Définitions

- Document $[w_1,...,w_d]$: une liste de mots
  - pourrait être tout un texte
  - pourrait être une seule phrase
  - pourrait être quelques mots

- Mots w_i : un mot ou une ponctuation
  - on suppose que nos documents ont déjà été segmentés en mots
  - généralement facile à faire en anglais (on sépare en fonction des espaces et des ponctuations)
  - difficile en chinois ou en japonais
    (pas d'espaces entre les mots)

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* ?

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, »

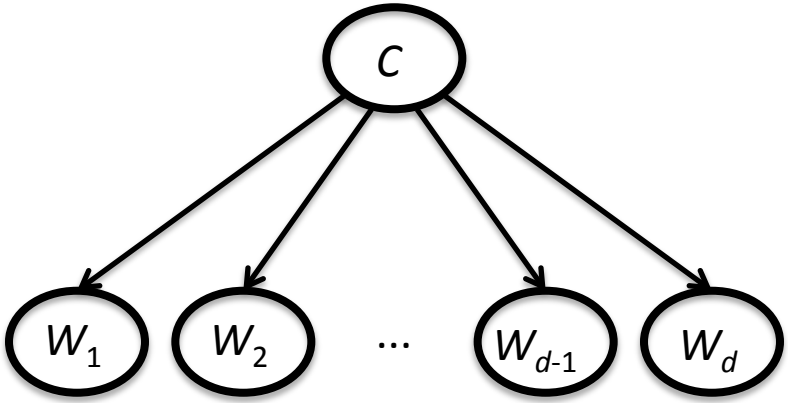
- Laquelle est une question d'examen *final* ?

Classification de documents

- Les mots individuels sont très informatifs du sujet (catégorie) d'un 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 **réseau bayésien**

Modèle bayésien naïf multinomial

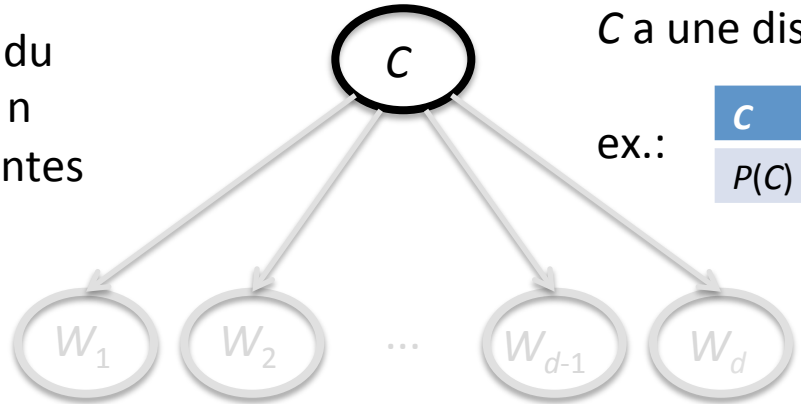
- Réseau bayésien: **modèle bayésien naïf multinomial**



Modèle bayésien naïf multinomial

- Réseau bayésien: **modèle bayésien naïf multinomial**

C est la catégorie du document, parmi n catégories différentes



C a une distribution a priori

ex.:

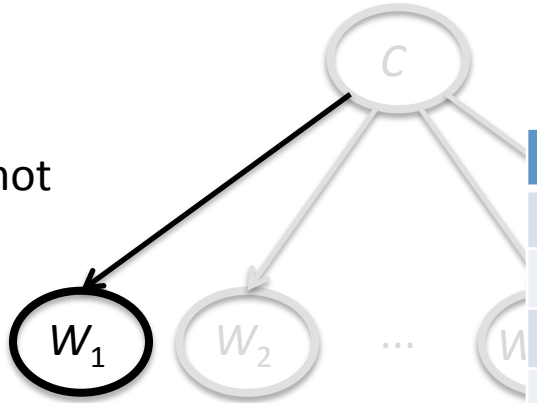
C	*intra*	*final*
P(*C*)	0.5	0.5

somme à 1

Modèle bayésien naïf multinomial

- Réseau bayésien: **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 »$\mid C)$	0.01	0.01
$P(W_1$=« qui »$\mid C)$	0.02	0.02
...	...	...
$P(W_1$=« perceptron »$\mid C)$	10^{-6}	0.002

