IFT 615 – Intelligence Artificielle

Hiver 2022

Raisonnement probabiliste

Inférences avec une distribution conjointe et classifieur bayésien naïf

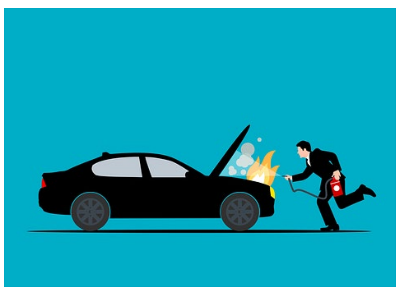
Professeur: Froduald Kabanza

Assistants: D’Jeff Nkashama & Jean-Charles Verdier

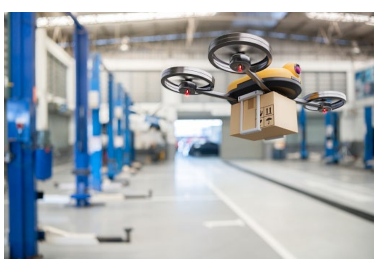
Université de Sherbrooke

Motivation

Détection de pourriels
Classification de documents

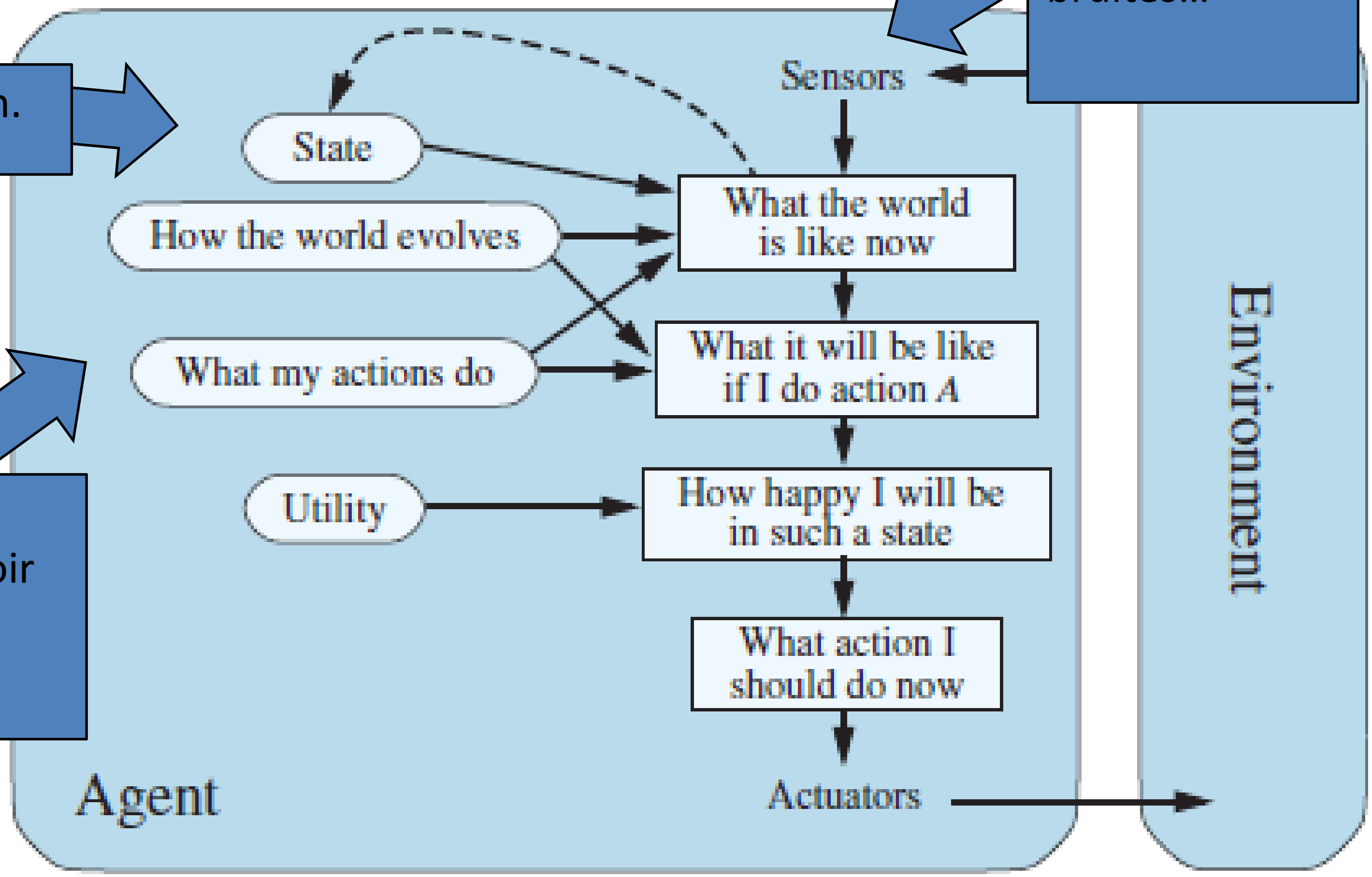


Assurance de dommages



Localisation robotique

Utility-based agents



Les capteurs peuvent être bruités…

État incertain.

Les actions peuvent avoir des effets incertains !

Théorie des probabilités en IA

- Permet de modéliser la vraisemblance d'événements
  - l'information sur la vraisemblance est dérivée
    - » des croyances/certitudes d'un agent, ou
    - » d'observations empiriques de ces événements

- Donne un cadre théorique pour mettre à jour la vraisemblance d'événements après l'acquisition d'observations

- Facilite la modélisation en permettant de considérer l'influence de phénomènes complexes comme du « bruit »

Decision-Theoretic Agent

function DT-AGENT(*percept*) **returns** an *action*
  persistent: *belief_state*, probabilistic beliefs about the current state of the world
    action, the agent's action

  update *belief_state* based on *action* and *percept*
  calculate outcome probabilities for actions,
    given action descriptions and current *belief_state*
  select *action* with highest expected utility
    given probabilities of outcomes and utility information
  return *action*

Figure 12.1 A decision-theoretic agent that selects rational actions.

Sujets couverts

- Inférence probabiliste avec une distribution conjointe
- Classifieur bayésien naïf

Exemple – Détection de pourriels

- On souhaite raisonner sur la possibilité qu'un courriel soit un pourriel tenant compte de l'incertitude associée à une telle classification

- Pour ce faire, notre modèle (« base de connaissances ») est une **distribution conjointe des probabilités** de variables aléatoires
  - ***Inconnu*** : l'adresse de l'expéditeur n'est pas connue du destinataire
  - ***Sensible*** : le courriel contient un mot sensible
  - ***Pourriel*** : le courriel est un pourriel

	Inconnu = vrai		*Inconnu = faux*	
	Sensible = vrai	*Sensible = faux*	*Sensible = vrai*	*Sensible = faux*
Pourriel = vrai	0.108	0.012	0.072	0.008
Pourriel = faux	0.016	0.064	0.144	0.576

Distribution de probabilités

- **Distribution de probabilités :** l’énumération des probabilités pour toutes les valeurs possibles de variables aléatoires

	Inconnu = vrai		*Inconnu = faux*	
	Sensible = vrai	*Sensible = faux*	*Sensible = vrai*	*Sensible = faux*
Pourriel = vrai	0.108	0.012	0.072	0.008
Pourriel = faux	0.016	0.064	0.144	0.576

Toutes ces probabilités somment à 1

- Exemples :
  - **P**(*Pourriel, Inconnu, Sensible*)
  - **P**(*Pourriel*) = [*P*(*Pourriel=faux*), *P*(*Pourriel=vrai*)] = [0.8, 0.2]
  - P(Pourriel, Inconnu)

    = [[Pourriel= faux, Inconnu=faux), [Pourriel= vrai, Inconnu=faux)],

    [Pourriel= faux, Inconnu=vrai), [Pourriel= vrai, Inconnu=vrai)]]
- La somme est toujours égale à 1
- J’utilise le symbole **P** pour les distributions et *P* pour les probabilités
  - *P*(*Pourriel*) désignera la probabilité *P*(*Pourriel*=*x*) pour une valeur *x* non-spécifiée

Probabilité conjointe

- **Probabilité conjointe** : probabilité d'une assignation de valeurs à toutes la variables
  - *P*(*Inconnu=vrai*, *Sensible=vrai*, *Pourriel=vrai*) = 0.108 (10.8%)
  - *P*(*Inconnu=faux*, *Sensible=faux*, *Pourriel=vrai*) = 0.008 (0.8%)

	Inconnu = vrai		*Inconnu = faux*	
	Sensible = vrai	*Sensible = faux*	*Sensible = vrai*	*Sensible = faux*
Pourriel = vrai	0.108	0.012	0.072	0.008
Pourriel = faux	0.016	0.064	0.144	0.576

Probabilité marginale

- **Probabilité marginale** : probabilité sur un sous-ensemble des variables
  - ***$P(Y) = \Sigma_z P(Y, Z=z)$* -** Pour n'importe quelle ensemble de variable **Y** et **Z**
  - *P*(*Inconnu=vrai*, *Pourriel=vrai*)
    = *P*(*Inconnu=vrai*, *Sensible=vrai*, *Pourriel=vrai*) + *P*(*Inconnu=vrai*, *Sensible=faux*, *Pourriel=vrai*)
    = $\Sigma_{z \in \{vrai,\ faux\}}$ *P*(*Inconnu=vrai*, *Sensible=z*, *Pourriel=vrai*) = 0.108 + 0.012 = **0.12**

