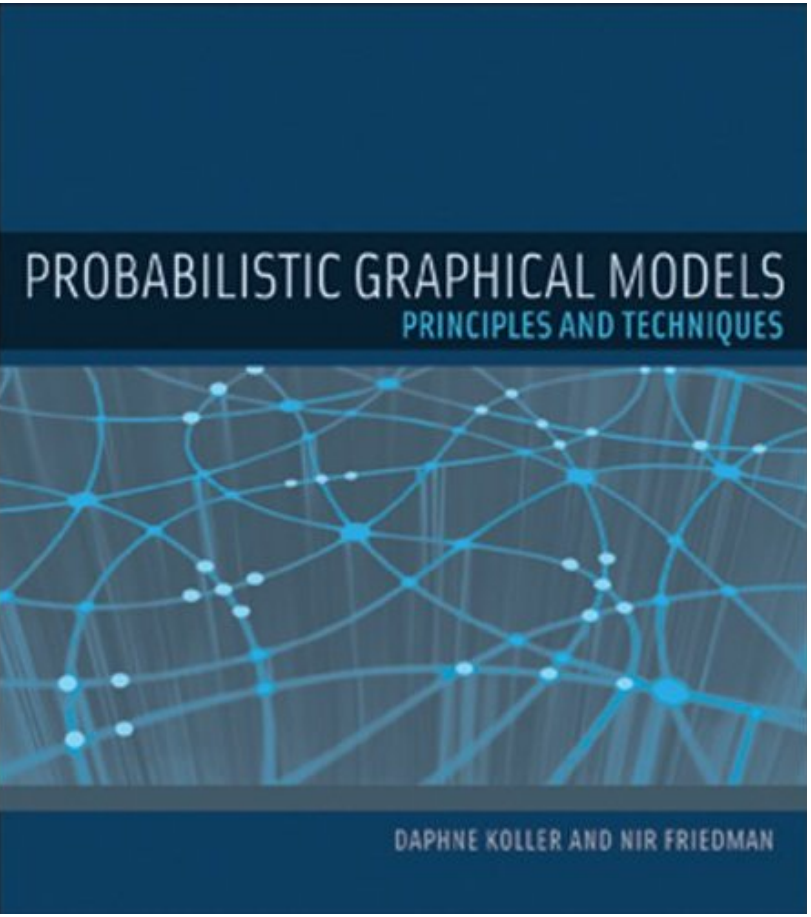
Introduction aux modèles graphiques probabilistes

Philippe LERAY
philippe.leray@univ-nantes.fr

Equipe COnnaissances et Décision – LINA – UMR 6241
Site de l'Ecole Polytechnique de l'Université de Nantes

