Modèles probabilistes pour l'accès à l'information à grande échelle

Introduction et Généralités

François Yvon

LIMSI — CNRS and Université Paris Sud

2015 / 2016

Questions pratiques

- Qui ? `francois.yvon@limsi.fr`
- Quand ? Les mercredi AM - 14 :00-17 :15 du 25/11 au 27/01
  sauf le 2 décembre
- Infos ? `http://tinyurl.com/phz7ulj` : pensez à vous inscrire pour être informés

Plan du cours

- séance 1 : introduction, les problèmes d'accès à l'information, les modèles graphiques
- séance 2 : EM et les modèles de thèmes : multinomial, PLSA, LDA, et autres
- séance 3 : modèles génératifs pour la syntaxe et la traduction automatique
- séance 4 : modèles de séquences, CRFs pour l'annotation de séquence, représentations non-orientées pour les MG
- séance 5 : inférence dans les MG, inférence exacte, algorithme d''élimination des variables, algorithme max-sum max-product
- séance 6 : inférence approchée ; méthodes d'échantillonage
- séance 7 : inférence approchée, propagation de croyance avec cycles, méthodes variationnelles

Pré-requis, Evaluation

Pré-requis

- bases de probabilités
- estimation supervisée (Bayésien naïf, HMM)
- un peu d'optimisation

Evaluation

- Contrôle continu ?
- Contrôle de connaissance

\+ travail personnel

L’avalanche des contenus non structurés

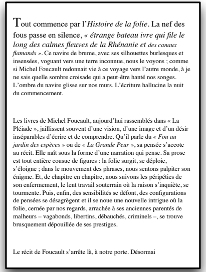
- 30 000 milliards de pages sont indexées par Google
- 3,3 milliards de requêtes chaque jour (100 milliards par mois)
- plus de traductions chaque jour tous les traducteurs en une année
- e-mails envoyés chaque jour : >200 milliards
- Messages envoyés chaque jour : 10 milliards
- Photos ajoutées chaque jour : 350 millions
- Contenus partagés chaque jour : 4,75 milliards
- 500 millions de tweets chaque jour
- > 10 milliards de whatsapp chaque jour
- > 180 milliard de SMS chaque année
- > 4M posts de blog chaque jour
- 1.7 millions de pages en français

+vidéos Youtube, produits Amazon / E-bay, etc

Des traitements automatiques typiques

Et les analyses nécessaires

Indexer / Classer



Tout commence par l'*Histoire de la folie*. La nef des fous passe en silence, *« étrange bateau ivre qui file le long des calmes fleuves de la Rhénanie et des canaux flamands »*. Ce navire de brume, avec ses silhouettes burlesques et insensées, voguant vers une terre inconnue, nous le voyons ; comme si Michel Foucault redonnait vie à ce voyage vers l'autre monde, à je ne sais quelle sombre croisade qui a peut-être hanté nos songes. L'ombre du navire glisse sur nos murs. L'écriture hallucine la nuit du commencement.

Les livres de Michel Foucault, aujourd'hui rassemblés dans « La Pléiade », jaillissent souvent d'une vision, d'une image et d'un désir inséparables d'écrire et de comprendre. Qu'il parle du *« Fou au jardin des espèces »* ou de *« La Grande Peur »*, sa pensée s'accote au récit. Elle naît sous la forme d'une narration qui pense. Sa prose est tout entière cousue de figures : la folie surgit, se déploie, s'éloigne ; dans le mouvement des phrases, nous sentons palpiter son énigme. Et, de chapitre en chapitre, nous suivons les péripéties de son enfermement, le lent travail souterrain où la raison s'inquiète, se tourmente. Puis, enfin, des sensibilités se défont, des configurations de pensées se désagrègent et il se noue une nouvelle intrigue où la folie, cernée par nos regards, arrachée à ses anciennes parentés de malheurs – vagabonds, libertins, débauchés, criminels –, se trouve brusquement dépouillée de ses prestiges.

Le récit de Foucault s'arrête là, à notre porte. Désormai

SPAM / HAM
Pertinent / Non-Pertinent
Positive / Negative

Sport / Business / Politique

Panne (O/N); Achat (O/N) ...

Document: texte, page, mail, tweet, ...

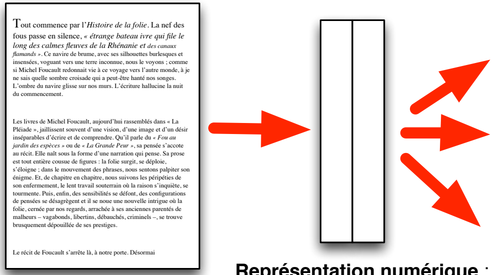
Classe:, thème,catégorie, polarité ...

s

Des traitements automatiques typiques

Et les analyses nécessaires

Indexer / Classer



SPAM / HAM
Pertinent / Non-Pertinent
Positive / Negative

Sport / Business / Politique

Panne (O/N); Achat (O/N) ...

Représentation numérique : BOW

Document: texte, page, mail, tweet, ...

Classe:, thème,catégorie, polarité ...

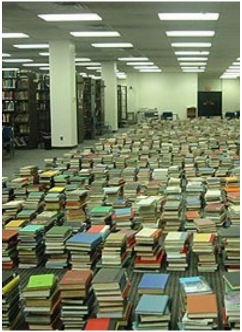
s

Des traitements automatiques typiques

Et les analyses nécessaires

Structurer / Segmenter

Collection: ensemble de documents



Thèmes: groupes, clusters thématiques

s

Des traitements automatiques typiques

Et les analyses nécessaires

Structurer / Segmenter

Sunday, February 3, 2008

Microsoft Responds To Google Missive (That Was Quick) de es fr it nl pt

Microsoft General Counsel Brad Smith has responded to today's missive from Google on the Microsoft-Yahoo acquisition by highlighting Google's dominance in search and advertising: The combination of Microsoft and Yahoo! will create a more competitive marketplace by establishing a compelling number....
techcrunch 11:32:00 PM CET

Is Microsoft gearing up for yesterday's battle?
IHT 12:45:00 PM CET

Yahoo Needs Time To Mull Microsoft Offer
iran-daily 5:49:00 PM CET

Google warns on Yahoo-Microsoft
news.com 10:41:00 PM CET

Google fires back at rival Microsoft
msnbc 10:16:00 PM CET

Microsoft Goes After Yahoo: Too Late?
ABCnews 4:30:00 AM CET

Microsoft bid for Yahoo! under scrutiny
GulfDailyNews 6:22:00 AM CET

Open-source silver lining in Microsoft's $44.6 billion wedding vow to Yahoo?
news.com 4:31:00 AM CET

Yahoo sale could hurt tech start-ups
IHT-tech 9:15:00 PM CET

Can Google Still Claim To Be David To Microsoft's Goliath? No.
techcrunch 10:39:00 PM CET

Microsoft-Yahoo deal poses antitrust issues: Google (Reuters)
news-yahoo 9:56:00 PM CET

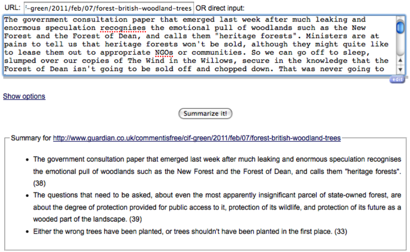
Microsoft bids $44.6 billion for Yahoo
msnbc 8:45:00 AM CET

Google balks at Microsoft bid for Yahoo (AFP)
news-yahoo 11:57:00 PM CET

Des traitements automatiques typiques

Et les analyses nécessaires

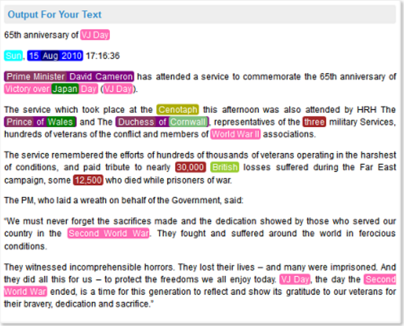
Visualiser / Résumer



Des traitements automatiques typiques

Et les analyses nécessaires

Comprendre / Distiller



Output For Your Text

65th anniversary of VJ Day

Sun, 15 Aug 2010 17:16:36

Prime Minister David Cameron has attended a service to commemorate the 65th anniversary of Victory over Japan Day (VJ Day).

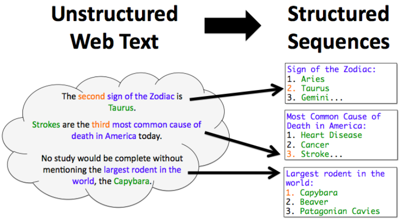
The service which took place at the Cenotaph this afternoon was also attended by HRH The Prince of Wales and The Duchess of Cornwall, representatives of the three military Services, hundreds of veterans of the conflict and members of World War II associations.

The service remembered the efforts of hundreds of thousands of veterans operating in the harshest of conditions, and paid tribute to nearly 30,000 British losses suffered during the Far East campaign, some 12,500 who died while prisoners of war.

The PM, who laid a wreath on behalf of the Government, said:

"We must never forget the sacrifices made and the dedication showed by those who served our country in the Second World War. They fought and suffered around the world in ferocious conditions.

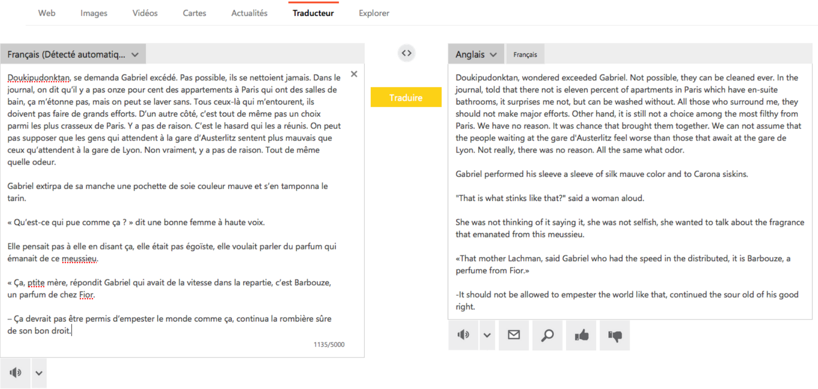
They witnessed incomprehensible horrors. They lost their lives – and many were imprisoned. And they did all this for us – to protect the freedoms we all enjoy today. VJ Day, the day the Second World War ended, is a time for this generation to reflect and show its gratitude to our veterans for their bravery, dedication and sacrifice."



Des traitements automatiques typiques

Et les analyses nécessaires

Traduire



Stucturer / Analyser / Enrichir / Annoter

Analyse Lexicale

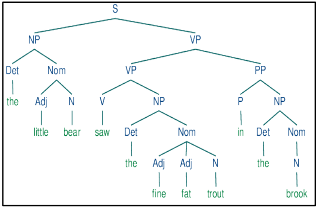
Lemmatisation : donne, donnera, donnerons --> donner

Racinisation : donne, donnera, donnerons, donation --> don+

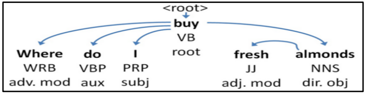
Etiquetage en Parties du discours (POS)

La/DET coronarographie/N met/V en/PREP évidence/N des/DET lésions/N bitronculaires/ADJ ./POINTFINAL

Analyse en constituants



Analyse en dépendances

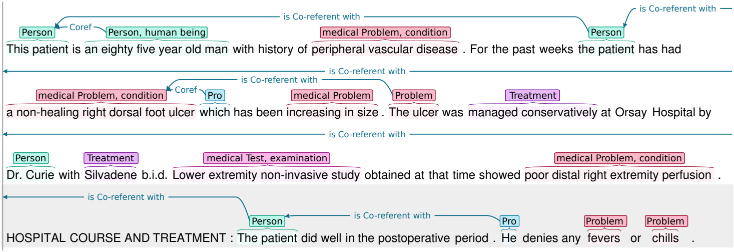


Des caractères aux représentations syntaxiques (...)