	Inconnu = vrai		*Inconnu = faux*	
	Sensible = vrai	*Sensible = faux*	*Sensible = vrai*	*Sensible = faux*
Pourriel = vrai	0.108	0.012	0.072	0.008
Pourriel = faux	0.016	0.064	0.144	0.576

Probabilité marginale

- **Probabilité marginale** : probabilité sur un sous-ensemble des variables
  - $P(Pourriel{=}vrai)$
    $= \Sigma_{x \in \{vrai,\, faux\}} \Sigma_{y \in \{vrai,\, faux\}}\ P(\mathrm{Pourriel{=}vrai}, Inconnu{=}x,\ Sensible{=}y,\ Pourriel{=}vrai)$
    $= 0.108 + 0.012 + 0.072 + 0.008 = \mathbf{0.2}$

	Inconnu = vrai		*Inconnu = faux*	
	Sensible = vrai	*Sensible = faux*	*Sensible = vrai*	*Sensible = faux*
Pourriel = vrai	0.108	0.012	0.072	0.008
Pourriel = faux	0.016	0.064	0.144	0.576

Probabilité d'un événement arbitraire

- **Probabilité de disjonction (« ou ») d'événements :**
  - *P*(*Pourriel=vrai* ou *Inconnu=faux*) – Six états (mondes) possibles

    = 0.108 + 0.012 +0.072 + 0.008 + 0.144 + 0.576

    = ***0.92***

  - *P*(*Pourriel=vrai* ou *Inconnu=faux*) – Une autre façon de le calculer

    = *P*(*Pourriel=vrai*) + *P*(*Inconnu=faux*) - *P(Pourriel=vrai, Inconnu=faux)*

    = 1 - *P*(*Pourriel=**faux**, Inconnu=**vrai***) = 1 – 0.016 – 0.064 = **0.92**

	Inconnu = vrai		*Inconnu = faux*	
	Sensible = vrai	*Sensible = faux*	*Sensible = vrai*	*Sensible = faux*
Pourriel = vrai	0.108	0.012	0.072	0.008
Pourriel = faux	0.016	0.064	0.144	0.576

Probabilité d'un événement arbitraire

- On peut calculer la probabilité d'événements arbitrairement complexes
  - il suffit d'additionner les probabilités des événements élémentaires associés
  - *P*((*Pourriel=vrai, Inconnu=faux*) ou (*Sensible=faux, Pourriel=faux*))
    = 0.072 + 0.008 + 0.064 + 0.576 = **0.72**

	Inconnu = vrai		*Inconnu = faux*	
	Sensible = vrai	*Sensible = faux*	*Sensible = vrai*	*Sensible = faux*
Pourriel = vrai	0.108	0.012	0.072	0.008
Pourriel = faux	0.016	0.064	0.144	0.576

Probabilité conditionnelle

- **Probabilité conditionnelle :**

  - ***P(X|Y) = P(X,Y) / P(Y)*** *si P(Y)* ***≠ 0***

  - *P(Pourriel=faux | Inconnu=vrai)*
    = *P(Pourriel=faux, Inconnu=vrai) / P(Inconnu=vrai)*
    = (0.016 + 0.064) / (0.016 + 0.064 + 0.108 + 0.012) = **0.4**

	Inconnu = vrai		*Inconnu = faux*	
	Sensible = vrai	*Sensible = faux*	*Sensible = vrai*	*Sensible = faux*
Pourriel = vrai	0.108	0.012	0.072	0.008
Pourriel = faux	0.016	0.064	0.144	0.576

Distribution conditionnelle

- On a vu que $P(Y) = \Sigma_z \, P(Y, Z{=}z)$
- On peut en déduire: $\boldsymbol{P(Y) = \Sigma_z \, P(Y|Z)P(Z{=}z)}$

Distribution conditionnelle

	Inconnu = vrai		*Inconnu = faux*	
	Sensible = vrai	*Sensible = faux*	*Sensible = vrai*	*Sensible = faux*
Pourriel = vrai	0.108	0.012	0.072	0.008
Pourriel = faux	0.016	0.064	0.144	0.576

- Exemple :
  - **P**(*Pourriel* | *Inconnu=faux*)
    = [*P*(*Pourriel=faux* | *Inconnu=faux*), *P*(*Pourriel=vrai* | *Inconnu=faux*)]
    = [0.9, 0.1]
  - **P**(*Pourriel* | *Inconnu*)
    = [[*P*(*Pourriel=faux* | *Inconnu=faux*), *P*(*Pourriel=vrai* | *Inconnu=faux*)],
    [*P*(*Pourriel=faux* | *Inconnu=vrai*), *P*(*Pourriel=vrai* | *Inconnu=vrai*)]]
    = [[0.9, 0.1], → somme à 1
    [0.4, 0.6]] → somme à 1
- **Chaque sous-ensemble de probabilités** associé aux mêmes valeurs des variables sur lesquelles on conditionne somme à 1
- **P**(*Pourriel* | *Inconnu*) contient deux distributions de probabilités sur la variable *Pourriel* : une dans le cas où *Inconnu=faux,* l'autre lorsque *Inconnu=vrai*

Distribution conditionnelle

- Une distribution conditionnelle peut être vue comme une distribution **renormalisée** afin de satisfaire les conditions de sommation à 1
    - $\boldsymbol{P(X|e) = \alpha\, \Sigma_y\, P(X,e,y)}$

Distribution conditionnelle

- Une distribution conditionnelle peut être vue comme une distribution **renormalisée** afin de satisfaire les conditions de sommation à 1
- Exemple :
  - **P**(*Pourriel* | *Inconnu=vrai*)
    = α **P**(*Pourriel, Inconnu=vrai*)
    = α [0.08, 0.12]
    = (1/ (0.08 + 0.12)) [0.08, 0.12]
    = [0.4, 0.6]
  - **P**(*Pourriel* | *Inconnu*)
    = [α_{faux} **P**(*Pourriel, Inconnu=faux*) ,
    α_{vrai} **P**(*Pourriel, Inconnu=vrai*)]
    = [[0.72, 0.08] / (0.72 + 0.08),
    [0.08, 0.12] / (0.08 + 0.12)]
    = [[0.9, 0.1],
    [0.4, 0.6]]

	Inconnu = vrai		*Inconnu = faux*	
	Sensible = vrai	*Sensible = faux*	*Sensible = vrai*	*Sensible = faux*
Pourriel = vrai	0.108	0.012	0.072	0.008
Pourriel = faux	0.016	0.064	0.144	0.576

Règle du produit

- **Règle du produit :**
  - ***P(X,Y)=P(X|Y)P(Y)***
  - *P*(*Pourriel=faux*, *Inconnu=vrai*)
= *P*(*Pourriel=faux* | *Inconnu=vrai*) *P*(*Inconnu=vrai*)
= *P*(*Inconnu=vrai* | *Pourriel=faux*) *P*(*Pourriel=faux*)
  - En général :
P(*Pourriel*, *Inconnu*) = *P*(*Pourriel* | *Inconnu*) *P*(*Inconnu*)
= *P*(*Inconnu* | *Pourriel*) *P*(*Pourriel*)

Règle de chaînage

- **Règle du produit :**
  - ***$P(X,Y)=P(X|Y)P(Y)$***

- **Règle de chaînage** (***chain rule***) pour n variables $X_1 \dots X_n$:
  - $P(X_1, \dots, X_n) = P(X_1,\dots,X_{n-1})\, P(X_n \mid X_1,\dots,X_{n-1})$
    $= P(X_1,\dots,X_{n-2})\, P(X_{n-1} \mid X_1,\dots,X_{n-2})\, P(X_n \mid X_1,\dots,X_{n-1})$
    $= \dots$
    $= \prod_{i=1..n} P(X_i \mid X_1, \dots ,X_{i-1})$

Règle de chaînage

- La règle du chaînage est vraie, quelle que soit la distribution de $X_1 \ldots X_n$
- Plutôt que de spécifier toutes les probabilités jointes $P(X_1, \ldots, X_n)$, on pourrait plutôt spécifier $P(X_1)$, $P(X_2|X_1)$, $P(X_3|X_1, X_2)$, ..., $P(X_n \mid X_1,\ldots,X_{n-1})$
- Exemple, on aurait pu spécifier :
  - $P(Pourriel{=}faux) = 0.8$, $P(Pourriel{=}vrai) = 0.2$
  - $P(Inconnu{=}faux|\ Pourriel{=}faux) = 0.9$, $P(Inconnu{=}vrai|\ Pourriel{=}faux) = 0.1$
    $P(Inconnu{=}faux|\ Pourriel{=}vrai) = 0.4$, $P(Inconnu{=}vrai \mid Pourriel{=}vrai) = 0.6$
- On aurait tous les ingrédients pour calculer les $P(Pourriel, Inconnu)$:
  - $P(Pourriel{=}faux, Inconnu{=}vrai) = P(Inconnu{=}vrai \mid Pourriel{=}faux)\ P(Pourriel{=}faux)$
    $= 0.1 * 0.8 = 0.08$
  - $P(Pourriel{=}vrai, Inconnu{=}vrai) = P(Inconnu{=}vrai|Pourriel{=}vrai)\ P(Pourriel{=}vrai)$
    $= 0.6 * 0.2 = 0.12$