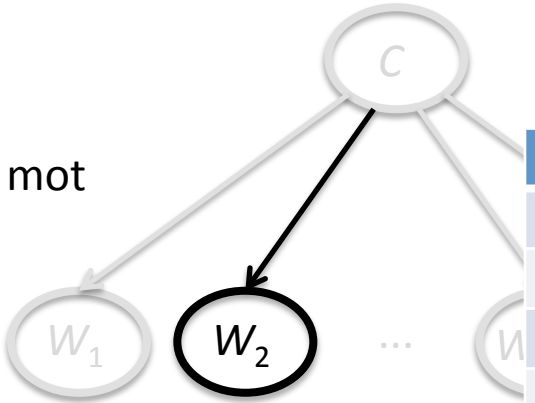
somme à 1

somme à 1

Modèle bayésien naïf multinomial

- Réseau bayésien: **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

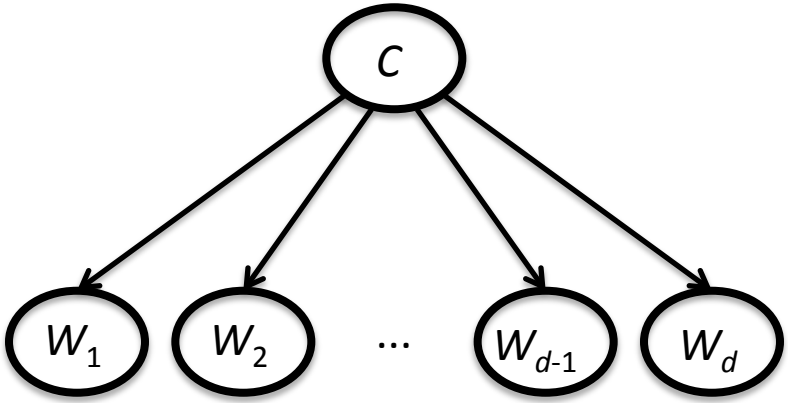
C	*intra*	*final*
$P(W_2$=« de »$\mid C)$	0.01	0.01
$P(W_2$=« qui »$\mid C)$	0.02	0.02
…	…	…
$P(W_2$=« perceptron »$\mid C)$	10^{-6}	0.002

somme à 1

somme à 1

Modèle bayésien naïf multinomial

- Réseau bayésien: **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*
$P(W_i$=« , »$\mid C)$	0.01	0.01
$P(W_i$=« un »$\mid C)$	0.02	0.02
$P(W_i$=« d' »$\mid C)$	0.01	0.02
$P(W_i$=« Perceptron »$\mid C)$	10^{-6}	0.002
$P(W_i$=« algorithme »$\mid C)$	0.005	0.005
$P(W_i$=« apprentissage »$\mid C)$	10^{-5}	0.001
$P(W_i$=« . »$\mid C)$	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 $[w_1,...,w_d]$, on choisit la classe c ayant la plus grande **probabilité a posteriori** $P(C=c|[w_1,...,w_d])$

$$
\begin{aligned}
& \underset{c}{\text{argmax}}\ P(C=c|[w_1,...,w_d]) \\
= \ & \underset{c}{\text{argmax}}\ P(C=c,[w_1,...,w_d]) \,/\, \alpha \\
= \ & \underset{c}{\text{argmax}}\ P(C=c,[w_1,...,w_d]) \quad \leftarrow \text{pour simplifier les calculs} \\
= \ & \underset{c}{\text{argmax}}\ \log P(C=c,[w_1,...,w_d]) \\
= \ & \underset{c}{\text{argmax}}\ \log \left(P(C=c) \prod_i P(W_i = w_i \mid C=c) \right) \\
= \ & \underset{c}{\text{argmax}}\ \log P(C=c) + \sum_i \log P(W_i = w_i \mid C=c) \quad \leftarrow \text{pour éviter le « } \textit{underflow} \text{ »}
\end{aligned}
$$

Décision de la catégorie d'un document

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

- Exemple:

argmax $\log P(C=c) + \sum_i \log P(W_i = w_i \mid C=c)$

= 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 = intra

C = final

= *final*