Des traitements automatiques typiques

Et les analyses nécessaires

Stucturer / Analyser / Enrichir / Annoter



(...) aux représentations sémantiques

La décision probabiliste

Caractéristiques récurrentes

- Grandes dimensions (données, représentations, ensembles de catégories)
- Décision multi-factorielle
- Données structurées

La décision probabiliste

- Probabilise les décisions : $P(C|d)$ plutôt que $\text{argmax}_c \, \text{score}(c, d)$
- Facilite :
  - la formulation des a priori
  - l'enchainement des modules
  - l'interprétation des résultats
  - l'expression de l'incertitude, les mesures de confiance

La décision probabiliste

Caractéristiques récurrentes

- Grandes dimensions (données, représentations, ensembles de catégories)
- Décision multi-factorielle
- Données structurées

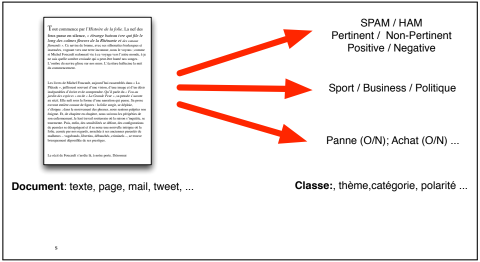
La décision probabiliste

- Probabilise les décisions : $P(C|d)$ plutôt que $\mathrm{argmax}_c\ \mathrm{score}(\mathrm{c}, \mathrm{d})$
- Facilite :
  - la formulation des a priori
  - l'enchainement des modules
  - l'interprétation des résultats
  - l'expression de l'incertitude, les mesures de confiance

Classification de documents et décision probabiliste

Un exemple emblématique

- tâche : assigner automatiquement à chaque document une ou plusieurs étiquette(s) (thème, mot-clés, pertinence, destinataires, etc.)



- hypothèses :
  - des documents étiquetés avec certitude sont disponibles (base d'apprentissage)
  - les étiquettes sont connues à l'avance, en nombre fini, éventuellement structurées (hiérarchie de thèmes ou d'index) ;
- mesure de succès : proportion de documents correctement étiquetés

La catégorisation supervisée

Formalisation du problème

Ce qui est donné

- un ensemble fini de catégories $y \in \mathcal{Y} = \{1 \dots n_K\}$
  pour $n_T = 2$: catégorisation binaire, $y \in \{+1, -1\}$ ou $y \in \{0, 1\}$
- un espace d'observables $x \in \mathcal{X}$;
- une fonction $f : \mathcal{X} \to \{0 \dots n_K\}$
- un ensemble d'exemples étiquetés $\mathcal{C} = \{(x_i, y_i), i = 1 \dots N\}$

Objectif

- construire $h : \mathcal{X} \to \{0 \dots n_K\}$
- tq. $h \approx f$ (au sens d'une mesure de succès)
- h dépend de $\mathcal{C}$

La catégorisation supervisée : l'approche probabiliste

Hypothèses

- chaque observation x_i est la réalisation d'une V.A. X_i
- chaque observation y_i est la réalisation d'une V.A. Y_i
- les couples de V.A. (X_i, Y_i) sont indépendantes entre elles et de même loi
- tirées sous une distribution inconnue $\mathcal{D}$

Démarche

- Modéliser la dépendance entre Y et X selon $\mathrm{P}(X, Y) \propto \mathrm{P}(Y)\,\mathrm{P}(X \mid Y)$
  - X est dans $\mathbb{R}^d$: $X|Y \sim \mathcal{N}(\mu, \Sigma)$
  - X est dans $\mathbb{N}^d$: $X|Y \sim \mathrm{Mult}(l, d)$
  - X est dans $\mathcal{S}(d)$: $X|Y \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_d)$
  - etc.
- estimer les paramètres de ces distributions
- règle de décision $y = h(x) = \mathrm{argmax}_{y' \in \mathcal{Y}}\, \mathrm{P}(Y = y' \mid X = x) \propto \mathrm{P}(X = x, Y = y')$

La catégorisation supervisée : l'approche probabiliste

Hypothèses

- chaque observation x_i est la réalisation d'une V.A. X_i
- chaque observation y_i est la réalisation d'une V.A. Y_i
- les couples de V.A. (X_i, Y_i) sont indépendantes entre elles et de même loi
- tirées sous une distribution inconnue $\mathcal{D}$

Démarche

- Modéliser la dépendance entre Y et X selon $\mathrm{P}(X, Y) \propto \mathrm{P}(Y)\,\mathrm{P}(X \mid Y)$
  - X est dans $\mathbb{R}^d$: $X|Y \sim \mathcal{N}(\mu, \Sigma)$
  - X est dans $\mathbb{N}^d$: $X|Y \sim \mathrm{Mult}(l, d)$
  - X est dans $\mathcal{S}(d)$: $X|Y \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_d)$
  - etc.
- estimer les paramètres de ces distributions
- règle de décision $y = h(x) = \mathrm{argmax}_{y' \in \mathcal{Y}}\, \mathrm{P}(Y = y' \mid X = x) \propto \mathrm{P}(X = x, Y = y')$

Bayésien Naïf, modèle de Bernoulli

Un document est comme un sac de pièces

Hypothèses :

- vocabulaire fixé et connu à l'avance (de taille n_W)
- documents représentants $\boldsymbol{x} \in \{0,1\}^{n_W}$ indépendants entre eux
- x_w encode la présence/absence du mot w dans $\boldsymbol{x}$: $x_w \sim \mathcal{B}(\beta_w)$
- les composants de $\boldsymbol{x}$ indépendantes entre elles
  $\Rightarrow n_W$ paramètres (par classe)

$$\mathrm{P}(\boldsymbol{x}; (\beta_1, \ldots, \beta_{n_W})) = \prod_{w=1}^{n_W} \beta_w^{x_w} (1 - \beta_w)^{(1-x_w)}$$

Le modèle de Bernoulli en couleurs

Avec un vocabulaire de 9 mots

1 $\beta_{1,blue}$	2 $\beta_{2,blue}$
1 $\beta_{1,brown}$	2 $\beta_{2,brown}$
1 $\beta_{1,cyan}$	2 $\beta_{2,cyan}$
1 $\beta_{1,green}$	2 $\beta_{2,green}$
1 $\beta_{1,mag}$	2 $\beta_{2,mag}$
1 $\beta_{1,oran}$	2 $\beta_{2,oran}$
1 $\beta_{1,pink}$	2 $\beta_{2,pink}$
1 $\beta_{1,red}$	2 $\beta_{2,red}$
1 $\beta_{1,yell}$	2 $\beta_{2,yell}$

Le modèle

blue	0
brown	0
cyan	0
green	0
magenta	0
orange	0
pink	0
red	0
yellow	0

document $\boldsymbol{x}$

On choisit au hasard un tas de pièces : avec proba $P(Y = 1)$. Supposons $Y = 1$

Le modèle de Bernoulli en couleurs

Avec un vocabulaire de 9 mots

1	$\beta_{1,blue}$
1	$\beta_{1,brown}$
1	$\beta_{1,cyan}$
1	$\beta_{1,green}$
1	$\beta_{1,mag}$
1	$\beta_{1,oran}$
1	$\beta_{1,pink}$
1	$\beta_{1,red}$
1	$\beta_{1,yell}$

Le modèle (classe 1)

blue	1
brown	0
cyan	0
green	0
magenta	0
orange	0
pink	0
red	0
yellow	0

document $\boldsymbol{x}$

On prend blue avec $P(blue = 1|Y = 1) = \beta_{1,blue}$

Le modèle de Bernoulli en couleurs

Avec un vocabulaire de 9 mots

1	$\beta_{1,blue}$
1	$\beta_{1,brown}$
1	$\beta_{1,cyan}$
1	$\beta_{1,green}$
1	$\beta_{1,mag}$
1	$\beta_{1,oran}$
1	$\beta_{1,pink}$
1	$\beta_{1,red}$
1	$\beta_{1,yell}$

Le modèle (classe 1)

blue	1
brown	0
cyan	0
green	0
magenta	0
orange	0
pink	0
red	0
yellow	0

document $\boldsymbol{x}$

On rejette brown avec $P(brown = 0 | Y = 1) = 1 - \beta_{1,brown}$

Le modèle de Bernoulli en couleurs

Avec un vocabulaire de 9 mots

1	$\beta_{1,blue}$
1	$\beta_{1,brown}$
1	$\beta_{1,cyan}$
1	$\beta_{1,green}$
1	$\beta_{1,mag}$
1	$\beta_{1,oran}$
1	$\beta_{1,pink}$
1	$\beta_{1,red}$
1	$\beta_{1,yell}$

Le modèle (classe 1)

blue	1
brown	0
cyan	1
green	0
magenta	0
orange	0
pink	0
red	0
yellow	0

document $\boldsymbol{x}$

On prend cyan avec $P(cyan = 1 | Y = 1) = \beta_{1,cyan}$

Le modèle de Bernoulli en couleurs

Avec un vocabulaire de 9 mots

1	$\beta_{1,blue}$
1	$\beta_{1,brown}$
1	$\beta_{1,cyan}$
1	$\beta_{1,green}$
1	$\beta_{1,mag}$
1	$\beta_{1,oran}$
1	$\beta_{1,pink}$
1	$\beta_{1,red}$
1	$\beta_{1,yell}$

Le modèle (classe 1)

blue	1
brown	0
cyan	1
green	1
magenta	0
orange	0
pink	0
red	0
yellow	0

document $\boldsymbol{x}$

On prend green avec $P(green|Y=1) = \beta_{1,green}$

Le modèle de Bernoulli en couleurs

Avec un vocabulaire de 9 mots

1	$\beta_{1,blue}$
1	$\beta_{1,brown}$
1	$\beta_{1,cyan}$
1	$\beta_{1,green}$
1	$\beta_{1,mag}$
1	$\beta_{1,oran}$
1	$\beta_{1,pink}$
1	$\beta_{1,red}$
1	$\beta_{1,yell}$

Le modèle (classe 1)

blue	1
brown	0
cyan	1
green	1
magenta	0
orange	0
pink	0
red	0
yellow	0

document $\boldsymbol{x}$

Bilan : $P(\boldsymbol{x}, Y = 1) \propto P(Y = 1)\beta_{1,blue}(1 - \beta_{1,brown})\beta_{1,cyan}\beta_{1,green}\cdots$

Classificateur à n_K classes

Sans information *a priori* sur les classes, la classe optimale pour $\boldsymbol{x}_\star$

$$y_\star = \underset{y=1\ldots n_K}{\operatorname{argmax}} \mathrm{P}(\boldsymbol{x} \,|\, y) \propto \prod_{w=1}^{w=n_W} \beta_{wy}^{x_{w\star}} \left(1 - \beta_{wy}\right)^{(1-x_{w\star})}$$

D'où viennent les paramètres β_{wy} ?

- corpus de documents indépendants $\mathcal{C} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$
- estimer β_{wy} pour chacune des classes y (peut être fait séparément)

Classificateur bayésien Bernoulli

Classificateur à n_K classes

Sans information *a priori* sur les classes, la classe optimale pour $\boldsymbol{x}_\star$

$$y_\star = \operatorname*{argmax}_{y=1\ldots n_K} \mathrm{P}(\boldsymbol{x} \,|\, y) \propto \prod_{w=1}^{w=n_W} \beta_{wy}^{x_{w\star}} (1 - \beta_{wy})^{(1-x_{w\star})}$$

D'où viennent les paramètres β_{wy} ?