Règle de Bayes

- **$P(X|Y) = P(Y|X)P(X)/P(Y)$**

- Donne une probabilité **diagnostique** à partir d'une probabilité **causale** :
  - $P(Cause|Effect) = P(Effect|Cause)\ P(Cause)\ /\ P(Effect)$

Règle de Bayes

- **$P(X|Y) = P(Y|X)P(X)/P(Y)$**
- Donne une probabilité **diagnostique** à partir d'une probabilité **causale** :
  - $P(\mathit{Cause}|\mathit{Effet}) = P(\mathit{Effet}|\mathit{Cause})\, P(\mathit{Cause}) / P(\mathit{Effet})$
- On pourrait calculer $P(\mathit{Pourriel}{=}\mathit{faux} \mid \mathit{Inconnu}{=}\mathit{vrai})$:
  - $P(\neg\text{pourriel} \mid \mathit{inconnu})$
    $= P(\neg\text{pourriel}, \mathit{inconnu}) / P(\mathit{inconnu})$
    $= P(\neg\text{pourriel}, \mathit{inconnu}) / (\, P(\mathit{inconnu}, \neg\mathit{pourriel}) + P(\mathit{inconnu}, \mathit{pourriel}))$
    $= \alpha\, P(\mathit{inconnu}|\neg\text{pourriel})\, P(\neg\text{pourriel})$
    $= 0.08 / (0.08 + 0.12) = \mathbf{0.4}$

Pourriel=faux ⇔	$\neg$pourriel
Pourriel=vrai ⇔	pourriel

- On appelle $P(\mathit{Pourriel})$ une **probabilité a priori**
  - c'est notre croyance p/r à la présence d'une Pourriel **avant** toute observation
- On appelle $P(\mathit{Pourriel}|\, \mathit{Inconnu})$ une probabilité a **posteriori**
  - c'est notre croyance mise à jour après avoir observé *Inconnu*
- **La règle de Bayes** lie ces deux probabilités ensemble
  - $\underline{P(\neg\text{pourriel} \mid \mathit{inconnu})} = \alpha\, P(\mathit{inconnu}|\neg\text{pourriel})\, \underline{P(\neg\text{pourriel})}$

Indépendance

- Soit les variables A et B, elles sont **indépendantes** si et seulement si
  - $P(A|B) = P(A)$ ou
  - $P(B|A) = P(B)$ ou
  - $P(A, B) = P(A)\ P(B)$

Indépendance

- Soit les variables A et B, elles sont **indépendantes** si et seulement si
  - $P(A|B) = P(A)$ ou
  - $P(B|A) = P(B)$ ou
  - $P(A, B) = P(A)\ P(B)$

- Exemple : $P(Pluie, Pourriel) = P(Pluie)\ P(Pourriel)$

***P*(*Pluie = vrai*) = 0.3**

***P*(*Pourriel = vrai*) = 0.1**

Pluie	***Pourriel***	**Probabilité**	
vrai	*vrai*	0.03	$= P(pluie)\ P(Pourriel) = 0.3 * 0.1$
vrai	*faux*	0.27	$= P(pluie)\ P(\neg Pourriel) = 0.3 * 0.9$
faux	*vrai*	0.07	$= P(\neg pluie)\ P(Pourriel) = 0.7 * 0.1$
faux	*faux*	0.63	$= P(\neg pluie)\ P(\neg Pourriel) = 0.7 * 0.9$

Indépendance

- L'indépendance totale est puissante mais rare

- L'indépendance entre les variables permet de réduire la taille de la distribution de probabilités et rendre les inférences plus efficaces
  - dans l'exemple précédent, on n'a qu'à stocker en mémoire *P*(*Pluie = vrai*) = 0.3 et *P*(*Pourriel = vrai*) = 0.1, plutôt que la table au complet

- Mais il est rare d'être dans une situation où toutes les variables sont réellement indépendantes

Pluie	***Pourriel***	**Probabilité**
vrai	*vrai*	0.03
vrai	*faux*	0.27
faux	*vrai*	0.07
faux	*faux*	0.63

Indépendance conditionnelle

- Si je sais déjà que le courriel est un pourriel, ma croyance (probabilité) qu'il contienne un mot sensible ne dépend plus du fait que l'expéditeur me soit inconnu ou non :
  - $P(Sensible \mid Inconnu, Pourriel{=}vrai) = P(Sensible \mid Pourriel{=}vrai)$

- On dit que *Sensible* est **conditionnellement indépendante** de *Inconnu* étant donné *Pourriel*, puisque :
  - $P(Sensible \mid Inconnu, Pourriel) = P(Sensible \mid Pourriel)$

- Formulations équivalentes :
  - $P(Inconnu \mid Sensible, Pourriel) = P(Inconnu \mid Pourriel)$
  - $P(Inconnu, Sensible \mid Pourriel) = P(Inconnu \mid Pourriel)\, P(Sensible \mid Pourriel)$

Indépendance conditionnelle

- Réécrivons la distribution conjointe en utilisant la **règle de chaînage** (***chain rule***) :

$$
\begin{aligned}
&P(\mathit{Inconnu}, \mathit{Sensible}, \mathit{Pourriel}) \\
&\quad = P(\mathit{Inconnu} \mid \mathit{Sensible}, \mathit{Pourriel})\, P(\mathit{Sensible}, \mathit{Pourriel}) \\
&\quad = P(\mathit{Inconnu} \mid \mathit{Sensible}, \mathit{Pourriel})\, P(\mathit{Sensible} \mid \mathit{Pourriel})\, P(\mathit{Pourriel}) \\
&\quad = P(\mathit{Inconnu} \mid \mathit{Pourriel})\, P(\mathit{Sensible} \mid \mathit{Pourriel})\, P(\mathit{Pourriel})
\end{aligned}
$$

- C-à-d., 2 + 2 + 1 = 5 **paramètres individuels/distincts**

- Dans des cas idéals, l'exploitation de l'indépendance conditionnelle réduit la complexité de représentation de la distribution conjointe de exponentielle ($O(2^n)$) en linéaire ($O(n)$)

En bref

- **Probabilité jointe :** $P(X_1, \ldots, X_n)$
- **Probabilité marginale :** $P(X_i)$, $P(X_i, X_j)$, etc.
- **Probabilité conditionnelle :** $P(X_1, \ldots, X_k \mid X_{k+1}, \ldots, X_n) = \dfrac{P(X_1, \ldots, X_k, X_{k+1}, \ldots, X_n)}{P(X_{k+1}, \ldots, X_n)}$
- **Régle de chaînage :** $P(X_1, \ldots, X_n) = \prod_{i=1..n} P(X_i \mid X_1, \ldots, X_{i-1})$
- **Indépendance :** X_i et X_j sont indépendantes si
  $P(X_i, X_j) = P(X_i)\, P(X_j)$, ou $P(X_i \mid X_j) = P(X_i)$ ou $P(X_j \mid X_i) = P(X_j)$
- **Indépendance conditionnelle :** X_i et X_j sont indépendante sachant X_k si
  $P(X_i, X_j \mid X_k) = P(X_i \mid X_k)\, P(X_j \mid X_k)$ ou $P(X_i \mid X_j, X_k) = P(X_i \mid X_k)$ ou $P(X_j \mid X_i, X_k) = P(X_j \mid X_k)$
- **Règle de Bayes :** $P(X_1, \ldots, X_k \mid X_{k+1}, \ldots, X_n) = \dfrac{P(X_{k+1}, \ldots, X_n \mid X_1, \ldots, X_k)\, P(X_1, \ldots, X_k)}{P(X_{k+1}, \ldots, X_n)}$

Autres types de variables aléatoires

- On s'est concentré sur des variables aléatoires Booléennes ou binaires
  - le **domaine**, c.-à-d. l'ensemble des valeurs possibles de la variable, était toujours {*vrai, faux*}

- On pourrait avoir d'autres types de variables, avec des domaines différents :
  - **Discrètes** : le domaine est énumérable
    - » *Météo* $\in$ {*soleil*, *pluie*, *nuageux*, *neige*}
    - » lorsqu'on marginalise, on doit sommer sur toutes les valeurs :
      $P(\mathit{Température}) = \Sigma_{x \in \{soleil,\ pluie,\ nuageux,\ neige\}} P(\mathit{Température}, \mathit{Météo}{=}x)$
  - **Continues** : le domaine est continu (par exemple, l'ensemble des réels)
    - » exemple : *PositionX* = 4.2
    - » le calcul des probabilités marginales nécessite des intégrales

Classifieur bayésien naïf

- Le classifieur (modèle) bayésien naïf est définit comme suit

$$\boldsymbol{P}(\text{Cause}, \textit{Effet}_1, \ldots, \textit{Effet}_n) = \boldsymbol{P}(\text{Cause}) = \Pi_{\text{i=1..n}}\, \boldsymbol{P}(\textit{Effet}_i \mid \textit{Cause})$$

- Naïf parce qu'on suppose l'Independence conditionnel. Mais fonctionne dans beaucoup d'applications

Classifieur bayésien naïf

- Le modèle (Classifieur) bayésien naïf est définit comme suit

$$\boldsymbol{P}(\text{Cause}, \mathit{Effet}_1, \ldots, \mathit{Effet}_n) = \boldsymbol{P}(\text{Cause}) \prod_{i=1..n} \boldsymbol{P}(\mathit{Effet}_i \mid \mathit{Cause})$$

- Pour l'appliquer, en général, on observe des effets (e) et on veut diagnostiquer la cause.