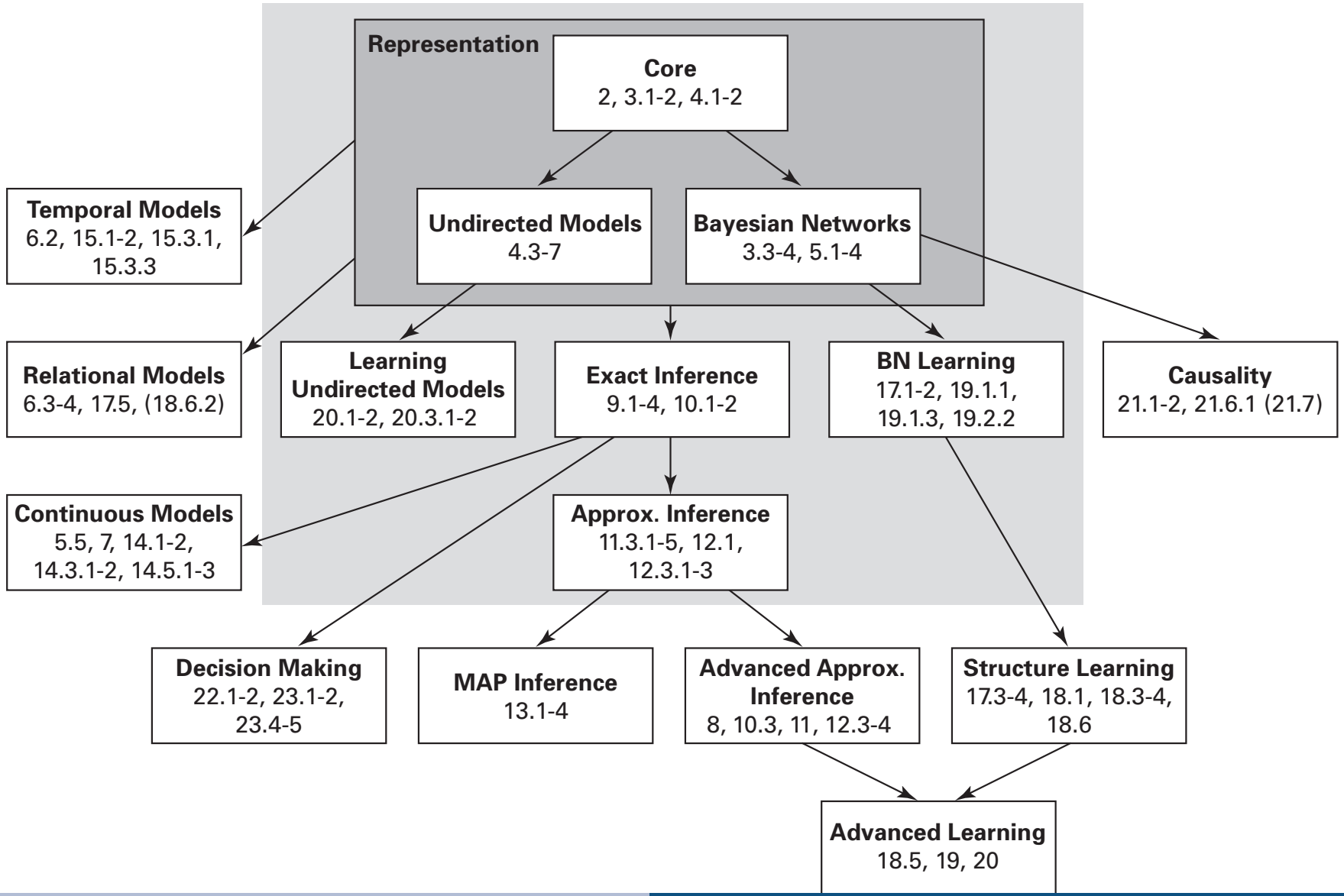
Introduction



Un domaine vaste

- Présentation et figures inspirées de [Koller & Friedman 09]
- $\simeq 1200p.$ à résumer en mois d'une heure :-)

Un domaine vaste ... suite



Plan

- Rappels : Probabilités et Graphes
- 3 étapes ...
  1. représentation
  2. inférence
  3. apprentissage
- ... pour 3 familles de PGM
  1. graphes dirigés : réseaux bayésiens
  2. graphes non dirigés : réseaux de Markov (MRF)
  3. graphes partiellement dirigés : chain graphs

Rappels Probabilités

Indépendance

- A et B sont indépendants ssi :
  $P(A, B) = P(A) \times P(B)$
  $P(A|B) = P(A)$
  $P(B|A) = P(B)$

Indépendance conditionnelle

- A et B sont indépendants conditionnellement à C ssi :
  $P(A|B, C) = P(A|C)$

Rappels Graphes

Terminologie

- Un graphe = un ensemble de nœuds et d'arêtes
- Graphes orientés (dirigés), non dirigés, partiellement dirigés
- Graphes orientés sans circuit

Principe des PGM

Représentation des connaissances

- Un graphe comme modèle d'indépendance

Raisonnement

- Des algorithmes d'inférence probabiliste tirant partie de la structure graphique du modèle