- corpus de documents indépendants $\mathcal{C} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$
- estimer β_{wy} pour chacune des classes y (peut être fait séparément)

Estimation des paramètres

Un problème d'optimisation

L'estimateur du maximum de vraisemblance

- Considère la vraisemblance $P(\mathcal{C};\boldsymbol{\theta})$ (ou son log) comme une fonction de $\boldsymbol{\theta}$
- le point maximisant $P(\mathcal{C};\boldsymbol{\theta})$ est l'estimateur du maximum de vraisemblance (ML) $\widehat{\boldsymbol{\theta}}_{ML}$,

Propriétés sous des conditions très générales

- consistance (converge vers la vraie valeur... si le modèle est bien spécifié)
- de variance asymptotique minimale (efficacité asymptotique)

Estimation des paramètres

Un problème d'optimisation

L'estimateur du maximum de vraisemblance

- Considère la vraisemblance $\mathrm{P}(\mathcal{C}; \boldsymbol{\theta})$ (ou son log) comme une fonction de $\boldsymbol{\theta}$
- le point maximisant $\mathrm{P}(\mathcal{C}; \boldsymbol{\theta})$ est l'estimateur du maximum de vraisemblance (ML) $\widehat{\boldsymbol{\theta}}_{ML}$,

Propriétés sous des conditions très générales

- consistance (converge vers la vraie valeur... si le modèle est bien spécifié)
- de variance asymptotique minimale (efficacité asymptotique)

Maximum de vraisemblance

- La vraisemblance des documents d'une classe :

$$\mathrm{P}(\mathcal{C};\boldsymbol{\theta}) = \prod_{d=1}^{d=n_D} \prod_{w=1}^{n_W} \beta_w^{x_{dw}} (1-\beta_w)^{(1-x_{dw})}$$

- La fonction à maximiser (en $\boldsymbol{\theta}$) :

$$f(\boldsymbol{\theta}) = \log \mathrm{P}(\boldsymbol{x_1} \ldots \boldsymbol{x}_{n_D};\beta) = \sum_d \sum_{w=1}^{n_W} x_{dw} \log \beta_w + (1-x_{dw}) \log(1-\beta_w)$$

- Conditions d'optimalité :

$$\forall w = 1 \ldots n_W, \ \frac{df}{d\beta_w} = 0, \Rightarrow \widehat{\beta}_w = \frac{\sum_d x_{dw}}{n_D}$$

En pratique, attention à $\widehat{\beta}_w = 0$

Maximum de vraisemblance

- La vraisemblance des documents d'une classe :

$$\mathrm{P}(\mathcal{C};\boldsymbol{\theta}) = \prod_{d=1}^{d=n_D} \prod_{w=1}^{n_W} \beta_w^{x_{dw}} (1-\beta_w)^{(1-x_{dw})}$$

- La fonction à maximiser (en $\boldsymbol{\theta}$) :

$$f(\boldsymbol{\theta}) = \log \mathrm{P}(\boldsymbol{x_1} \ldots \boldsymbol{x}_{n_D}; \beta) = \sum_d \sum_{w=1}^{n_W} x_{dw} \log \beta_w + (1-x_{dw}) \log(1-\beta_w)$$

- Conditions d'optimalité :

$$\forall w = 1 \ldots n_W, \ \frac{df}{d\beta_w} = 0, \Rightarrow \widehat{\beta_w} = \frac{\sum_d x_{dw}}{n_D}$$

En pratique, attention à $\widehat{\beta_w} = 0$

Maximum de vraisemblance

- La vraisemblance des documents d'une classe :

$$\mathrm{P}(\mathcal{C}; \boldsymbol{\theta}) = \prod_{d=1}^{d=n_D} \prod_{w=1}^{n_W} \beta_w^{x_{dw}} (1 - \beta_w)^{(1-x_{dw})}$$

- La fonction à maximiser (en $\boldsymbol{\theta}$) :

$$f(\boldsymbol{\theta}) = \log \mathrm{P}(\boldsymbol{x_1} \ldots \boldsymbol{x}_{n_D}; \beta) = \sum_d \sum_{w=1}^{n_W} x_{dw} \log \beta_w + (1 - x_{dw}) \log(1 - \beta_w)$$

- Conditions d'optimalité :

$$\forall w = 1 \ldots n_W, \frac{df}{d\beta_w} = 0, \Rightarrow \widehat{\beta}_w = \frac{\sum_d x_{dw}}{n_D}$$

En pratique, attention à $\widehat{\beta}_w = 0$

Maximum de vraisemblance

- La vraisemblance des documents d'une classe :

$$P(\mathcal{C};\boldsymbol{\theta}) = \prod_{d=1}^{d=n_D} \prod_{w=1}^{n_W} \beta_w^{x_{dw}} (1-\beta_w)^{(1-x_{dw})}$$

- La fonction à maximiser (en $\boldsymbol{\theta}$) :

$$f(\boldsymbol{\theta}) = \log P(\boldsymbol{x_1} \ldots \boldsymbol{x}_{n_D};\beta) = \sum_d \sum_{w=1}^{n_W} x_{dw} \log \beta_w + (1-x_{dw}) \log(1-\beta_w)$$

- Conditions d'optimalité :

$$\forall w = 1 \ldots n_W, \frac{df}{d\beta_w} = 0, \Rightarrow \widehat{\beta}_w = \frac{\sum_d x_{dw}}{n_D}$$

En pratique, attention à $\widehat{\beta}_w = 0$

Maximum *a Posteriori* (MAP)

Un estimateur alternatif

Point de vue bayésien

$\boldsymbol{\theta}$ est aussi une variable aléatoire

Le modèle complet :

$$\mathrm{P}(\mathcal{C}, \boldsymbol{\theta}) = \mathrm{P}(\mathcal{C} \,|\, \boldsymbol{\theta})\, \mathrm{P}(\boldsymbol{\theta})$$

Estimation MAP

$$\max_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) = \log(\mathrm{P}(\mathcal{C}, \boldsymbol{\theta})) = \log(\mathrm{P}(\mathcal{C} \,|\, \boldsymbol{\theta})) + \log(\mathrm{P}(\boldsymbol{\theta}))$$

Loi *a priori* conjuguée

$$\mathrm{P}(\beta_1, \ldots, \beta_{n_W}) = \prod_{w=1}^{n_W} \mathrm{Beta}(\beta_w; \gamma_\beta, \delta_\beta), \text{ avec}$$

$$\mathrm{Beta}(\beta; \gamma_\beta, \delta_\beta) = \frac{\Gamma(\gamma_\beta + \delta_\beta)}{\Gamma(\gamma_\beta)\Gamma(\delta_\beta)} \beta^{\gamma_\beta - 1}(1 - \beta)^{\delta_\beta - 1} \mathbb{I}(\beta \in [0, 1])$$

$\Gamma()$ est la loi Gamma d'Euler (factorielle généralisée aux réels)

Lois conjuguées

Définition

La loi a posteriori de $\boldsymbol{\theta}$ est de la forme : $\mathrm{P}(\theta \,|\, \mathcal{C}) = \frac{\mathrm{P}(\mathcal{C} \,|\, \boldsymbol{\theta})\,\mathrm{P}(\boldsymbol{\theta})}{\int \mathrm{P}(\mathcal{C} \,|\, \boldsymbol{\theta})\,\mathrm{P}(\boldsymbol{\theta}) d\boldsymbol{\theta}}$. Si $\mathrm{P}(\boldsymbol{\theta})$ est choisi pour que $\mathrm{P}(\theta \,|\, \mathcal{C})\,\mathrm{P}(\boldsymbol{\theta})$ et $\mathrm{P}(\boldsymbol{\theta})$ aient la même forme, on dit que $\mathrm{P}(\theta)$ est la loi a priori conjuguée.

Illustrations

Bernoulli , Binomiale	Beta
Discrète, Multinomiale	Dirichlet
Poisson	Gamma

NB. Toutes les distributions de la famille exponentielle ont une distribution conjuguée.

Maximum *a posteriori* (MAP)

Un estimateur alternatif

Estimation MAP (suite)

$$g(\boldsymbol{\theta}) = \log(\mathrm{P}(\mathcal{C}, \boldsymbol{\theta})) = \log\Big(\prod_{d=1}^{n_D}\prod_{w=1}^{n_W} \beta_w^{x_{dw}}(1-\beta_w)^{(1-x_{dw})}\prod_{w=1}^{n_W}\mathrm{Beta}(\beta_w;\gamma_\beta,\delta_\beta)\Big)$$

$$= \sum_{w=1}^{n_W}\Big(\sum_d x_{dw}\log\beta_w + (1-x_{dw})\log(1-\beta_w)\Big) + (\gamma_\beta - 1)\log\beta_w + (\delta_\beta - 1)\log(1-\beta_w) + \mathrm{Cste}$$

Le terme d'*a priori* fait apparaître un pseudo document :

$$\widehat{\beta}_w = \frac{\sum_d x_{dw} + \gamma_\beta - 1}{n_D + \gamma_\beta - 1 + \delta_\beta - 1} \quad \text{plus de zéros } (\gamma_\beta, \delta_\beta \geq 1)$$

Conclusion

Partant d'une connaissance *a priori* modélisée par $\mathrm{P}(\boldsymbol{\theta})$, le traitement de $\mathcal{C}$ produit un modèle *a posteriori* $\mathrm{P}(\boldsymbol{\theta}|\mathcal{C}) \propto \mathrm{P}(\mathcal{C}|\boldsymbol{\theta})\mathrm{P}(\boldsymbol{\theta})$

Rq : Quand la loi *a priori* est Beta(), la loi *a posteriori* est aussi une loi Beta() : propriété générale des lois conjuguées. Mais comment choisir alors les méta-paramètres ? Comment les estimer ?

Maximum *a posteriori* (MAP)

Un estimateur alternatif

Estimation MAP (suite)

$$g(\boldsymbol{\theta}) = \log(\mathrm{P}(\mathcal{C},\boldsymbol{\theta})) = \log\Big(\prod_{d=1}^{n_D}\prod_{w=1}^{n_W}\beta_w^{x_{dw}}(1-\beta_w)^{(1-x_{dw})}\prod_{w=1}^{n_W}\mathrm{Beta}(\beta_w;\gamma_\beta,\delta_\beta)\Big)$$

$$= \sum_{w=1}^{n_W}\Big(\sum_d x_{dw}\log\beta_w + (1-x_{dw})\log(1-\beta_w)\Big) + (\gamma_\beta-1)\log\beta_w + (\delta_\beta-1)\log(1-\beta_w) + \mathrm{Cste}$$

Le terme d'*a priori* fait apparaître un pseudo document :

$$\widehat{\beta}_w = \frac{\sum_d x_{dw} + \gamma_\beta - 1}{n_D + \gamma_\beta - 1 + \delta_\beta - 1} \text{ plus de zéros } (\gamma_\beta, \delta_\beta \geq 1)$$

Conclusion

Partant d'une connaissance *a priori* modélisée par $\mathrm{P}(\boldsymbol{\theta})$, le traitement de $\mathcal{C}$ produit un modèle *a posteriori* $\mathrm{P}(\boldsymbol{\theta}\,|\,\mathcal{C}) \propto \mathrm{P}(\mathcal{C}\,|\,\boldsymbol{\theta})\,\mathrm{P}(\boldsymbol{\theta})$

Rq : Quand la loi *a priori* est Beta(), la loi *a posteriori* est aussi une loi Beta() : propriété générale des lois conjuguées. Mais comment choisir alors les méta-paramètres ? Comment les estimer ?

Encore plus bayésien : la loi prédictive

Choisir $\boldsymbol{\theta}$ ou l'intégrer ?