- Noton **E=e** les effets observés. On a vu que $\boldsymbol{P(Cause|e)} = \alpha\, \Sigma_y\, \boldsymbol{P}(Cause,e,y)$
- On a donc:

$$\begin{aligned}
\boldsymbol{P}(\text{Cause} \mid \text{e}) &= \alpha\, \Sigma_y\, \boldsymbol{P}(Cause) P(y \mid Cause) \prod_{j=1..n} \boldsymbol{P}(e_i \mid Cause) \\
&= \alpha\, \boldsymbol{P}(Cause) \prod_{j=1..n} \boldsymbol{P}(e_i \mid Cause)\, \Sigma_y\, P(y \mid Cause) \\
&= \boldsymbol{\alpha\, P(Cause) \prod_{j=1..n} P(e_i \mid Cause)}
\end{aligned}$$

Classification de documents

- Classifieur bayésien naïf

$$P(Cause|e) = \alpha\, P(Cause)\, \Pi_{j=1..n}\, P(e_i\, |Cause)$$

- Classification de documents: étant donné un document texte, déterminer dans laquelle des catégories prédéfinie il appartient (ex: sport, économie)

- Exemples de textes (documents):
  - Apple a fait état jeudi d'un chiffre d'affaires et d'un bénéfice net supérieur aux attentes pour le trimestre allant d'octobre à décembre l'année dernière, la forte hausse des ventes d'iPhone, notamment en Chine, ayant plus que compensé les difficultés des chaînes d'approvisionnement … (*Tiré de Radio Canada / Économie – 2022-01-08*)
  - Le Canada s'est rapproché davantage de son objectif en défaisant le Honduras 2-0, jeudi, en match de qualification pour la Coupe du monde au Qatar (*Tiré de Radio Canada / Sport – 2022-01-08*)

Classification de documents

- Classifieur bayésien naïf

$$P(Cause|e) = \alpha\, P(Cause)\, \Pi_{j=1..n}\, P(e_i\,|Cause)$$

- Étant donné un document, déterminer dans laquelle des catégories prédéfinie il appartient (e.g., sport, politique, économie, etc.)

  - **Cause** correspond à la catégorie (classe) des documents (sport, politique, etc.)
  - $\mathbf{e_j}$ correspond à la présence ou absence de certains mots clés, *keyWord$_{i.}$.*
  - *Donc* $\mathbf{e} \equiv \{keyWord_1, .., keyWord_n\}$*, c.-à-.d, les mots-clés observés*

$$P(Class|keyWord_1, .., keyWord_n) = \alpha\, P(Class)\, \Pi_{j=1..n}\, P(keyWorld_i\,|Class)$$

- Le modèle bayésien naïf consiste des probabilités à priori *P(Class)* et des probabilités conditionnelles $P(KeyWorld_i\,|Class)$.

Classification de documents

$$\boldsymbol{P(Category|\mathrm{ObservedKeyWords}) = \alpha\, P(Catgeory)\, \Pi_{j=1..n}\, P(HasWorld_i\, |Category)}$$

- Le modèle bayésien naïf consiste des probabilités à priori *P(Category)* et des probabilités conditionnelles $P(HasWorld_i\,|Category)$.

- Pour classifier un document

  - On vérifie quels mots clés apparaissent dans le document, ce qui donne ObservedKeywords
  - On applique ensuite l'équation pour obtenir la distribution des probabilités à postériori des catégories, c.-a.-d., ***P****(Category|ObservedKeyWords)*
  - On choisit finalement $argmax_c\, P(Category = c\ |ObservedKeyWords)$, c.-à-d., la catégorie avec la probabilité à postériori la plus elevée.

Classification de documents

$$P(Category|ObservedKeywords) = \alpha\ P(Catgeory)\ \Pi_{j=1..n}\ P(HasWorld_i\ |Category)$$

- Le modèle bayésien naïf consiste des probabilités à priori *P(Category)* et des probabilités conditionnelles $P(HasWorld_i\ |Category)$.

- Pour apprendre le modèle:

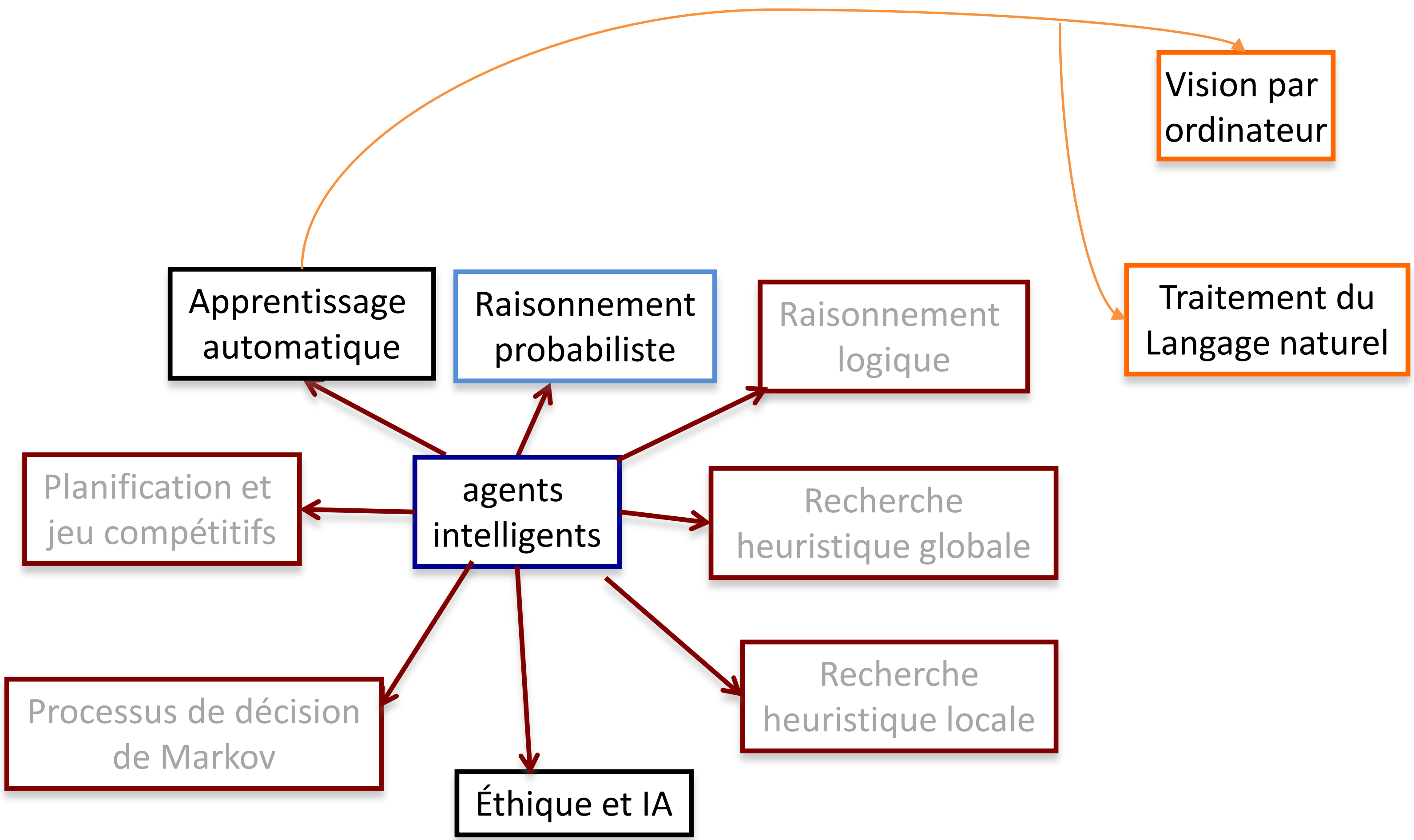
  - *P(Category = c)* est la fraction des documents de cette catégorie vue jusqu'à présent.

  - $P(HasWord_i \mid Category = c)$ est la fraction de documents de catégorie *c* de la catégorie qui contient le mot $Word_i$.

Sujets couverts par le cours

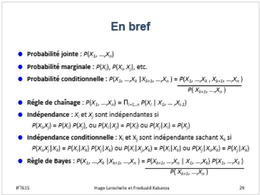
Concepts et algorithmes

Applications



Vous devriez être capable de...