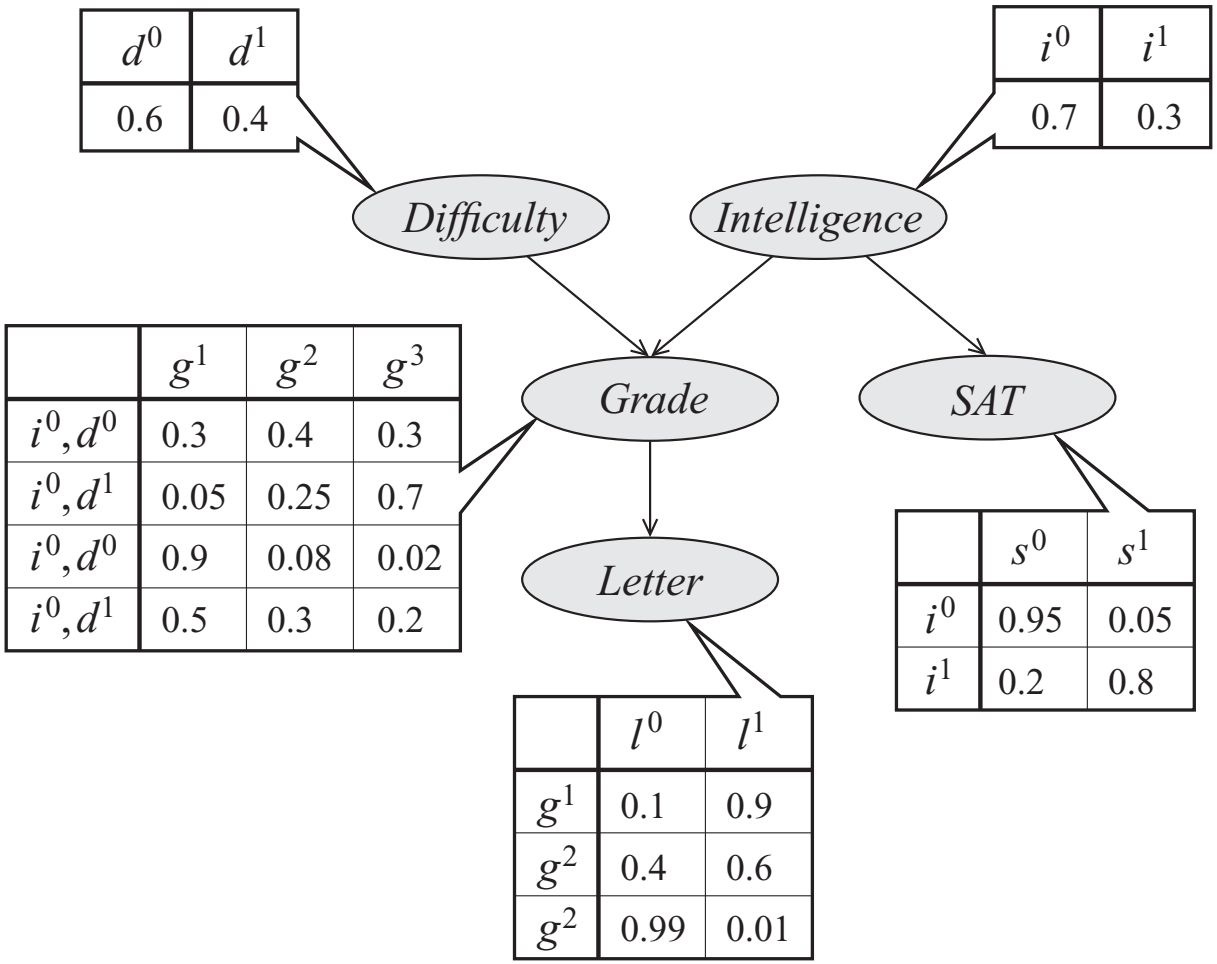
Construction

- Des connaissances a priori pouvant déterminer tout ou partie de la structure graphique
- Des algorithmes d'apprentissage déterminant le reste du modèle à partir de données

Plan

- Rappels : Probabilités et Graphes
- 3 étapes ...
  1. représentation
  2. inférence
  3. apprentissage
- ... pour 3 familles de PGM
  1. graphes dirigés : réseaux bayésiens
  2. graphes non dirigés : réseaux de Markov (MRF)
  3. graphes partiellement dirigés : chain graphs

les réseaux bayésiens [Pearl 88]



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^0, d^0	0.9	0.08	0.02
i^0, 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^2	0.99	0.01

RB comme modèles d'indépendance

La dépendance est symétrique, alors pourquoi utiliser un graphe orienté ?

