire des probabilités
  - On peut alors chercher un réseau dont la structure reflète au mieux la situation observée

Apprentissage

- Le problème ci-dessus est donc un problème d'*apprentissage*
- On distingue habituellement trois types d'apprentissage dans les réseaux bayésiens
  - Apprentissage par lots
  - Adaptation
  - *Tuning*
- Tous ces algorithmes font encore l'objet de recherches actives et aucun ne peut être considéré comme "standard"
  - Nous ne verrons donc pas ces algorithmes en détail !

Apprentissage par lots

ou *batch learning*

- On dispose d'une grande base de données et on désire construire un réseau bayésien qui représente ces données au mieux
  - On connaît déjà les noeuds du graphe, mais on veut trouver les arcs (et en déduire les tables)...
- En principe, la théorie des probabilités nous donne facilement la réponse
- En pratique, les algorithmes directs sont *largement* trop lourds
  - Il "suffirait" de parcourir tous les graphes possibles et d'étudier les conséquences... mais il y en a *beaucoup* trop !

Apprentissage par lots

Idée générale

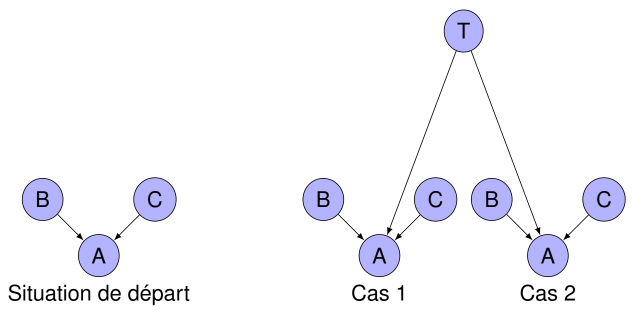
1. On exprime des *contraintes* pour réduire le nombre de graphes possibles
   - Relations de causalité (un symptôme ne va jamais causer une maladie...)
   - Relations d'indépendance conditionnelle affirmées par un expert
   - ...
2. On va ensuite parcourir l'espace des graphes restants par "réparation"
   - On ajoute, modifie, ou retire des arcs pour essayer d'obtenir la distribution la plus proche des données à disposition...

- À noter que ceci ressemble furieusement à un problème de recherche dans un espace d'états ! (A*, ...)

Adaptation

- Dans d'autres cas, on peut disposer d'une structure de réseau connue, mais ne pas être sûr des valeurs des tables de probabilité
  - Souvent, on dispose plutôt d'une *plage de valeurs* possibles
  - Par exemple dans le cas où un produit doit pouvoir s'adapter à différents contextes
  - Cette incertitude est qualifiée d'*incertitude du second ordre*
- Un solution est de représenter explicitement le contexte par un nouveau noeud et d'adapter au fur et à mesure notre "croyance" sur ce noeud

Adaptation – exemple



Tuning

- Le *tuning* s'applique dans des situations où
  - la structure du réseau est déjà fixée
  - les probabilités dépendent de paramètres
  - On connaît déjà certaines probabilités conditionnelles
- Le problème est alors de trouver les valeurs des paramètres qui "collent" le mieux aux valeurs connues
- On utilise des techniques de calcul différentiel pour se déplacer dans l'espace des paramètres de manière à minimiser la distance entre les probabilités connues et les probabilités calculées. . .

Graphes de décision

- Jusqu'à maintenant, nous avons considéré des réseaux permettant de calculer les probabilités de différents évènements
- La question de savoir comment utiliser cette information pour prendre des décisions restait une question *meta* :
  - La réflexion sur les actions est une réflexion *sur* le réseau, pas *dans* le réseau
- À noter qu'on peut distinguer (en tout cas) deux types de prise de décision :
  - Une décision de test : quel est le prochain test à effectuer pour accroître utilement ma connaissance ?
  - Une décision d'action : comment vais-je agir dans le monde pour avoir une bonne probabilité d'obtenir l'effet recherché ?

Action et observation

- Notons qu'on peut distinguer deux types d'actions :
  - Les actions *internes*, qui modifient l'état de certaines variables du réseau
  - les actions *externes*, dont l'impact n'est pas modélisé dans le réseau

Attention !

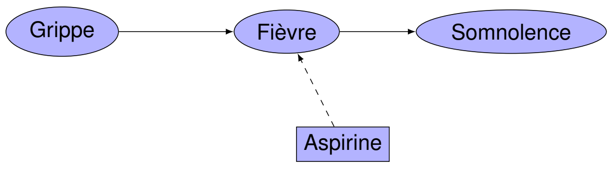
Il y a une différence capitale entre observation et action interne !

- Une observation propage des nouvelles probabilités en aval *et en amont* du noeud observé
- Une action interne ne peut modifier que des noeuds en aval du noeud affecté.

Action et observation

Un exemple

- On sait que l'observation de *Fièvre* modifie les probabilités de *Grippe* et de *Somnolence*
- La prise d'une aspirine diminuera directement la fièvre et aura donc une influence sur la somnolence, *mais pas sur la grippe*



Utilité

- Pour quantifier l'effet des actions effectuées dans un réseau bayésien, on peut introduire des noeuds d'utilité
  - Noeuds en forme de losange
  - Les états sont des valeurs numériques représentant l'utilité de cet état
- On peut donc chercher à maximiser la probabilité des états d'utilité maximale.

Utilité attendue

Avec un seul noeud de décision

- Pour ramener la prise de décision à l'intérieur du modèle, on peut introduire la notion d'*utilité attendue* (*expected utility*) d'une décision :

Utilité attendue

Soient $X_1, \ldots, X_n$ les utilités d'un réseau bayésien, D un noeud de décision et O l'ensemble des observations effectuées. L'utilité attendue de D est

$$EU(D|e) = \sum_{X_1} U_1(X_1)P(X_1|D,e) + \ldots + \sum_{X_n} U_n(X_n)P(X_n|D,e)$$

Arbres de décision

- Si on a plusieurs décisions successives à prendre, une solution est de les représenter sous forme d'arbre
- On peut alors propager les utilités attendues des feuilles vers la racine...
- ... dans un algorithme qui ressemble beaucoup à un *Minimax* probabiliste (mais où on maximise à chaque étage) !

Conclusion – 1

- Les réseaux bayésien couvrent un domaine très large...
  - Du plus appliqué (commerce électronique, filtres à spam, ...)
  - Au plus philosophique (inférence de causes, explication de la méthode scientifique ou de l'apprentissage humain (!), ...)
- ... sur un large ensemble de techniques
  - propagation de croyances
  - apprentissage structurel ou quantitatif
  - aide à la décision

Conclusion – 2

- Certaines techniques sont mûres et bien connues
  - propagation *exacte* de croyances
- Et d'autres font encore l'objet de recherche active
  - propagation approchée (mais efficace) de croyances
  - apprentissage
  - ...
- L'un dans l'autre, un domaine en pleine évolution promis à un bel avenir !

Sources

- Finn V. Jensen, “Bayesian Networks and Decision Graphs”, Springer, 2001
- Judea Pearl, “Causality / Models, Reasoning and Inference”, Cambridge University Press, 2000
- `http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html`
- `http://yudkowsky.net/bayes/bayes.html`