- À partir d'une distribution conjointe ou des distributions conditionnelles et a priori nécessaires :
  - calculer une probabilité conjointe
  - calculer une probabilité marginale
  - déterminer si deux variables sont indépendantes
  - déterminer si deux variables sont conditionnellement indépendantes sachant une troisième
  - Appliquer la règle du chainage
  - Appliquer la règle de Bayes



En bref

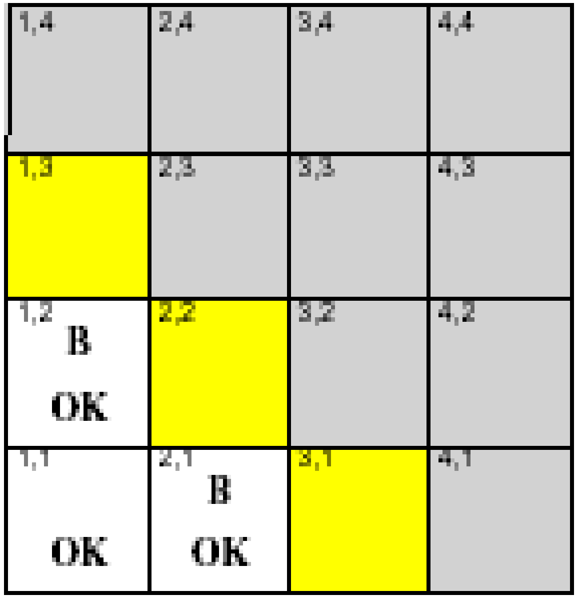
- **Probabilité jointe :** $P(X_1, \ldots, X_n)$
- **Probabilité marginale :** $P(X_i)$, $P(X_i, X_j)$, etc.
- **Probabilité conditionnelle :** $P(X_1, \ldots, X_k \mid X_{k+1}, \ldots, X_n) = \dfrac{P(X_1, \ldots, X_k, X_{k+1}, \ldots, X_n)}{P(X_{k+1}, \ldots, X_n)}$
- **Régle de chaînage :** $P(X_1, \ldots, X_n) = \prod_{i=1..n} P(X_i \mid X_1, \ldots, X_{i-1})$
- **Indépendance :** X_i et X_j sont indépendantes si
  $P(X_i, X_j) = P(X_i)\, P(X_j)$, ou $P(X_i \mid X_j) = P(X_i)$ ou $P(X_j \mid X_i) = P(X_j)$
- **Indépendance conditionnelle :** X_i et X_j sont indépendante sachant X_k si
  $P(X_i, X_j \mid X_k) = P(X_i \mid X_k)\, P(X_j \mid X_k)$ ou $P(X_i \mid X_j, X_k) = P(X_i \mid X_k)$ ou $P(X_j \mid X_i, X_k) = P(X_j \mid X_k)$
- **Règle de Bayes :** $P(X_1, \ldots, X_k \mid X_{k+1}, \ldots, X_n) = \dfrac{P(X_{k+1}, \ldots, X_n \mid X_1, \ldots, X_k)\, P(X_1, \ldots, X_k)}{P(X_{k+1}, \ldots, X_n)}$

IFT615 Hugo Larochelle et Froduald Kabanza 29

Le monde des Wumpus

Problème: calculer la probabilité que [1,3], [2,2] et [3,1] contienne une fosse

1. Identifier l'ensemble de variables aléatoires nécessaires:
   - ***P_{ij}****=true* ssi il y a une fosse dans *[i,j]* (*Pij=0.2* partout sauf dans *[1,1]*).
   - ***B_{ij}****=true* ssi il y a une brise dans *[i,j]*



Inclure seulement les variables observées B_{11}, B_{12}, B_{21} dans la distribution des probabilités (modèle).

Spécifier la distribution des probabilités

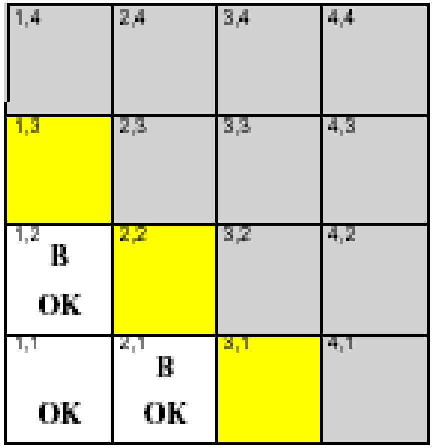
2. **Spécifier la distribution conjointe** ($\mathbf{P}(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1})$)

- appliquer la règle du produit : $\mathbf{P}(B_{1,1}, B_{1,2}, B_{2,1} | P_{1,1}, \ldots, P_{4,4})\, \mathbf{P}(P_{1,1}, \ldots, P_{4,4})$ (on spécifie une forme $\mathbf{P}(Effect|Cause)$)
- premier terme : $\mathbf{P}(B_{1,1}, B_{1,2}, B_{2,1} | P_{1,1}, \ldots, P_{4,4})$
  - » probabilité conditionnelle d'une configuration/état de brises, étant donnée une configuration de fosses
  - » 1 si les fosses sont adjacentes aux brises, 0 sinon
- second terme : $\mathbf{P}(P_{1,1}, \ldots, P_{4,4})$
  - » probabilité a priori des configurations des fosses
  - » les fosses sont placées aléatoirement, avec une probabilité de 0.2 par chambre
  - » si $P_{1,1}, \ldots, P_{4,4}$ sont telles qu'il y a exactement n fosses, on aura $\mathbf{P}(P_{1,1}, \ldots, P_{4,4}) = \Pi_{(i,j)=(1,1) \ldots (4,4)} \mathbf{P}(P_{i,j}) = 0.2^n * 0.8^{16-n}$

Observations et requête

3. Identifier les observations

- on sait ce qui suit :
  - » $b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$
  - » $known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$



4. Identifier les variables de requête

- y a-t-il une fosse à la position 1,3?
- $\mathbf{P}(P_{1,3} \mid known, b)$?

3. Identifier les variables cachées

- on définit *Unknown* comme étant l'ensemble des variables $P_{i,j}$ autres que celles qui sont connues (*known*) et la variable de requête $P_{1,3}$

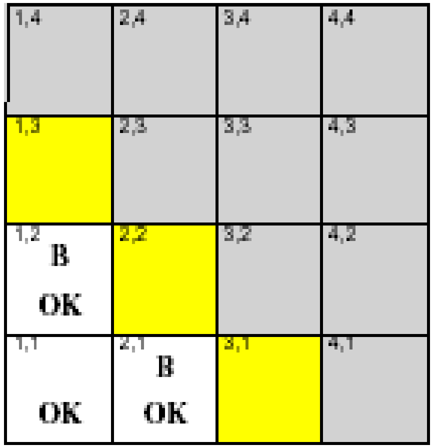
Observations et requête

6. Faire l'inférence

- avec l'inférence par énumération, on obtient :

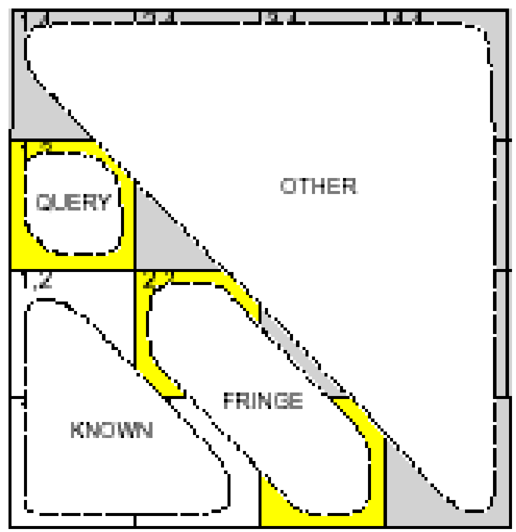
  $\mathbf{P}(P_{1,3} | \ known, b) =$

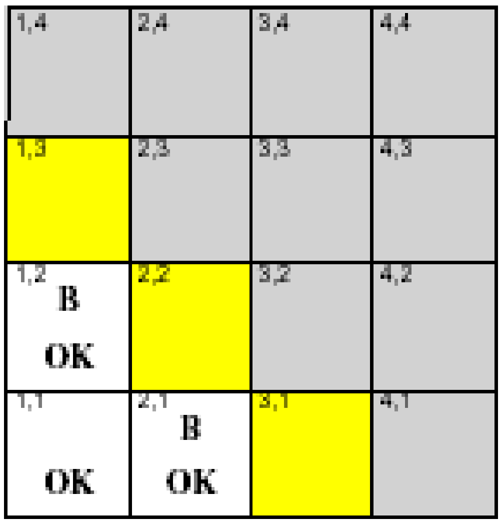
  $\alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b)$
- croît exponentiellement avec le nombre de chambres!
  - » avec 12 chambres *unknown* : 2^{12}=4096 termes



Utiliser l’indépendance conditionnelle

- Idée de base: les observations sont conditionnellement indépendantes des chambres cachées étant données les chambres adjacentes.
  - C.-à-d., les autres chambres ne sont pas pertinentes.





- Définir *Unknown = Fringe* $\cup$ *Other*

- $\mathbf{P}(b \mid P_{1,3}, known, Unknown) = \mathbf{P}(b \mid P_{1,3}, known, Fringe, Other)$

- Réécrire la probabilité d’interrogation $P(P_{1,3} \mid known, b)$ pour exploiter cette indépendance.