Exemple avec 3 nœuds, et 3 structures simples

- $A \rightarrow C \rightarrow B$: connexion série
  - A et B sont dépendants, mais indépendants conditionnement à C
- $A \leftarrow C \rightarrow B$: connexion divergente
  - pareil
- $A \rightarrow C \leftarrow B$: connexion convergente (V-structure)
  - **A et B sont indépendants, mais dépendants conditionnement à C**

Factorisation de la loi jointe

Avantage

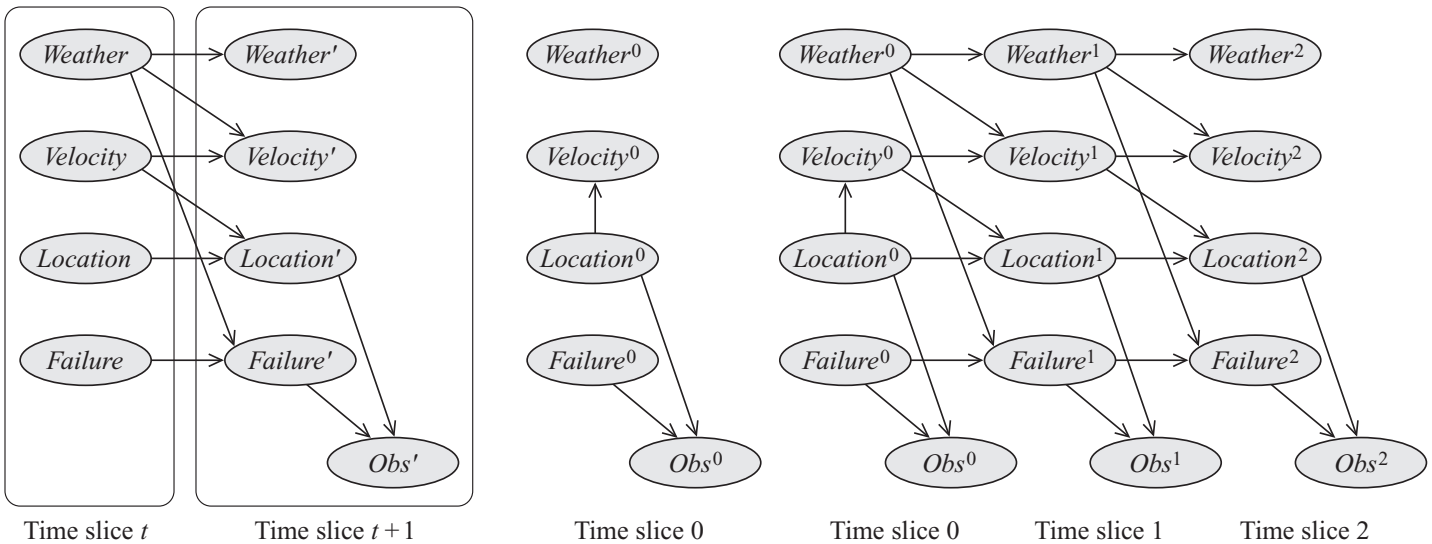
- Décomposition de la loi jointe (globale) en un produit de distributions conditionnelles locales