- l'estimation ML/MAP choisit une valeur de $\boldsymbol{\theta}$
- utiliser un estimateur ponctuel $\boldsymbol{\theta}$ est sous-optimal : meilleure idéee : moyenner sur toutes les valeurs possibles de $\boldsymbol{\theta}$

Loi prédictive bayésienne

$$
\begin{aligned}
h(x^*) = \operatorname*{argmax}_{y'=1\dots n_K} \mathrm{P}(x^*, y' \mid \mathcal{C}) &= \int_\theta \mathrm{P}(x^*, y', \boldsymbol{\theta} \mid \mathcal{C})\, dp\boldsymbol{\theta} \\
&= \int_\theta \mathrm{P}(x^* \mid y'; \boldsymbol{\theta})\, \mathrm{P}(\boldsymbol{\theta} \mid \mathcal{C})\, dp\boldsymbol{\theta} \text{ [quand } P(y) \text{ uniforme]}
\end{aligned}
$$

Encore plus bayésien : la loi prédictive

Choisir $\boldsymbol{\theta}$ ou l'intégrer ?

- l'estimation ML/MAP choisit une valeur de $\boldsymbol{\theta}$
- utiliser un estimateur ponctuel $\boldsymbol{\theta}$ est sous-optimal : meilleure idéee : moyenner sur toutes les valeurs possibles de $\boldsymbol{\theta}$

Loi prédictive bayésienne

$$h(x^*) = \operatorname*{argmax}_{y'=1\ldots n_K} \mathrm{P}(x^*, y' \mid \mathcal{C}) = \int_\theta \mathrm{P}(x^*, y', \boldsymbol{\theta} \mid \mathcal{C}) dp\boldsymbol{\theta}$$

$$= \int_\theta \mathrm{P}(x^* \mid y'; \boldsymbol{\theta}) \, \mathrm{P}(\boldsymbol{\theta} \mid \mathcal{C}) dp\boldsymbol{\theta} \text{ [quand } P(y) \text{ uniforme]}$$

Calculer la loi prédictive

Un modèle plus simple : choisir une pièce, puis la lancer

$\beta \sim Beta(\gamma, \delta); x \sim Bernoulli(\beta)$. La loi prédictive :

$$
\begin{aligned}
\mathrm{P}(x) &= \int_0^1 \beta^x (1-\beta)^{1-x}\, \mathrm{P}(\beta) d\beta \\
&= \frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma)\Gamma(\delta)} \int \beta^x (1-\beta)^{1-x} \beta^{\gamma-1} (1-\beta)^{\delta-1} d\beta \\
&= \frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma)\Gamma(\delta)} \int \beta^{x+\gamma-1} (1-\beta)^{1-x+\delta-1} d\beta \\
&= \frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma)\Gamma(\delta)} \frac{\Gamma(x+\gamma)\Gamma(1-x+\delta)}{\Gamma(\gamma+\delta+1)} \\
&= \frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma)\Gamma(\delta)} \frac{\gamma^x \Gamma(\gamma) \delta^{1-x} \Gamma(\delta)}{(\gamma+\delta)\Gamma(\gamma+\delta)} = \frac{\gamma+1^x \delta^{1-x}}{(\gamma+\delta)} \qquad (*)
\end{aligned}
$$

(*) car $\Gamma(x+1) = x\Gamma(x)$

Calculer la loi prédictive

Pour simplifier : $n_W = 1$, classes uniformes

$$\mathrm{P}(y \mid x, \mathcal{C}) \propto \mathrm{P}(x \mid y, \mathcal{C})$$

$$\begin{aligned}
\mathrm{P}(x, y \mid \mathcal{C}) &= \int \mathrm{P}(x \mid y; \boldsymbol{\theta})\, \mathrm{P}(\boldsymbol{\theta} \mid \mathcal{C}) d\boldsymbol{\theta} \\
&\propto \int \beta_y^x (1-\beta)^{1-x}\, \mathrm{P}(\mathcal{C} \mid \boldsymbol{\theta})\, \mathrm{P}(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int \beta_y^x (1-\beta)^{1-x} \prod_{k=1}^{n_K} \beta_k^{n_{k1}} (1-\beta_k)^{n_{k0}} \beta_k^{\gamma_\beta - 1} (1-\beta_k)^{\delta_\beta - 1} \\
&= \int \beta_y^{n_{y1}+x+\gamma_\beta-1} (1-\beta)^{n_{y0}+1-x+\delta_\beta-1} \prod_{k \neq y} \beta_k^{n_{k1}+\gamma_\beta-1} (1-\beta_k)^{n_{k0}+\delta_\beta-1} \\
&= \frac{(n_{y1}+\gamma_\beta)^x (n_{y0}+\delta_\beta)^{1-x}}{n_y+\gamma_\beta+\gamma_\delta} \prod_k \frac{\Gamma(n_{1k}+\gamma_\beta)\Gamma(n_{0k}+\delta_\beta)}{\Gamma(n_k+\gamma_\beta+\delta_\beta)}
\end{aligned}$$

Classifieur bayésien Bernoulli : un résumé

1. le modèle :

$$\mathrm{P}(\boldsymbol{x}; (\beta_1, \ldots, \beta_{n_W})) = \prod_{w=1}^{n_W} \beta_w^{x_w} (1 - \beta_w)^{(1-x_w)}$$

2. estimation

$$\text{ML} : \forall w, y, \widehat{\beta}_{wy} = \frac{\sum_{d \in y} x_{dw}}{\sum_{d \in y} 1}$$

$$\text{MAP} : \forall w, y, \widehat{\beta}_{wy} = \frac{\sum_{d \in y} x_{wd} + \gamma_\beta}{1 + \gamma_\beta - \delta_\beta + \sum_{d \in y} 1}$$

3. décisions :

$$\begin{aligned} y^* &= \operatorname*{argmax}_{y'=1\ldots n_K} \mathrm{P}(x^* \,|\, y', \boldsymbol{\theta}) \propto \prod_{w=1}^{w=n_W} \widehat{\beta}_{wy'}^{x_{*w}} (1 - \widehat{\beta}_{wy'})^{(1-x_{*w})} \\ &= \operatorname*{argmax}_{y'=1\ldots n_K} \mathrm{P}(x^* \,|\, y', \boldsymbol{\theta}) \propto \prod_{w=1}^{w=n_W} \widehat{\beta}_{wy'}^{x_{*w}} (1 - \widehat{\beta}_{wy'})^{(1-x_{*w})} \mathrm{P}(y') \\ &= \operatorname*{argmax}_{y'=1\ldots n_K} \mathrm{P}(x \,|\, y', \mathcal{C}) \propto \prod_{w=1}^{w=n_W} \mathrm{P}(x_w \,|\, y', \mathcal{C}) \, \mathrm{P}(y') \end{aligned}$$

Classifieur bayésien Bernoulli : une ouverture

1. estimation (MAP) :

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathrm{P}(\mathcal{C}, \boldsymbol{\theta}) \propto \mathrm{P}(\mathcal{C} \,|\, \boldsymbol{\theta})\, \mathrm{P}(\boldsymbol{\theta})$$

2. décision :

$$y^* = \underset{y'}{\operatorname{argmax}}\, \mathrm{P}(x^*, y' \,|\, \boldsymbol{\theta}; \mathcal{C}) = \mathrm{P}(x^* \,|\, y', \boldsymbol{\theta})\, \mathrm{P}(y')$$

Problèmes généraux

Calculer la distribution marginale, le mode ou l'espérance d'une VA dans un modèle impliquant de nombreuses variables, avec des dépendances complexes

Classifieur bayésien Bernoulli : une ouverture

1. estimation (MAP) :

$$\operatorname*{argmax}_{\boldsymbol{\theta}} \mathrm{P}(\mathcal{C}, \boldsymbol{\theta}) \propto \mathrm{P}(\mathcal{C} \,|\, \boldsymbol{\theta})\, \mathrm{P}(\boldsymbol{\theta})$$

2. décision :

$$y^* = \operatorname*{argmax}_{y'} \mathrm{P}(x^*, y' \,|\, \boldsymbol{\theta}; \mathcal{C}) = \mathrm{P}(x^* \,|\, y', \boldsymbol{\theta})\, \mathrm{P}(y')$$

Problèmes généraux

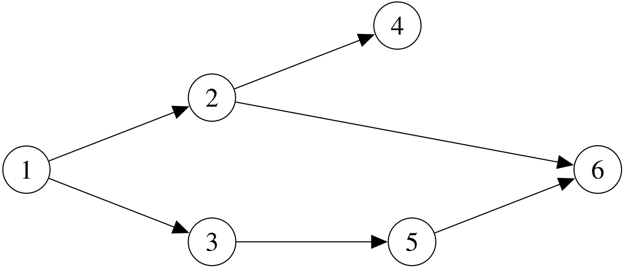
Calculer la distribution marginale, le mode ou l'espérance d'une VA dans un modèle impliquant de nombreuses variables, avec des dépendances complexes

Graphes orientés : définitions et notations de base

Définition

Un graphe $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$ fini est défini par :

- $\boldsymbol{V}$ un ensemble fini de sommets,
- $\boldsymbol{E} \subset \boldsymbol{V} \times \boldsymbol{V}$ un ensemble fini d'arcs (u, v)

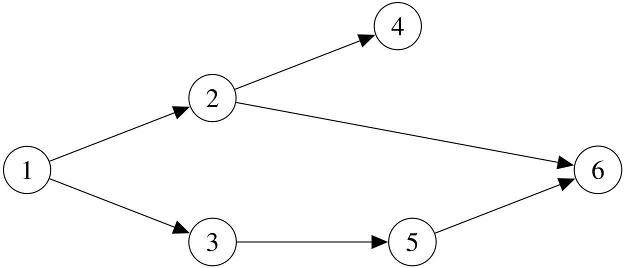


$$\boxed{V = 1, 2, 3, 4, 5, 6, E = \{(1, 2), \ldots, (5, 6)\}}$$

Relations de parenté

Généalogie graphique dans $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$

- parents de v : $pa(v) = \{u, (u, v) \in E\}$
- ancêtres de v : clotûtre transitive de $pa(v)$
- enfants de v : $en(v) = \{u, (v, u) \in E\}$ et les descendants
- degré (entrant, sortant) de v ($\Delta_+(v), \Delta_-(v)$) : nombre de parents, d'enfants de v

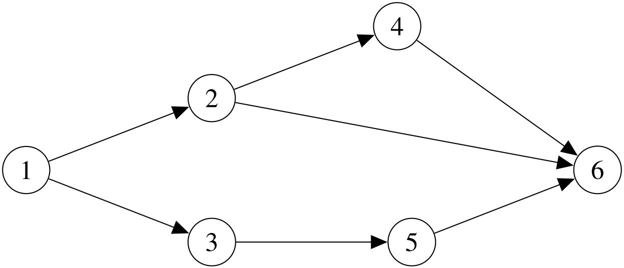


$\boxed{pa(1) = \varnothing; pa(2) = \{1\}; pa(6) = \{2, 4, 6\}; \Delta_+(5) = 1}$

Chemins, cycles

Chemins, cycles

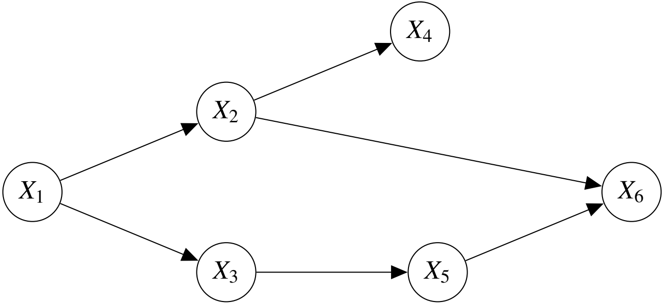
- Un chemin de longueur l dans $G = (V, E)$ est une séquence d'arcs $\pi = (u_1, v_1) \ldots \pi = (u_l, v_l)$ avec $\forall i > 1, u_i = v_{i-1}$.
- un cycle est un chemin π t.q. $\exists (i, j) \in V \times V, u_i = u_j$
- un graphe acyclique ne contient aucun cycle
- un graphe acyclique peut être trié topologiquement (chaque père avant ses fils)



Modèles graphiques

Modèle graphique

Un modèle graphique (a.k.a réseau bayésien) est un graphe orienté acyclique $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$ dans lequel chaque sommet est associé à une variable aléatoire.