Utiliser l'indépendance conditionnelle

$\mathbf{P}(P_{1,3} | \, known, b) = \alpha \, \Sigma_{unknown} \, \mathbf{P}(P_{1,3}, \; unknown, known, b)$

$= \alpha \, \Sigma_{unknown} \, \mathbf{P}(b | P_{1,3}, known, unknown) \, \mathbf{P}(P_{1,3}, known, unknown)$

$= \alpha \, \Sigma_{frontier} \, \Sigma_{other} \, \mathbf{P}(b | \, known, P_{1,3}, frontier, other) \, \mathbf{P}(P_{1,3}, known, frontier, other)$

$= \alpha \, \Sigma_{frontier} \, \Sigma_{other} \, \mathbf{P}(b | \, known, P_{1,3}, frontier) \, \mathbf{P}(P_{1,3}, known, frontier, other)$

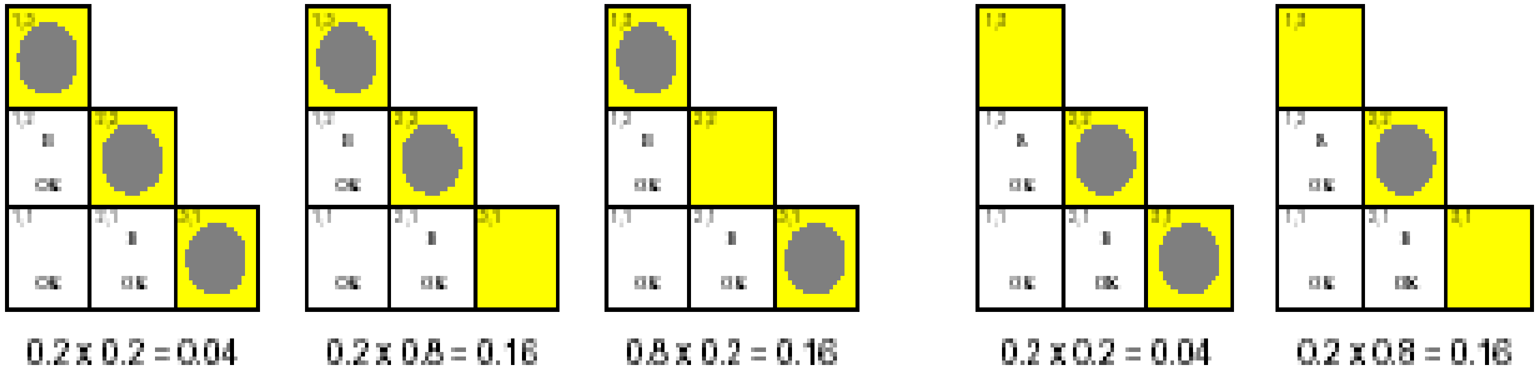
$= \alpha \, \Sigma_{frontier} \, \mathbf{P}(b | \, known, P_{1,3}, frontier) \, \Sigma_{other} \, \mathbf{P}(P_{1,3}, known, frontier, other)$

$= \alpha \, \Sigma_{frontier} \, \mathbf{P}(b | \, known, P_{1,3}, frontier) \, \Sigma_{other} \, \mathbf{P}(P_{1,3}) \, P(known) \, P(frontier) \, P(other)$

$= \alpha \, P(known) \, \mathbf{P}(P_{1,3}) \, \Sigma_{frontier} \, \mathbf{P}(b | \, known, P_{1,3}, frontier) \, P(frontier) \, \Sigma_{other} \, P(other)$

$= \alpha' \, \mathbf{P}(P_{1,3}) \, \Sigma_{frontier} \, \mathbf{P}(b | \, known, P_{1,3}, frontier) \, P(frontier)$

Utiliser l'indépendance conditionnelle



(a) (b)

- Événements cohérents pour les variables $P_{2,2}$ et $P_{3,1}$, montrant **P**(*frontier*)
- Pour chaque événement :
  a) 3 événements avec $P_{1,3}$= *vrai*, montrant 2 ou 3 fosses.
  b) 2 événements avec $P_{1,3}$= *faux*, montrant 1 ou 2 fosses.

$$\mathbf{P}(P_{1,3} | known, b) = \alpha' \langle 0.2(0.04+0.16+0.16), 0.8(0.04+0.16) \rangle$$

$$\approx \langle 0.31, 0.69 \rangle$$

Classification de documents

- Soit les deux documents (question d'examen) suivants:

« Dessinez la partie de l'espace d'états qui serait explorée par l'algorithme alpha-beta pruning, en supposant qu'il explore l'espace d'états de la gauche vers la droite. »

« En utilisant l'algorithme d'apprentissage du perceptron et un pas d'apprentissage de 0.3, donnez la sortie et les poids des connexions à la fin de la deuxième itération. »

- Laquelle est une question d'examen final, en IFT 615?

Classification de documents

- Soit les deux documents (question d'examen) suivants:

« d'états d'états de qui explore qu'il explorée gauche l'algorithme pruning, l'espace par en Dessinez alpha-beta droite. la la supposant l'espace partie serait la de vers »	« un pas de l'algorithme fin sortie de perceptron donnez la deuxième En à poids du et et des d'apprentissage connexions les itération. la la d'apprentissage utilisant 0.3, »

- Laquelle est une question d'examen final, en IFT 615?

Classification de documents

- Les mots individuels sont très informatifs du sujet (catégorie) d'une document

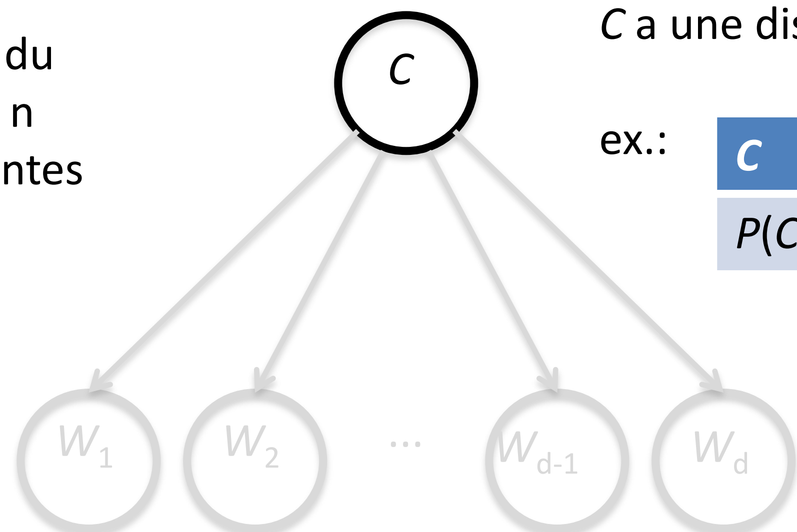
- L'ordre des mots n'est souvent pas utile
  - l'ordre reflète surtout la syntaxe d'une langue
  - on suppose que la catégorie n'influence que la probabilité d'observer un mot dans un document

- Ignorer l'ordre des mots va permettre de simplifier le système, sans trop compromettre sa précision

- On va formaliser ces hypothèses à l'aide d'un **modèle bayésien**

Modèle bayésien naïf multinomial

C est la catégorie du document, parmi n catégories différentes



C a une distribution a priori

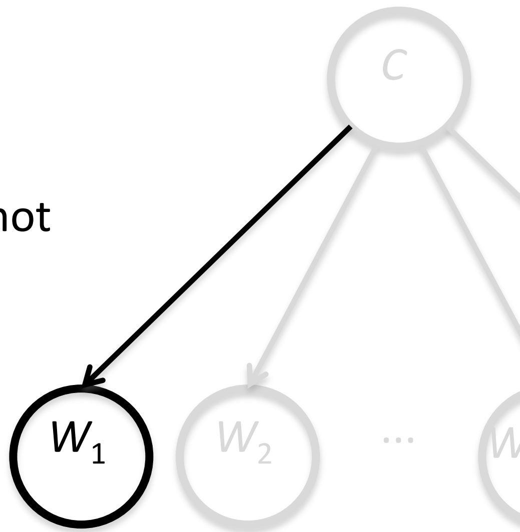
ex.:

C	*intra*	*final*
P(*C*)	0.5	0.5

somme à 1

Modèle bayésien naïf multinomial

W_1 est le premier mot d'un document, contenant d mots



W_1 a une distribution **conditionnelle multinomiale**

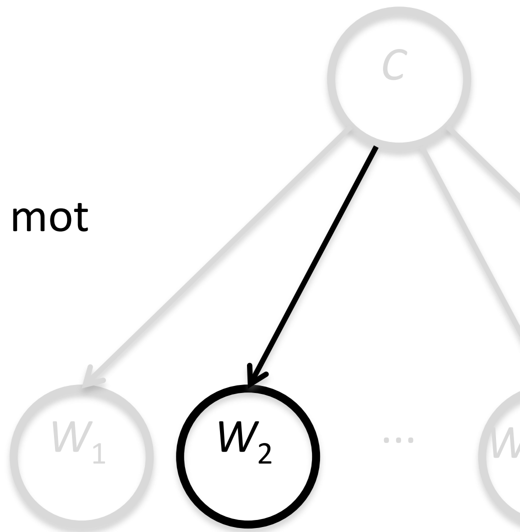
C	*intra*	*final*
$P(W_1$=« de »$\|C)$	0.01	0.01
$P(W_1$=« qui »$\|C)$	0.02	0.02
...	...	...
$P(W_1$=« perceptron »$\|C)$	10^{-6}	0.002

somme à 1

somme à 1

Modèle bayésien naïf multinomial

W_2 est le deuxième mot d'un document, contenant d mots



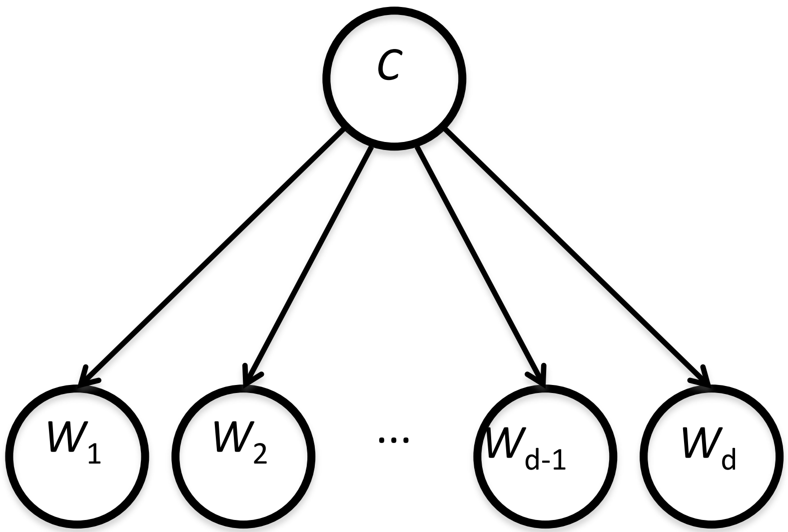
W_2 a **la même** une distribution conditionnelle multinomiale