$$P(S) = \Pi_{i=1}^{n} P(X_i | parents(X_i))$$

Des extensions

pour de nombreux problèmes

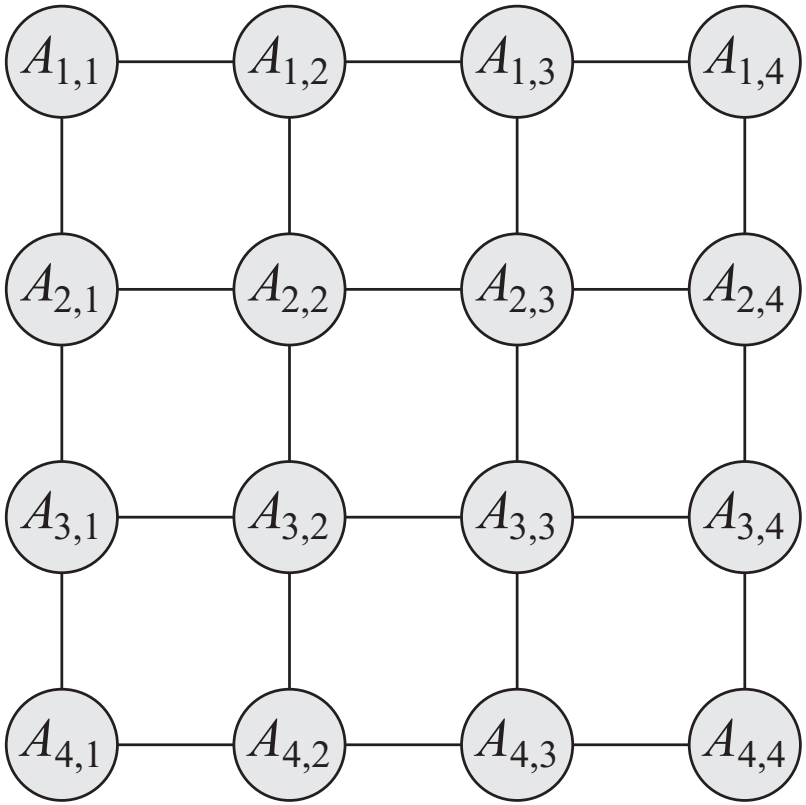
- Causalité : RB causal
- Variables continues : RB gaussien, hybride (CG)
- Temporalité : **RB temporel** , HMM, Filtre de Kalman
- Décision : Diagramme d'influence
- Classification : Naive Bayes, multinets, ...



Plan

- Rappels : Probabilités et Graphes
- 3 étapes ...
  1. représentation
  2. inférence
  3. apprentissage
- ... pour 3 familles de PGM
  1. graphes dirigés : réseaux bayésiens
  2. graphes non dirigés : réseaux de Markov (MRF)
  3. graphes partiellement dirigés : chain graphs

les MRF ...[Kindermann & Snell 80]



Factorisation de la loi jointe

Avantage

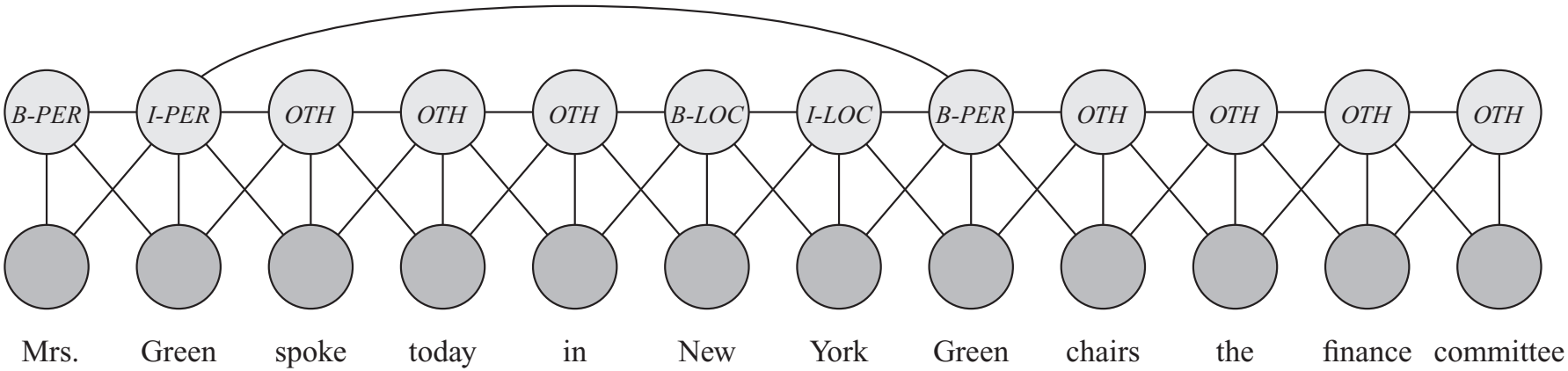
- Décomposition de la loi jointe (globale) en un produit de potentiels locaux
- Z constante de normalisation globale