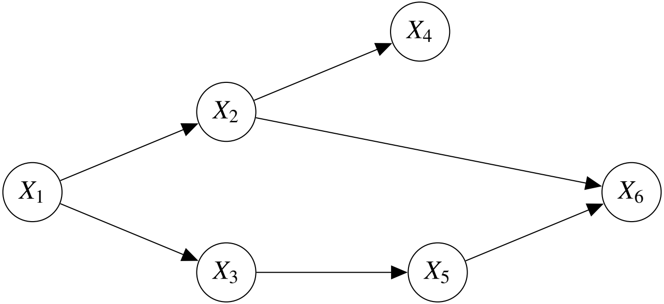
d'après un exemple de Michael Jordan

Factorisation de la loi jointe

Modèle graphique

Un modèle graphique représente une factorisation de la loi jointe

$$P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i \,|\, pa(X_i))$$



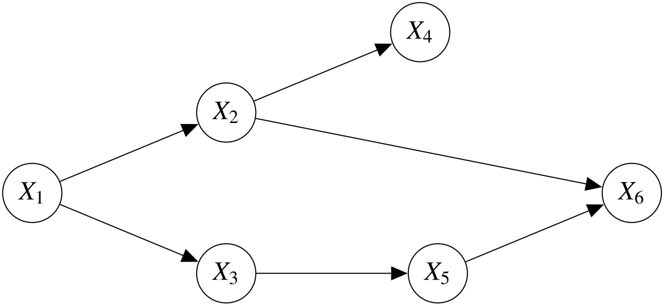
$$\boxed{P(X_1, X_2, X_3, X_4, X_5, X_6) = P(X_1)\,P(X_2 \,|\, X_1)\,P(X_3 \,|\, X_1)\,P(X_4 \,|\, X_2)\,P(X_5 \,|\, X_3)\,P(X_6 \,|\, X_2, X_5)}$$

Factorisation de la loi jointe

Modèle graphique

Un modèle graphique raconte une histoire générative

$$\mathrm{P}(X_1, \ldots, X_N) = \prod_{i=1}^{N} \mathrm{P}(X_i \mid pa(X_i))$$



choisir X_1 avec $\mathrm{P}(X_1)$ puis séparément X_2 avec $\mathrm{P}(X_2 \mid X_1)$ et X_3 avec $\mathrm{P}(X_3 \mid X_1)$...

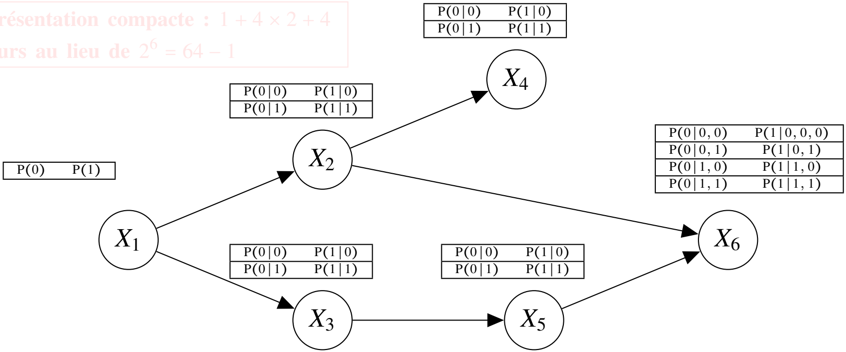
Factorisation de la loi jointe

Modèle graphique

Un modèle graphique représente une famille de distributions jointes

$$P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i \,|\, pa(X_i))$$

Représentation compacte : $1 + 4 \times 2 + 4$ valeurs au lieu de $2^6 = 64 - 1$



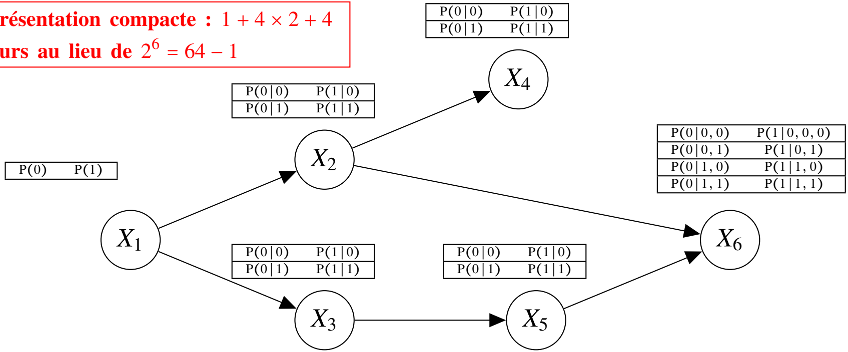
Factorisation de la loi jointe

Modèle graphique

Un modèle graphique représente une famille de distributions jointes

$$\mathrm{P}(X_1, \ldots, X_N) = \prod_{i=1}^{N} \mathrm{P}(X_i \,|\, pa(X_i))$$

Représentation compacte : $1 + 4 \times 2 + 4$ **valeurs au lieu de** $2^6 = 64 - 1$



Un réseau bayésien définit une distribution

Conditions stochastiques

1. $\mathrm{P}(X_1, \ldots, X_N) > 0$ comme produit de facteurs positifs ;
2. $\mathrm{P}(X_1, \ldots, X_N)$ est normalisé :

$$\begin{aligned}
\sum_{x_1,\ldots,x_n} \mathrm{P}(X_1 = x_1, \ldots, X_N = x_n) &= \sum_{x_1,\ldots,x_n} \prod_i \mathrm{P}(X_i = x_i \,|\, pa(X_i)) \\
&= \sum_{x_1,\ldots,x_{n-1}} \prod_i \mathrm{P}(X_i = x_i \,|\, pa(X_i)) \sum_{x_N} \mathrm{P}(X_N = x_N \,|\, pa(X_N)) \\
&= \sum_{x_1,\ldots,x_{n-1}} \prod_i \mathrm{P}(X_i = x_i \,|\, pa(X_i)) \underbrace{\sum_{x_N} \mathrm{P}(X_N = x_N \,|\, pa(X_N))}_{=1} \\
&= \sum_{x_1,\ldots,x_{n-1}} \prod_i \mathrm{P}(X_i = x_i \,|\, pa(X_i)) \\
&= \ldots \\
&= \sum_{x_1} \mathrm{P}(X_1 = x_1) = 1
\end{aligned}$$

Sous-réseau

Un sous-graphe de $\mathcal{G}$ induit un sous-réseau bayésien.

Un réseau bayésien définit une distribution

Conditions stochastiques

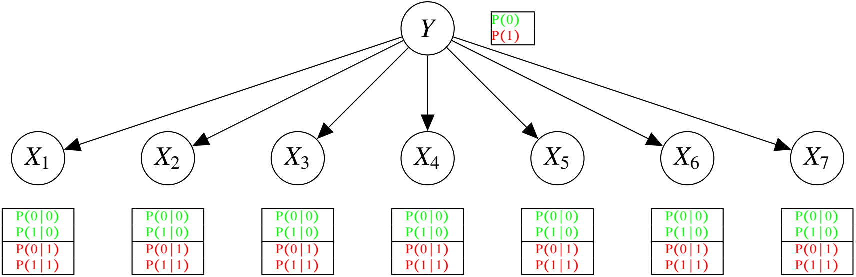
1. $\mathrm{P}(X_1, \ldots, X_N) > 0$ comme produit de facteurs positifs ;
2. $\mathrm{P}(X_1, \ldots, X_N)$ est normalisé :

$$
\begin{aligned}
\sum_{x_1,\ldots,x_n} \mathrm{P}(X_1 = x_1, \ldots, X_N = x_n) &= \sum_{x_1,\ldots,x_n} \prod_i \mathrm{P}(X_i = x_i \,|\, pa(X_i)) \\
&= \sum_{x_1,\ldots,x_{n-1}} \prod_i \mathrm{P}(X_i = x_i \,|\, pa(X_i)) \sum_{x_N} \mathrm{P}(X_N = x_N \,|\, pa(X_N)) \\
&= \sum_{x_1,\ldots,x_{n-1}} \prod_i \mathrm{P}(X_i = x_i \,|\, pa(X_i)) \underbrace{\sum_{x_N} \mathrm{P}(X_N = x_N \,|\, pa(X_N))}_{=1} \\
&= \sum_{x_1,\ldots,x_{n-1}} \prod_i \mathrm{P}(X_i = x_i \,|\, pa(X_i)) \\
&= \ldots \\
&= \sum_{x_1} \mathrm{P}(X_1 = x_1) = 1
\end{aligned}
$$

Sous-réseau

Un sous-graphe de $\mathcal{G}$ induit un sous-réseau bayésien.

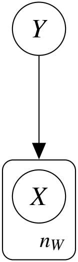
Retour à la classification de textes



$$\mathrm{P}(X_1, \ldots, X_{n_W}, Y) = \mathrm{P}(Y) \prod_{i=1}^{n_W} \mathrm{P}(X_i \mid Y)$$

Retour à la classification de textes

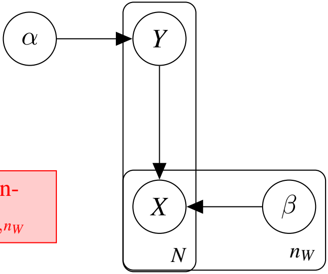
Représentation compacte (*plate notation*)



$$P(X_1, \ldots, X_{n_W}, Y) = P(Y) \prod_i P(X_i \mid Y)$$

Retour à la classification de textes

Vision bayésienne



n_W paramètres : $\beta_1 \ldots \beta_{n_W}$

N vecteurs de dimension n_W : $x_{1,1} \ldots x_{N,n_W}$

$$\mathrm{P}(\mathcal{C}, \alpha, \boldsymbol{\beta}) = \mathrm{P}(\alpha) \prod_{1}^{n_W} \mathrm{P}(\beta_i) \prod_{j=1}^{N} \mathrm{P}(Y_j \mid \alpha) \prod_{k=1}^{n_W} \mathrm{P}(X_{jk} \mid Y_j, \beta_k)$$

Question

$\boldsymbol{E}, \boldsymbol{F} \subset \boldsymbol{G}$ deux ensembles disjoints de variables ($\boldsymbol{F}$ potentiellement vide)
calculer la distribution $\mathrm{P}(\boldsymbol{E} \mid \boldsymbol{F})$

Réponse

$$\mathrm{P}(\boldsymbol{E} \mid \boldsymbol{F}) = \frac{\mathrm{P}(\boldsymbol{E}, \boldsymbol{F})}{\sum_{E} \mathrm{P}(\boldsymbol{E}, \boldsymbol{F})} = \frac{\sum_{\boldsymbol{G} \setminus (\boldsymbol{E} \cup \boldsymbol{F})} \mathrm{P}(\boldsymbol{G})}{\sum_{E} \mathrm{P}(\boldsymbol{E}, \boldsymbol{F})}$$

Problème : marginaliser (partiellement ou totalement) efficacement

$$\sum_{\boldsymbol{G} \setminus (\boldsymbol{E} \cup \boldsymbol{F})} \mathrm{P}(\boldsymbol{G}) = \sum_{\boldsymbol{G} \setminus (\boldsymbol{E} \cup \boldsymbol{F})} \prod_{i} \mathrm{P}(X_i \mid pa(X_i))$$

Inférence dans les réseaux bayésiens

Mécanisation du raisonnement probabiliste

Question

$\boldsymbol{E}, \boldsymbol{F} \subset \boldsymbol{G}$ deux ensembles disjoints de variables ($\boldsymbol{F}$ potentiellement vide)
calculer la distribution $\mathrm{P}(\boldsymbol{E} \mid \boldsymbol{F})$