C	*intra*	*final*
$P(W_2$=« de »$\|C)$	0.01	0.01
$P(W_2$=« qui »$\|C)$	0.02	0.02
...	...	...
$P(W_2$=« perceptron »$\|C)$	10^{-6}	0.002

somme à 1

somme à 1

Modèle bayésien naïf multinomial



- En général la **probabilité conjointe** d'un document $[W_1,\ldots,W_d]$ ayant d mots et de sa catégorie C:

$$P([W_1,\ldots,W_d], C) = P(C) \prod_i P(W_i \mid C)$$

Modèle bayésien naïf multinomial

- Exemple:

C	*intra*	*final*
$P(C)$	0.5	0.5

C	*intra*	*final*
$P(W_i$=« , »$\|C)$	0.01	0.01
$P(W_i$=« un »$\|C)$	0.02	0.02
$P(W_i$=« d' »$\|C)$	0.01	0.02
$P(W_i$=« Perceptron »$\|C)$	10^{-6}	0.002
$P(W_i$=« algorithme »$\|C)$	0.005	0.005
$P(W_i$=« apprentissage »$\|C)$	10^{-5}	0.001
$P(W_i$=« . »$\|C)$	0.03	0.03
...	...	...

P(« Perceptron, un algorithme d'apprentissage. », C = *intra*) =

$0.5 * 10^{-6} * 0.01 * 0.02 * 0.005 * 0.01 * 10^{-5} * 0.03 = 1.5 * 10^{-21}$

P(« Perceptron, un algorithme d'apprentissage. », C = *final*) =

$0.5 * 0.002 * 0.01 * 0.02 * 0.005 * 0.02 * 0.001 * 0.03 = 6 * 10^{-16}$

Décision de la catégorie d'un document

- Pour classifier un document contenant les mots $[w_1,\ldots,w_d]$, on choisit la classe c ayant la plus grande **probabilité a posteriori** $P(C{=}c \mid [w_1,\ldots,w_d])$

$$
\begin{aligned}
&\operatorname*{argmax}_{c} P(C{=}c \mid [w_1,\ldots,w_d]) \\
&\quad = \operatorname{argmax} P(C{=}c,[w_1,\ldots,w_d]) \,/\, \alpha \\
&\quad = \operatorname{argmax} P(C{=}c,[w_1,\ldots,w_d]) \quad \longleftarrow \text{ pour simplifier les calculs} \\
&\quad = \operatorname{argmax} \log P(C{=}c,[w_1,\ldots,w_d]) \\
&\quad = \operatorname{argmax} \log P(C{=}c) \textstyle\prod_i P(W_i = w_i \mid C{=}c) \\
&\quad = \operatorname{argmax} \log P(C{=}c) + \textstyle\sum_i \log P(W_i = w_i \mid C{=}c) \quad \longleftarrow \text{ pour éviter le « underflow »}
\end{aligned}
$$

Décision de la catégorie d'un document

- Pour classifier un document fait des mots $[w_1,\ldots,w_d]$, on choisit la classe c ayant la plus grande **probabilité a posteriori** $P(C=c\,|\,[w_1,\ldots,w_d])$

- Exemple:

$$\text{argmax} \log P(C=c) + \textstyle\sum_i \log P(W_i = w_i \mid C=c)$$

C = intra

$$= \text{argmax}\{ \log(0.5) + \log(10^{-6}) + \log(0.01) + \log(0.02) + \log(0.005) + \log(0.01) + \log(10^{-5}) + \log(0.03),$$
$$\log(0.5) + \log(0.002) + \log(0.01) + \log(0.02) + \log(0.005) + \log(0.02) + \log(0.001) + \log(0.03) \}$$

C = final

= *final*

Apprentissage du modèle

- Comment obtient-on les distributions $P(C)$ et $P(W_i \mid C)$?
  - on les obtient à partir de vraies données
  - on choisit $P(C)$ et $P(W_i \mid C)$ pour quelles reflètent les statistiques de ces données

- Soit un **corpus**, c.-à-d. un ensemble de T documents $\{D_t, C_t\}$
  - chaque document D_t est une liste de mots $[w^t_1,\ldots,w^t_d]$ de taille variable
  - C_t est la catégorie de D_t

$$P(C=c) = (\text{nb. de documents de la catégorie } c) \,/\, (\text{nb. de documents total})$$
$$= |\{ t \mid C_t = c\}| \,/\, T$$

$$P(W_i = w \mid C=c) = \frac{\text{nb. de fois que } w \text{ apparaît dans les documents de la catégorie } c}{\text{nb. de mots total dans les documents de la catégorie } c}$$
$$= \frac{\sum_{t \mid Ct=c} \text{freq}(w, D_t)}{\sum_{t \mid Ct=c} |D_t|}$$

Lissage du modèle

- Selon la formule pour $P(W_i = w | C=c)$, un mot w aura une probabilité de 0 s'il n'apparaît jamais dans notre corpus

- Si un seul des $P(W_i = w | C=c) = 0$, alors tout $P(C=c,[w_1,\ldots,w_d]) = 0$!
  - les mots rares vont beaucoup faire varier $P(C=c,[w_1,\ldots,w_d])$ en général

- Pour éviter cette instabilité, deux trucs afin de **lisser la distribution *P*(*w*|*c*)**
  - on détermine un **vocabulaire** *V* de taille fixe, et on associe les mots qui ne sont pas dans ce vocabulaire au **symbole OOV** (*out of vocabulary*)
  - **lissage δ**: on ajoute une constante δ au numérateur, pour chaque mot

$$P(W_i = w | C=c) = \frac{\delta + \sum_{t \mid Ct=c} \mathrm{freq}(w, D_t)}{\delta\,(|V|+1) + \sum_{t \mid Ct=c} |D_t|}$$

Lissage du modèle

- Exemple: soit le vocabulaire

  V = { « Perceptron », « , », « un », « apprentissage »}

- La phrase

  « Perceptron, un algorithme d'apprentissage. »

  sera représentée par la liste de mots

  [« Perceptron », « , », « un », « OOV », « OOV », « apprentissage », « OOV »]

  w_1 w_2 w_3 w_4 w_5 w_6 w_7

- Les statistiques sont calculées à partir de cette représentation
  - on pourrait aussi enlever les mots « OOV » et les ignorer

Prétraitement des données

- Si, parmi tous les intra des années dernières (corpus de 426 mots)
  - « Perceptron » apparaît 0 fois
  - « , » apparaît 15 fois
  - « un » apparaît 10 fois
  - « apprentissage » apparaît 1 fois
  - « OOV » (tous les autres mots) apparaissent 400 fois

- Si on utilisait δ = 1, alors
  - *P*(« Perceptron » | *C=intra*) = (1 + 0) / (1 (4+1) + 426) = 1 / 431
  - *P*(« , » | *C=intra*) = (1 + 15) / (1 (4+1) + 426) = 16 / 431
  - *P*(« un » | *C=intra*) = (1 + 10) / (1 (4+1) + 426) = 11 / 431
  - *P*(« apprentissage » | *C=intra*) = (1 + 1) / (1 (4+1) + 426) = 2 / 431
  - *P*(« OOV » | *C=intra*) = (1 + 400) / (1 (4+1) + 426) = 401 / 431

somme à 1

Prétraitement des données

- Comment choisir *V*
  - ne garder que **les mots les plus fréquents** (ex.: apparaissent au moins 10 fois)
  - **ne pas garder les mots trop communs**
    - ne pas inclure la ponctuation
    - ne pas inclure les déterminants (« un », « des », etc.)
    - ne pas inclure les conjonction (« mais », « ou », etc.)
    - ne pas inclure les pronoms (« je », « tu », etc.)
    - ne pas inclure les verbes communs (« être », « avoir », « faire », etc.)
    - etc.
  - utiliser une **forme normalisée des mots** (fusion de mots différents en un seul)
    - **enlever les majuscules** (« Perceptron » → « perceptron »)
    - **lemmatiser** les mots (« marchons » → « marcher »,
      « suis » → « être », « est » → « être »)
- Il n'y a pas de recette universelle, le meilleure choix de *V* varie d'une application à l'autre

Probabilités

- Les assertions probabilistes facilitent la modélisation :
  - **des faits et de règles complexes** : comparée aux règles de production, l’approche est moins sensible à l’impossibilité d’énumérer toutes les exceptions, antécédents ou conséquences de règles
  - **de l’ignorance** : l’approche est moins sensible à l’omission/oubli des faits, de prémisses ou des conditions initiales à un raisonnement