$$P(S) = \frac{1}{Z}\Pi_{c=1}^{n_c}\phi(X_c)$$

Des extensions

pour de nombreux problèmes

- Des structures "historiques" : modèle d'Ising, machine de Boltzmann
- + Var. latentes : Deep Belief Networks
- Variables continues : Gaussian MRF
- Temporalité : Dynamic MRF
- Classification : **Conditional Random Field**



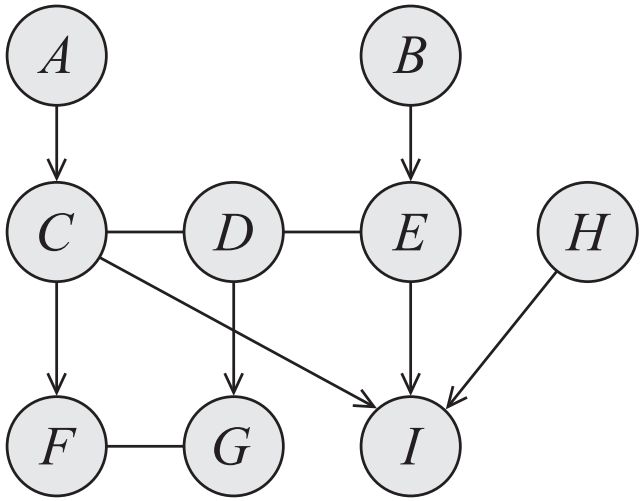
Plan

- Rappels : Probabilités et Graphes
- 3 étapes ...
  1. représentation
  2. inférence
  3. apprentissage
- ... pour 3 familles de PGM
  1. graphes dirigés : réseaux bayésiens
  2. graphes non dirigés : réseaux de Markov (MRF)
  3. graphes partiellement dirigés : chain graphs

les chains graphs ...[Lauritzen 96]

Modèle partiellement dirigé

- représentation de la loi jointe par un produit de facteurs "conditionnels"



Plan

- Rappels : Probabilités et Graphes
- 3 étapes ...
  1. représentation
  2. inférence
  3. apprentissage
- ... pour 3 familles de PGM
  1. graphes dirigés : réseaux bayésiens
  2. graphes non dirigés : réseaux de Markov (MRF)
  3. graphes partiellement dirigés : chain graphs

Inférence

$P(X|E)$?

RB, MRF, ... même combat

- problème NP-difficile
- heureusement, c'est dans le pire des cas
- pour des problèmes réels, il existe des algorithmes efficaces

inférence exacte

- élimination de variables
- conditionnement
- arbre de jonction

inférence approchée

- simulation : MCMC, filtrage particulaire, ...
- approximations variationnelles : Mean field, ...

Exemple : arbre de jonction

Principe

- convertir le PGM en un arbre de jonction de cliques
- faire circuler des messages dans cet arbre

A noter

- généralisation d'un "vieux" principe
  - HMM : forward-backward [Rabiner 89]
  - BN Polyarbres : Message Passing [Pearl 88]
- complexité : exponentielle par rapport à la taille des cliques

Plan

- Rappels : Probabilités et Graphes
- 3 étapes ...
  1. représentation
  2. inférence
  3. apprentissage
- ... pour 3 familles de PGM
  1. graphes dirigés : réseaux bayésiens
  2. graphes non dirigés : réseaux de Markov (MRF)
  3. graphes partiellement dirigés : chain graphs

Apprentissage : deux "philosophies"

Trouver le modèle optimal qui ...