Réponse

$$\begin{aligned}
\mathrm{P}(\boldsymbol{E} \mid \boldsymbol{F}) &= \frac{\mathrm{P}(\boldsymbol{E}, \boldsymbol{F})}{\sum_E \mathrm{P}(\boldsymbol{E}, \boldsymbol{F})} \\
&= \frac{\sum_{\boldsymbol{G} \setminus (\boldsymbol{E} \cup \boldsymbol{F})} \mathrm{P}(\boldsymbol{G})}{\sum_{\boldsymbol{E}} \mathrm{P}(\boldsymbol{E}, \boldsymbol{F})}
\end{aligned}$$

Problème : marginaliser (partiellement ou totalement) efficacement

$$\sum_{\boldsymbol{G} \setminus (\boldsymbol{E} \cup \boldsymbol{F})} \mathrm{P}(\boldsymbol{G}) = \sum_{\boldsymbol{G} \setminus (\boldsymbol{E} \cup \boldsymbol{F})} \prod_i \mathrm{P}(X_i \mid pa(X_i))$$

Inférence dans les réseaux bayésiens

Mécanisation du raisonnement probabiliste

Question

$\boldsymbol{E}, \boldsymbol{F} \subset \boldsymbol{G}$ deux ensembles disjoints de variables ($\boldsymbol{F}$ potentiellement vide)

calculer la distribution $\mathrm{P}(\boldsymbol{E} \mid \boldsymbol{F})$

Réponse

$$\begin{aligned} \mathrm{P}(\boldsymbol{E} \mid \boldsymbol{F}) &= \frac{\mathrm{P}(\boldsymbol{E}, \boldsymbol{F})}{\sum_E \mathrm{P}(\boldsymbol{E}, \boldsymbol{F})} \\ &= \frac{\sum_{\boldsymbol{G} \setminus (\boldsymbol{E} \cup \boldsymbol{F})} \mathrm{P}(\boldsymbol{G})}{\sum_{\boldsymbol{E}} \mathrm{P}(\boldsymbol{E}, \boldsymbol{F})} \end{aligned}$$

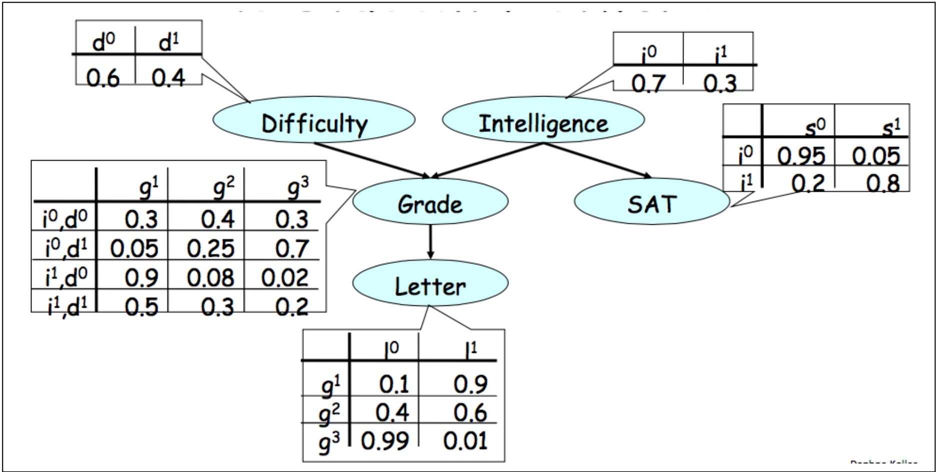
Problème : marginaliser (partiellement ou totalement) efficacement

$$\sum_{\boldsymbol{G} \setminus (\boldsymbol{E} \cup \boldsymbol{F})} \mathrm{P}(\boldsymbol{G}) = \sum_{\boldsymbol{G} \setminus (\boldsymbol{E} \cup \boldsymbol{F})} \prod_i \mathrm{P}(X_i \mid pa(X_i))$$

Raisonner sur un réseau exemplaire

d'après Daphne Koller

$$P(D, I, G, S, L) = P(D)\,P(I)\,P(G \mid D, I)\,P(S \mid I)\,P(L \mid G)$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

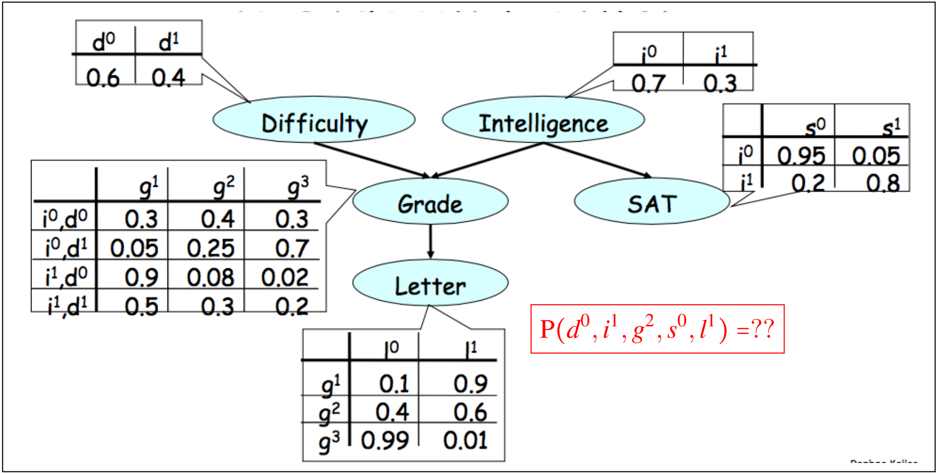
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Raisonner sur un réseau exemplaire

d’après Daphne Koller

$$P(D,I,G,S,L) = P(D)\,P(I)\,P(G \mid D,I)\,P(S \mid I)\,P(L \mid G)$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

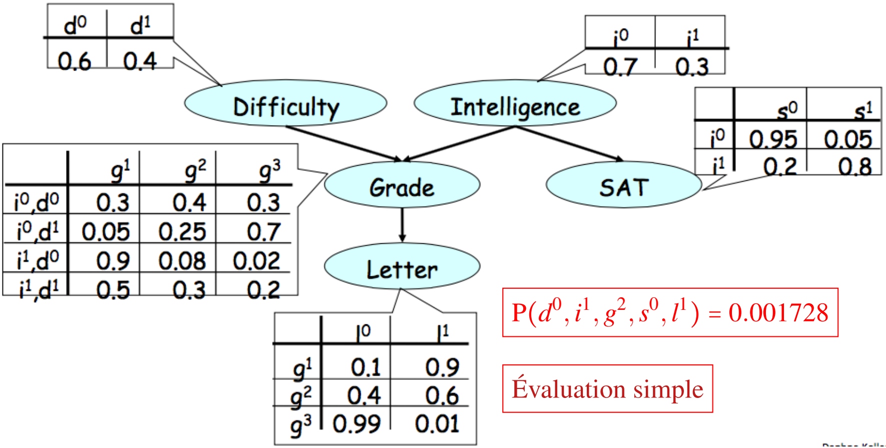
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

$$P(d^0, i^1, g^2, s^0, l^1) = ??$$

Raisonner sur un réseau exemplaire

d'après Daphne Koller

$$P(D,I,G,S,L) = P(D)\,P(I)\,P(G \mid D,I)\,P(S \mid I)\,P(L \mid G)$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

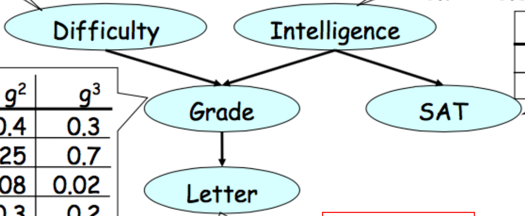
$P(d^0, i^1, g^2, s^0, l^1) = 0.001728$

Évaluation simple

Raisonner sur un réseau exemplaire

d'après Daphne Koller

$$P(D,I,G,S,L) = P(D)\,P(I)\,P(G \mid D,I)\,P(S \mid I)\,P(L \mid G)$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

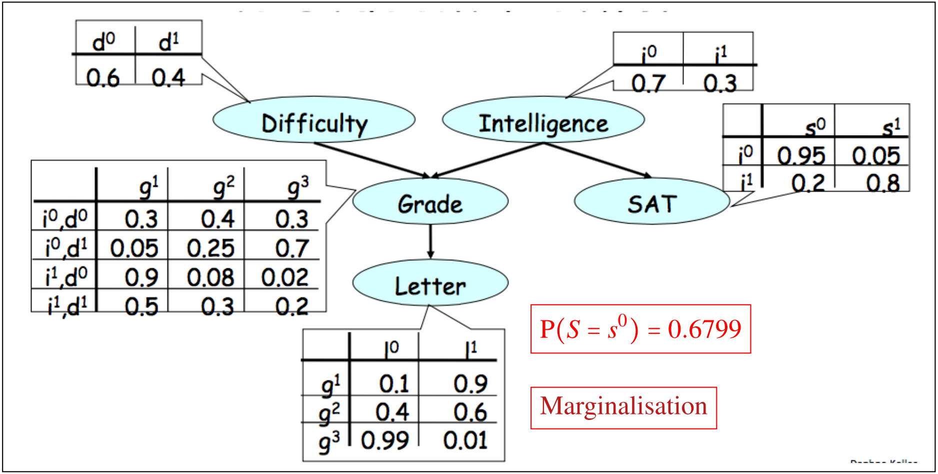
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

$$\boxed{P(S = s^0) = ?}$$

Raisonner sur un réseau exemplaire

d'après Daphne Koller

$$P(D, I, G, S, L) = P(D)\,P(I)\,P(G \mid D, I)\,P(S \mid I)\,P(L \mid G)$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

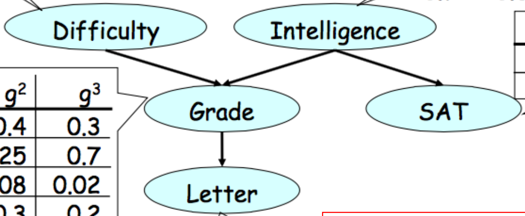
$P(S = s^0) = 0.6799$

Marginalisation

Raisonner sur un réseau exemplaire

d’après Daphne Koller

$$P(D, I, G, S, L) = P(D)\,P(I)\,P(G \mid D, I)\,P(S \mid I)\,P(L \mid G)$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

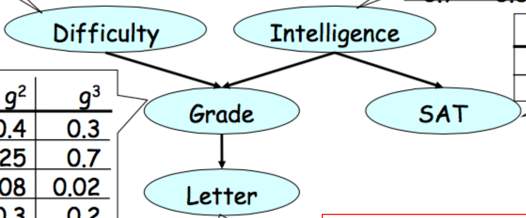
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

$$P(l^1 \mid g_2, d_1, i^0, s^1) = ??$$

Raisonner sur un réseau exemplaire

d’après Daphne Koller

$$P(D, I, G, S, L) = P(D)\,P(I)\,P(G \mid D, I)\,P(S \mid I)\,P(L \mid G)$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

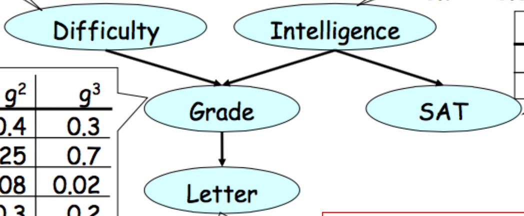
$$\boxed{P(l^1 \mid g_2, d_1, i^0, s^1) = 0.6}$$

Conditionnement

Raisonner sur un réseau exemplaire

d'après Daphne Koller

$$P(D, I, G, S, L) = P(D)\,P(I)\,P(G \mid D, I)\,P(S \mid I)\,P(L \mid G)$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