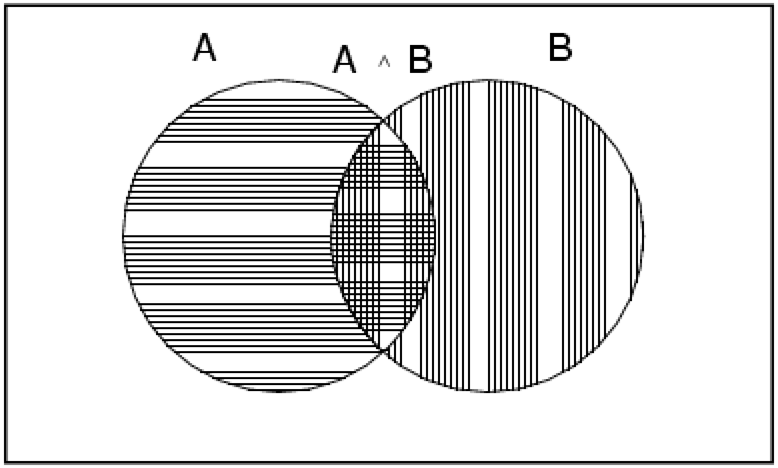
Probabilités

- Perspective **subjective/bayésienne** des probabilités :
  - les probabilités expriment le degré de croyance d'un agent dans des propositions/faits
    - » exemple : $P(A_{25}$ | aucun accident rapporté) = 0.06
  - les probabilités ne sont pas des assertions sur ce qui est vrai de façon absolue
  - n'expriment pas forcément des tendances/fréquences d'une situation, mais pourraient être apprises automatiquement à partir d'expériences
  - les probabilités des propositions changent avec l'acquisition de nouvelles informations
    - » exemple : $P(A_{25}$ | aucun accident rapporté, 5h du matin) = 0.15
- À l'opposée, il y a la perspective **objective/fréquentiste** des probabilités
  - les probabilités expriment des faits/propriétés sur des objets
  - on peut estimer ces probabilités en observant ces objets à plusieurs reprises
  - les physiciens diront que les phénomènes quantiques sont objectivement probabilistes

Axiomes de la théorie des probabilités : Axiomes de Kolmogorov

- Pour toute propositions *a*, *b*
  - $0 \leq P(a) \leq 1$
  - $P(vrai) = 1$ et $P(faux) = 0$
  - $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

True



Prise de décisions avec incertitude

- Supposons que je crois ceci :
  - $P(A_{25}$ me permet d'arriver à temps | …) = 0.04
  - $P(A_{90}$ me permet d'arriver à temps | …) = 0.70
  - $P(A_{120}$ me permet d'arriver à temps | …) = 0.95
  - $P(A_{240}$ me permet d'arriver à temps | …) = 0.999
  - $P(A_{1440}$ me permet d'arriver à temps | …) = 0.9999
- Quelle action devrais-je choisir?
  - cela dépend de mes **préférences** : manquer l'avion vs. trop d'attente
- **La théorie de l'utilité** est utilisée pour modéliser et inférer avec des préférences
  - une préférence exprime le degré d'utilité d'une action/situation

- **Théorie de la décision** = **théorie des probabilités** + **théorie de l'utilité**

Probabilités : traitement général

- On commence avec un ensemble Ω appelé **univers**
  - $\omega \in \Omega$ est un **événement élémentaire**
- Un **modèle de probabilités** est une distribution de probabilité $P(\omega)$ pour chaque élément $\omega \in \Omega$, telle que
  - $0 \leq P(\omega) \leq 1$
  - $\Sigma_{\omega \in \Omega} P(\omega) = 1$
- Un **événement** est un sous-ensemble de Ω
  - probabilité d'un événement A : $P(\mathrm{A}) = \Sigma_{\{\omega \in \mathrm{A}\}} P(\omega)$

- Exemple d'un dé :
  - $\Omega = \{1,2,3,4,5,6\}$ et $P(1)=P(2)=P(3)=P(4)=P(5)=P(6)=1/6$
- Événement A = « Dé est < 4 » :
  - $P(\mathrm{A}) = P(\omega=1)\ P(\omega=2) + P(\omega=3) = 1/6+1/6+1/6 = 1/2$

Variable aléatoire

- Une **variable aléatoire** est une variable décrivant **une partie** des connaissances incertaines (on la note avec une **première lettre majuscule**)
  - c'est une « fenêtre » sur l'univers

- Chaque variable a un **domaine** de valeurs qu'elle peut prendre

- **Types de variables aléatoires** :
  - **Booléennes** : le domaine est {*vrai*, *faux*}
    - » exemple : *Pourriel* $\in$ {*vrai*, *faux*} (ai-je la Pourriel?)
  - **Discrètes** : le domaine est énumérable
    - » *Météo* $\in$ {*soleil*, *pluie*, *nuageux*, *neige*}
  - **Continues** : le domaine est continu (par exemple, l'ensemble des réels)
    - » exemple : $X = 4.0$, $PositionX \leq 10.0$, $Speed \leq 20.5$

Variable aléatoire

- On peut voir une variable aléatoire X comme une fonction $X(\omega)$ donnant une valeur à chaque événement élémentaire ω possible
  - sauf si nécessaire, on va écrire X plutôt que $X(\omega)$

- P induit une **distribution de probabilités** pour chaque variable aléatoire X
  - la probabilité qu'une variable X ait la valeur x_i est la somme des probabilités d'événements élémentaires ω qui sont tels que $X(\omega) = \mathrm{x}$

$$P(X{=}x_i) = \Sigma_{\{\omega\ :\ X(\omega)=\mathrm{xi}\ \}}\ P(\omega)$$

- Exemple du dé :
  - $P(\ \mathit{NombreImpaire} = \mathit{vrai}) = P(1)+P(3)+P(5) = 1/6+1/6+1/6=1/2$

Propositions

- Une **proposition** est une assertion de ce qui est vrai, c.-à-d., une assertion sur la valeur d'une ou plusieurs variables
  - en d'autres mots, un événement (ensemble d'échantillons ou d'événements atomiques) pour lequel la proposition est vraie
    - exemple : *Pourriel* = *vrai* (noté parfois *Pourriel*) ou *Pourriel* = *faux* ($\neg$ *Pourriel*)

- Étant données deux variables booléennes *A* et *B* :
  - l'événement *a* est l'ensemble d'échantillons ω pour lesquels *A* = *vrai*
  - l'événement $\neg$ *a* est l'ensemble d'échantillons ω pour lesquels *A* = *faux*
  - l'événement $a \wedge b$ est l'ensemble des ω pour lesquels *A*=*vrai* et *B*=*vrai*
  - l'événement $a \vee b$ est l'ensemble des ω pour lesquels *A*=*vrai* ou *B*=*vrai*

Propositions

- Souvent nous aurons plusieurs variables aléatoires
  - toutes les variables aléatoires tiennent leur valeur d'un même échantillon ω
  - pour des variables distinctes, l'espace d'échantillonnage est alors le **produit cartésien** des domaines des variables aléatoires
- Un **événement atomique** est donc une spécification complète de l'état du « monde » pour lequel un agent est incertain
  - par exemple, si le « monde » de l'agent est décrit par seulement deux variables aléatoires booléennes (*Pourriel* et *Inconnu*), il y a exactement quatre états / événements atomiques possibles :
    - » *Pourriel = faux* $\wedge$ *Inconnu = faux*
    - » *Pourriel = faux* $\wedge$ *Inconnu = vrai*
    - » *Pourriel = vrai* $\wedge$ *Inconnu = faux*
    - » *Pourriel = vrai* $\wedge$ *Inconnu = vrai*
    - » on a donc Ω = { *<vrai,vrai>*, *<vrai,faux>*, *<faux,vrai>*, *<faux, faux>*}
- Les événements atomiques sont exhaustifs et mutuellement exclusifs

Syntaxe des propositions

- Élément de base : variable aléatoire
- Similaire à la logique propositionnelle

- **Variables aléatoires booléenne**
  - exemple : *Pourriel* = *vrai*

- **Variables aléatoires discrètes** (domaines **finis** or **infinis**)
  - exemple : *Météo* = *v*, avec $v \in$ { *soleil*, *pluie*, *nuageux*, *neige* }

- **Variables aléatoires continues** (**bornées** ou **non bornées**)
  - exemple : *Temp*=21.6 (la variable *Temp* a exactement la valeur 21.6)
  - exemple : *Temp* < 22.0 (la variable *Temp* a une valeur inférieure à 22)

Syntaxe des propositions

- En général, les **propositions élémentaires** sont définies en assignant une valeur ou un intervalle de valeurs aux variables
  - exemple : *Météo = soleil*, *Pourriel = faux* (notée aussi $\neg$*Pourriel*)

- Les **propositions complexes** sont définies par des combinaisons booléennes
  - exemple : (*Météo = soleil*) $\vee$ (*Pourriel = faux*)

Variable aléatoire

- **Variables aléatoires** :
  - ***Inconnu*** : l'adresse de l'expéditeur n'est pas connue du destinataire
  - ***Sensible*** : le courriel contient un mot sensible
  - ***Pourriel*** : le courriel est un pourriel

	Inconnu = vrai		*Inconnu = faux*	
	Sensible = vrai	*Sensible = faux*	*Sensible = vrai*	*Sensible = faux*
Pourriel = vrai	0.108	0.012	0.072	0.008
Pourriel = faux	0.016	0.064	0.144	0.576