Apprentissage génératif

- approche le mieux $P(X, Y)$
- pas de variable cible

Apprentissage génératif

- modèle plus général $\Rightarrow$ biais
- meilleur traitement des données incomplètes

Apprentissage discriminant

- approche le mieux $P(Y|X)$
- une variable cible Y privilégiée

Apprentissage génératif

- modèle plus spécifique
- meilleurs résultats si données importantes

Taxonomie des tâches d'apprentissage

MGP = un graphe et des paramètres

- apprentissage des paramètres / structure donnée
- apprentissage de la structure

... à partir de données

- données complètes
- données incomplètes
- variables latentes ?

Plan

- Rappels : Probabilités et Graphes
- 3 étapes ...
  1. représentation
  2. inférence
  3. apprentissage
- ... pour 3 familles de PGM
  1. graphes dirigés : réseaux bayésiens
  2. graphes non dirigés : réseaux de Markov (MRF)
  3. graphes partiellement dirigés : chain graphs

App. génératif et RB

Estimation de paramètres — Données complètes $\mathcal{D}$

- Approche statistique classique = *max. de vraisemblance (MV)*

$$\hat{\theta}^{MV} = \text{argmax } P(\mathcal{D}|\theta)$$

- Probabilité d'un événement = fréquence d'apparition de l'événement

Maximum de vraisemblance (MV)

$$\hat{P}(X_i = x_k | Pa(X_i) = x_j) = \hat{\theta}^{MV}_{i,j,k} = \frac{N_{i,j,k}}{\sum_k N_{i,j,k}}$$

$N_{i,j,k}$ = nb d'occurences de $\{X_i = x_k \text{ et } Pa(X_i) = x_j\}$

Apprentissage (données complètes)

Autre approche

- Approche bayésienne = *max. à posteriori (MAP)*

$$\hat{\theta}^{MAP} = \text{argmax } P(\theta|\mathcal{D}) = \text{argmax } P(\mathcal{D}|\theta)P(\theta)$$

- besoin d'une loi a priori sur les paramètres $P(\theta)$
- souvent distribution *conjuguée* à la loi de X
- si $P(X)$ multinomiale, $P(\theta)$ conjuguée = Dirichlet :

$$P(\theta) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} (\theta_{i,j,k})^{\alpha_{i,j,k}-1}$$

où $\alpha_{i,j,k}$ sont les cœfficients de la distribution de Dirichlet associée au coefficient $\theta_{i,j,k}$

Apprentissage (données complètes)

Maximum a Posteriori (MAP)

$$\hat{P}(X_i = x_k | Pa(X_i) = x_j) = \hat{\theta}^{MAP}_{i,j,k} = \frac{N_{i,j,k} + \alpha_{i,j,k} - 1}{\sum_k (N_{i,j,k} + \alpha_{i,j,k} - 1)}$$

Autre approche bayésienne

- *espérance à posteriori (EAP)* : calculer l'espérance a posteriori de $\theta_{i,j,k}$ au lieu du max.

$$\hat{P}(X_i = x_k | Pa(X_i) = x_j) = \hat{\theta}^{EAP}_{i,j,k} = \frac{N_{i,j,k} + \alpha_{i,j,k}}{\sum_k (N_{i,j,k} + \alpha_{i,j,k})}$$

Algorithme Expectation Maximisation

Apprentissage avec données incomplètes

- Principe très général [Dempster 77]

Principe

- Algorithme itératif
  - initialiser les paramètres $\theta^{(0)}$
  - E estimer la distribution des valeurs manquantes à partir des paramètres actuels $\theta^{(t)}$
    - = calculer $P(X_{\text{manquant}}|X_{\text{mesurés}})$ dans le RB actuel
    - = faire des inférences
  - M ré-estimer les paramètres $\theta^{(t+1)}$ à partir des données complétées
    - en utilisant MV, MAP, ou EAP

Génératif ou discriminant ?

apprentissage (génératif) des paramètres des RB

- données complètes
  - forme close calculable en une itération (MV, MAP, EAP)
- données incomplètes
  - algorithme itératif (EM), optimum local

apprentissage (discriminant) des paramètres des RB

- données complètes
  - algorithme itératif de type descente de gradient
- données incomplètes
  - algorithme "doublement" itératif (EM), optimum local

Et la structure ?