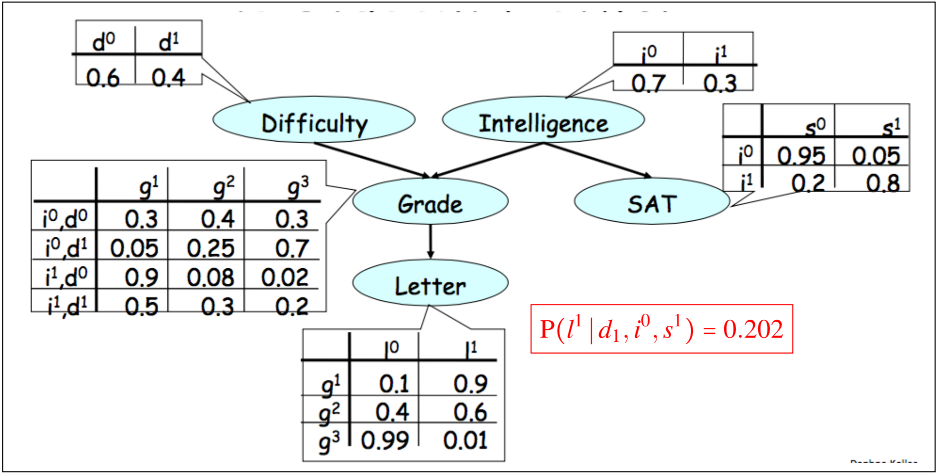
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

$$\boxed{P(l^1 \mid d_1, i^0, s^1) = ??}$$

Raisonner sur un réseau exemplaire

d’après Daphne Koller

$$P(D, I, G, S, L) = P(D)\, P(I)\, P(G \mid D, I)\, P(S \mid I)\, P(L \mid G)$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

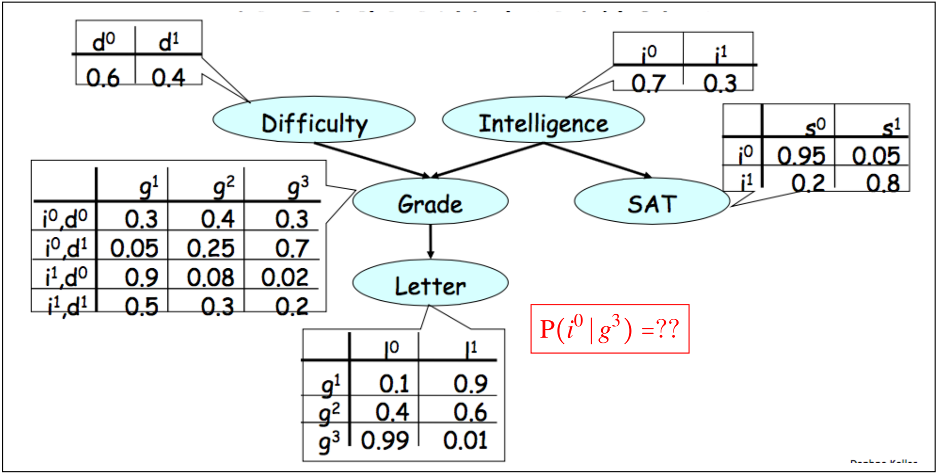
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

$$P(l^1 \mid d_1, i^0, s^1) = 0.202$$

Raisonner sur un réseau exemplaire

d’après Daphne Koller

$$P(D,I,G,S,L) = P(D)\,P(I)\,P(G\,|\,D,I)\,P(S\,|\,I)\,P(L\,|\,G)$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

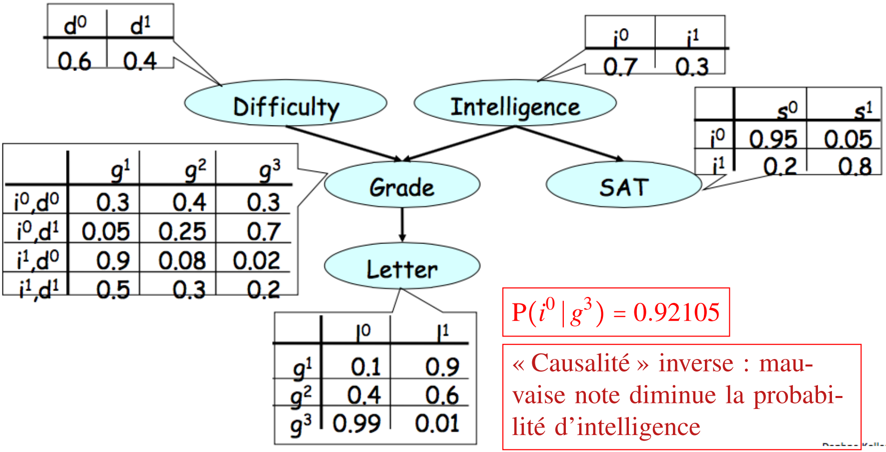
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

$$\boxed{P(i^0\,|\,g^3) = ??}$$

Raisonner sur un réseau exemplaire

d'après Daphne Koller

$$P(D,I,G,S,L) = P(D)\,P(I)\,P(G\,|\,D,I)\,P(S\,|\,I)\,P(L\,|\,G)$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

$P(i^0\,|\,g^3) = 0.92105$

« Causalité » inverse : mauvaise note diminue la probabilité d'intelligence

Indépendance conditionnelle

Indépendance marginale

$$X, Y, \text{ des VA}, X \perp\!\!\!\perp Y \Leftrightarrow \begin{aligned} \mathrm{P}(X, Y) &= \mathrm{P}(X)\,\mathrm{P}(Y) \\ \mathrm{P}(X \mid Y) &= \mathrm{P}(X) \\ \mathrm{P}(Y \mid X) &= \mathrm{P}(Y) \end{aligned}$$

Deux lectures :

- quantification universelle : $\forall x, y, \mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\,\mathrm{P}(Y = y)$
- vision factorielle : la matrice représentant $\mathrm{P}(X, Y)$ est un produit de deux vecteurs pour $\mathrm{P}(X)$ et $\mathrm{P}(Y)$

Indépendance conditionnelle

$$X, Y, Z \text{ des VA}, X \perp\!\!\!\perp Z \mid Y \Leftrightarrow \begin{aligned} \mathrm{P}(X \mid Y, Z) &= \mathrm{P}(X \mid Y) \\ \mathrm{P}(Z \mid X, Y) &= \mathrm{P}(Z \mid Y) \\ \mathrm{P}(X, Z \mid Y) &= \mathrm{P}(X \mid Y)\,\mathrm{P}(Z \mid Y) \end{aligned}$$

Indépendance conditionnelle

Indépendance marginale

$$X, Y, \text{ des VA}, X \perp\!\!\!\perp Y \Leftrightarrow \begin{aligned} \mathrm{P}(X, Y) &= \mathrm{P}(X)\,\mathrm{P}(Y) \\ \mathrm{P}(X \mid Y) &= \mathrm{P}(X) \\ \mathrm{P}(Y \mid X) &= \mathrm{P}(Y) \end{aligned}$$

Indépendance conditionnelle

$$X, Y, Z \text{ des VA}, X \perp\!\!\!\perp Z \mid Y \Leftrightarrow \begin{aligned} \mathrm{P}(X \mid Y, Z) &= \mathrm{P}(X \mid Y) \\ \mathrm{P}(Z \mid X, Y) &= \mathrm{P}(Z \mid Y) \\ \mathrm{P}(X, Z \mid Y) &= \mathrm{P}(X \mid Y)\,\mathrm{P}(Z \mid Y) \end{aligned}$$

Généralisation à des ensembles de variables $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$.

Indépendance conditionnelle

Indépendance marginale

$$X, Y, \text{ des VA}, X \perp\!\!\!\perp Y \Leftrightarrow \begin{aligned} \mathrm{P}(X, Y) &= \mathrm{P}(X)\,\mathrm{P}(Y) \\ \mathrm{P}(X \mid Y) &= \mathrm{P}(X) \\ \mathrm{P}(Y \mid X) &= \mathrm{P}(Y) \end{aligned}$$

Indépendance conditionnelle

$$X, Y, Z \text{ des VA}, X \perp\!\!\!\perp Z \mid Y \Leftrightarrow \begin{aligned} \mathrm{P}(X \mid Y, Z) &= \mathrm{P}(X \mid Y) \\ \mathrm{P}(Z \mid X, Y) &= \mathrm{P}(Z \mid Y) \\ \mathrm{P}(X, Z \mid Y) &= \mathrm{P}(X \mid Y)\,\mathrm{P}(Z \mid Y) \end{aligned}$$

Généralisation à des ensembles de variables $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$.

Modèles graphiques, indépendances conditionnelles

Factorisation canonique et factorisation graphique

$$P(X_1 \dots X_N) = \underbrace{P(X_1) \prod_i P(X_i \,|\, \{X_j, j < i\})}_{\text{toujours vrai}} = \underbrace{P(X_1) \prod_i P(X_i \,|\, \{pa(X_i)\})}_{\text{sous condition}}$$

Implications : indépendance conditionnelle

1. soit P(), tq. $\forall i \, P(X_i \,|\, \{X_j, j < i\}) = P(X_i \,|\, pa(X_i))$ alors P() se factorise selon $\mathcal{G}$
2. soit $pv(X_i) = \{X_j, j < i, X_J \notin pa(X_i)\}$, si $\forall i, X_i \perp\!\!\!\perp pv(X_i) | pa(X_i)$, alors P() se factorise selon $\mathcal{G}$

Quelles sont les relations d'indépendance conditionnelle exprimées par un NB ?
Comment les identifier automatiquement ?

Sémantique des arcs d'un RB

Proposition

Soit $\mathcal{G}$ un réseau bayésien, soit $pa(X_i)$ les parents de X_i et $nd(X_i)$ l'ensemble des nœuds qui ne sont pas des descendants de X_i, alors les énoncés suivants sont équivalents :

1. $\forall i, X_i \perp\!\!\!\perp nd(X_i) \,|\, pa(X_i)$
2. $\mathrm{P}(X)$ se factorise $\mathrm{P}(X) = \prod_i \mathrm{P}(X_i \,|\, pa(X_i))$

Sémantique des arcs d'un RB

Proposition

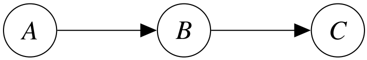
Soit $\mathcal{G}$ un réseau bayésien, soit $pa(X_i)$ les parents de X_i et $nd(X_i)$ l'ensemble des nœuds qui ne sont pas des descendants de X_i, alors les énoncés suivants sont équivalents :

1. $\forall i, X_i \perp\!\!\!\perp nd(X_i) \,|\, pa(X_i)$
2. $P(X)$ se factorise $P(X) = \prod_i P(X_i \,|\, pa(X_i))$

Preuve 2 implique 1 : (par récurrence sur i, on montre que $P(X_i \,|\, X_1 \ldots X_{i-1}) = P(X_I \,|\, pa(X_i))$, ce qui entraine que : $\forall i, X_i \perp\!\!\!\perp \{X_1 \ldots X_{i-1}\} \setminus pa(X_i) \,|\, pa(X_i)$.
Si X_j est non-descendant de X_i alors il existe une numérotation topologique qui ordonne j avant i.

Y-a-t-il d'autres dépendances dans G ? Comment les identifier automatiquement ?

3 configurations ternaires



Chaine linéaire (*tail-to-head*) : $\mathrm{P}(A, B, C) = \mathrm{P}(A)\,\mathrm{P}(B \mid A)\,\mathrm{P}(C \mid B)$

$$\mathrm{P}(A, C \mid B) = \frac{\mathrm{P}(A)\,\mathrm{P}(B \mid A)\,\mathrm{P}(C \mid B)}{\mathrm{P}(B)} = \mathrm{P}(A \mid B)\,\mathrm{P}(C \mid B)$$

$\Rightarrow$ Indépendance conditionnelle de A et de C

3 configurations ternaires

$$A \leftarrow B \rightarrow C$$

$$\textit{tail-to-tail} : \mathrm{P}(A, B, C) = \mathrm{P}(B)\,\mathrm{P}(A \mid B)\,\mathrm{P}(C \mid B)$$

$$\mathrm{P}(A, C \mid B) = \frac{\mathrm{P}(B)\,\mathrm{P}(A \mid B)\,\mathrm{P}(C \mid B)}{\mathrm{P}(B)} = \mathrm{P}(A \mid B)\,\mathrm{P}(C \mid B)$$

$\Rightarrow$ Indépendance conditionnelle de A et de C

3 configurations ternaires

$$A \longrightarrow B \longleftarrow C$$

$$\textit{head-to-head} : \mathrm{P}(A, B, C) = \mathrm{P}(A)\,\mathrm{P}(C)\,\mathrm{P}(B \mid A, C)$$

$$\mathrm{P}(A, C \mid B) = \frac{\mathrm{P}(A)\,\mathrm{P}(C)\,\mathrm{P}(B \mid A, C)}{\mathrm{P}(B)} \neq \mathrm{P}(A \mid B)\,\mathrm{P}(C \mid B)$$

$\Rightarrow$ pas d'indépendance conditionnelle de A et de C

Anti-causalité : explaining « *Explaining away* »

$$A \leftarrow B \rightarrow C$$

Connaitre A modifie la connaissance sur C (et réciproquement).

Application (Bishop, p377)

A, B, C sont binaires, avec : $\mathrm{P}(A = 1) = \mathrm{P}(C = 1) = 0.9$ et $\mathrm{P}(B = 1 \mid A, C)$ défini par :

$\mathrm{P}(B = 1 \mid A = 1, C = 1) = 0.8$	$\mathrm{P}(B = 1 \mid A = 1, C = 0) = 0.2$
$\mathrm{P}(B = 1 \mid A = 0, C = 1) = 0.2$	$\mathrm{P}(B = 1 \mid A = 0, C = 0) = 0.1$

Calculer :

- $\mathrm{P}(A = 0 \mid B = 1)$
- $\mathrm{P}(A = 0 \mid B = 1, C = 1)$

Savoir $C = 1$ rend $A = 0$ moins probable !

Anti-causalité : explaining « *Explaining away* »

$$A \longleftarrow B \longrightarrow C$$

Connaitre A modifie la connaissance sur C (et réciproquement).

Application (Bishop, p377)

A, B, C sont binaires, avec : $\mathrm{P}(A = 1) = \mathrm{P}(C = 1) = 0.9$ et $\mathrm{P}(B = 1 \,|\, A, C)$ défini par :

$\mathrm{P}(B = 1 \,\vert\, A = 1, C = 1) = 0.8$	$\mathrm{P}(B = 1 \,\vert\, A = 1, C = 0) = 0.2$
$\mathrm{P}(B = 1 \,\vert\, A = 0, C = 1) = 0.2$	$\mathrm{P}(B = 1 \,\vert\, A = 0, C = 0) = 0.1$

Calculer :

- $\mathrm{P}(A = 0 \,|\, B = 1)$
- $\mathrm{P}(A = 0 \,|\, B = 1, C = 1)$

Savoir $C = 1$ rend $A = 0$ moins probable !

Généralisation : le concept de d-séparation

Evaluer l'indépendance conditionnelle

Chemins séparés et d-séparation

Un chemin π entre les sommets A et B est bloqué (D-séparé) (sachant l'ensemble de sommets $\boldsymbol{C}$, disjoint de $\boldsymbol{A}$ et $\boldsymbol{B}$) s'il contient un sommet tel que soit :

1. les arcs rencontrent ce sommet « head-to-tailfg ou « tail-to-head » et le sommet est dans C, ou bien
2. les arcs rencontrent le sommet « head-to-head », et ni le sommet, ni aucun de ses dépendants, n'est dans $\boldsymbol{C}$

Theorem (Théorème)

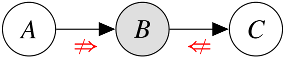
Si tous les chemins entre un ensemble $\mathbf{A}$ *et un ensemble* $\mathbf{B}$ *sont bloqués (relativement à* $\mathbf{C}$*) alors :*

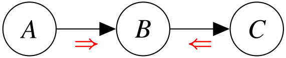
$$\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C}$$
$$\mathrm{P}(\boldsymbol{A}, \boldsymbol{B} \mid \boldsymbol{C}) = \mathrm{P}(\boldsymbol{A} \mid \boldsymbol{C})\, \mathrm{P}(\boldsymbol{B} \mid \boldsymbol{C})$$

Les règles du « Bayes-Ball »

La circulation de l'influence

$X \perp\!\!\!\perp Y | \mathcal{Z}$ si aucune balle issue du nœud X n'atteint Y lorsque les nœuds de $\mathcal{Z}$ sont connus et que s'appliquent les règles suivantes ($\Leftarrow$: passe ; $\not\Leftarrow$: bloque).

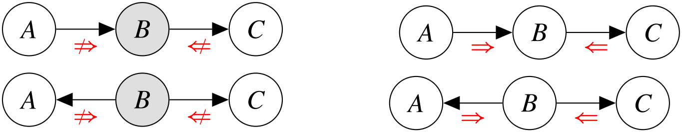




Les règles du « Bayes-Ball »

La circulation de l'influence

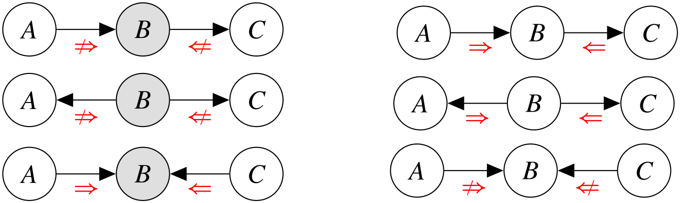
$X \perp\!\!\!\perp Y | \mathcal{Z}$ si aucune balle issue du nœud X n'atteint Y lorsque les nœuds de $\mathcal{Z}$ sont connus et que s'appliquent les règles suivantes ($\Leftarrow$: passe ; $\not\Leftarrow$: bloque).



Les règles du « Bayes-Ball »

La circulation de l'influence

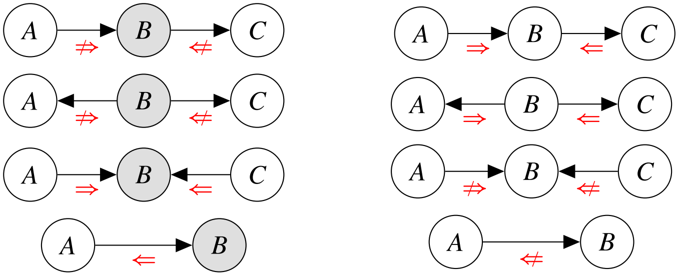
$X \perp\!\!\!\perp Y | \mathcal{Z}$ si aucune balle issue du nœud X n'atteint Y lorsque les nœuds de $\mathcal{Z}$ sont connus et que s'appliquent les règles suivantes ($\Leftarrow$: passe ; $\nLeftarrow$: bloque).



Les règles du « Bayes-Ball »

La circulation de l'influence

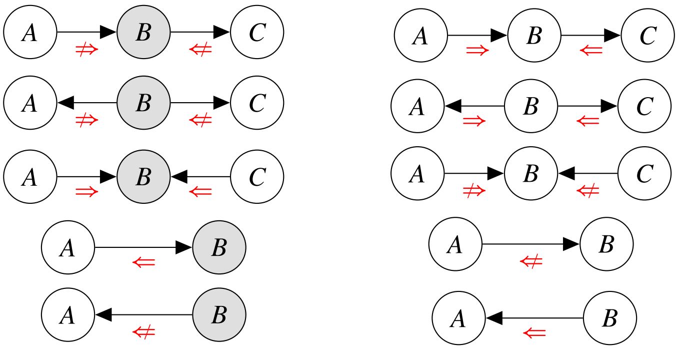
$X \perp\!\!\!\perp Y | \mathcal{Z}$ si aucune balle issue du nœud X n'atteint Y lorsque les nœuds de $\mathcal{Z}$ sont connus et que s'appliquent les règles suivantes ($\Leftarrow$: passe ; $\not\Leftarrow$: bloque).



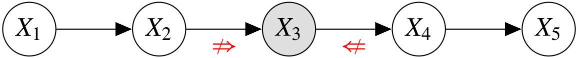
Les règles du « Bayes-Ball »

La circulation de l'influence

$X \perp\!\!\!\perp Y | \mathcal{Z}$ si aucune balle issue du nœud X n'atteint Y lorsque les nœuds de $\mathcal{Z}$ sont connus et que s'appliquent les règles suivantes ($\Leftarrow$: passe ; $\not\Leftarrow$: bloque).

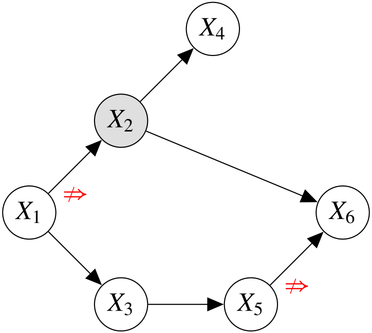


Jouons au « Bayes-Ball »



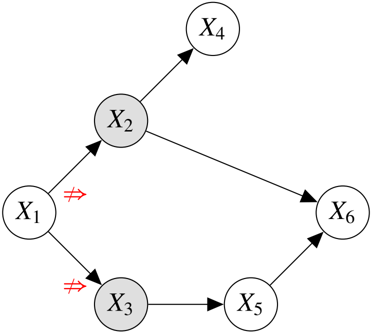
$\{X_1, X_2\} \perp\!\!\!\perp \{X_4, X_5\} \,|\, X_3$ (standard Markov)

Jouons au « Bayes-Ball »



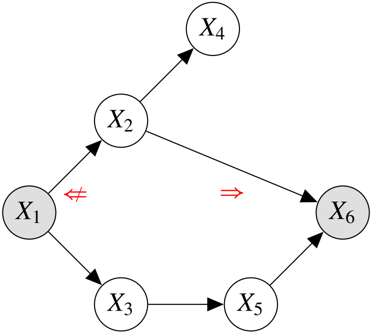
$$\{X_1, X_3, X_5\} \perp\!\!\!\perp \{X_4\} \,|\, X_2$$

Jouons au « Bayes-Ball »



$$\{X_1\} \perp\!\!\!\perp \{X_5\} \,|\, X_3$$

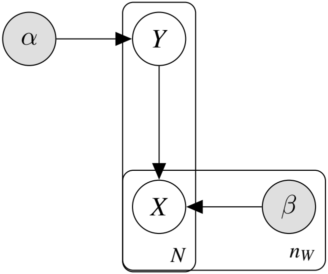
Jouons au « Bayes-Ball »



$$\wedge \{X_2\} \perp\!\!\!\perp \{X_3\} \mid \{X_1, X_6\}$$

Bayésien naïf Bernoulli

Qui dépend de quoi ?

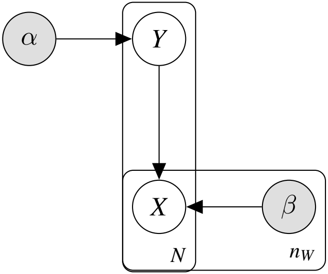


$X_{ik} \perp\!\!\!\perp X_{jk} \,|\, \{\alpha, \beta_k\}$

attention : on n'a pas $X_{ik} \perp\!\!\!\perp X_{jk}$!

Bayésien naïf Bernoulli

Qui dépend de quoi ?



$X_{ik} \perp\!\!\!\perp X_{jk} \mid \{\alpha, \beta_k\}$

attention : on n'a pas $X_{ik} \perp\!\!\!\perp X_{jk}$!