Deux problèmes :

Taille de l'espace de recherche

- le nombre de structures possibles à partir de n nœuds est super-exponentiel [Robinson 77]

$$NS(5) = 29281 \qquad NS(10) = 4.2 \times 10^{18}$$

Identifiabilité

- Les données reflètent la loi jointe et ses dépendances / indépendances entre variables
- Equivalence de Markov : plusieurs graphes peuvent représenter un même modèle d'indépendance
- Suffisance causale : et s'il y avait des variables latentes ?

Algorithmes existants

Apprentissage de la structure - données complètes

1. Recherche d'indépendances conditionnelles dans les données
2. Méthodes d'optimisation d'une fonction de score
   avantage : score décomposable localement
3. Méthodes hybrides de recherche de voisinage locale + optimisation globale

et ensuite ?

- données incomplètes
  - EM dans l'espace des structures (SEM) [Friedman 97]
- variables latentes
  - heuristiques de découverte + recherche gloutonne pour fixer leur cardinalité

Plan

- Rappels : Probabilités et Graphes
- 3 étapes ...
  1. représentation
  2. inférence
  3. apprentissage
- ... pour 3 familles de PGM
  1. graphes dirigés : réseaux bayésiens
  2. graphes non dirigés : réseaux de Markov (MRF)
  3. graphes partiellement dirigés : chain graphs

Et là, ca se complique ...

Apprentissage des paramètres, données complètes

RB

- $P(S) = \Pi_i P(X_i|pa(X_i))$
- chaque terme est une distribution de probabilité estimable séparément

MRF

- $P(S) = \frac{1}{Z}\Pi_c \phi(X_c)$
- la constante Z globale empêche l'estimation locale

Seule une classe de MRF (MRF cordaux) équivalente aux RB s'apprend aussi facilement que les RB.

App. génératif et MRF

Estimation de paramètres — Données complètes $\mathcal{D}$

- la fonction log-vraisemblance est unimodale
- problème : pas de forme close du maximum pour les MRF

$\Rightarrow$ descente de gradient et convergence vers optimum global

- problème : le calcul du gradient nécessite une étape d'inférence dans le réseau
- possibilité d'utiliser des méthodes d'inférence approchées ou d'utiliser une approximation de la vraisemblance plus sympathique (pseudo-likelihood, marge ...)

Et les données incomplètes ?

- perte de la concavité du log-vraisemblance
- utilisation possible d'EM mais convergence locale (idem. RB)

App. discriminant et CRF

Et dans le cas discriminant **Données complètes $\mathcal{D}$**

- la fonction log-vraisemblance conditionnelle est aussi unimodale
- par contre, le conditionnement par rapport à la variable cible nécessite plusieurs étapes d'inférence dans le réseau

\- plus d'étapes d'inférence

\+ inférences avec conditionnement sur $Y \Rightarrow$ calculs plus simples

Et la structure ?

Apprentissage de la structure - données complètes

1. Recherche d'indépendances conditionnelles dans les données
   - plus simple que pour les RB, car les indépendances se traduisent plus simplement en terme graphique
   - même problème de fiabilité du test / taille des données
2. Méthodes d'optimisation d'une fonction de score
   - problème : les scores sont basés sur la vraisemblance donc calculables plus difficilement et ne sont plus décomposables
   - nécessité d'approcher l'impact (variation de score) des opérateurs classiques permettant de parcourir l'espace des MRF

Pour conclure ...

Domaine vaste ... très vaste

- principes généraux
- spécificités liées à la nature de ces modèles
- peu de références indiquées

$\Rightarrow$ un bon point de départ = [Koller & Friedman 09]

Ce n'est qu'une introduction ... à suivre :

- des modèles spécifiques (MRF, CRF, Deep BN ...)
- appliqués à vos domaines d'intérêt :-)

Des questions ?