MLWG Session 2

Apprentissage Bayésien

Pierre Dangauthier

18 mai 2005

PLAN

- Apprentissage Statistique non supervisé
  - Probabilités subjectives
  - Règles de calcul
  - Exemple
- Modèle génératif / discriminatif
- Le Bayésien chez E-Motion
- Références

3 types d'apprentissage

Soient des entrées sensorielles x1,x2,x3,x4 …

- **Apprentissage supervisé**:
  - La machine reçoit aussi des sorties *désirées* y1; y2; …
  - But: apprendre à produire la sortie correcte à partir d'une nouvelle entrée (interpolation, réseaux de neurone)
- **Apprentissage par renforcement**:
  - La machine peut aussi faire des actions a1,a2… qui changent l'état du monde
  - elle reçoit en retour une récompense r1,r2…
  - Son but est de maximiser les récompenses sur le « long » terme.
- **Apprentissage non supervisé**
  - construire un modèle des données,
  - exploiter les régularités statistiques existantes
  - résumer l'information, trouver une représentation compacte
  - « comprendre » en introduisant des concepts de plus haut niveau
  - Afin de raisonner, prendre des décisions, prédire, communiquer…

Apprentissage non supervisé

Apprentissage bayésien: Méthode

1. Formuler notre connaissance du problème
   - Définir une classe de modèles avec paramètres inconnus
   - Spécifier une distribution a priori sur ces param. exprimant notre degré de croyance sur leur vraisemblance avant d'avoir vu les données
2. Collecter des données
3. Calculer la probabilités *a posteriori* des paramètres, sachant les données
4. Utiliser ce « posterior » pour
   - Conclure en tenant compte rigoureusement des incertitudes
   - Faire des prédictions
   - Prendre des décisions minimisant un coût espéré (expectation)

Probabilités subjectives

- Le hasard est une notion subjective, dépendant des informations possédées.

- Théorème de Cox:

  Si on veut exprimer les degrés de croyance par des réels, en respectant certaine contraintes de bon sens et de consistance, alors ces degrés doivent respecter les règles de la théorie des probabilités.

- Dutch Book Theorem:

  Si vous acceptez des paris basés sur vos degrés de croyance en ne respectant pas les règles des probabilités, alors il existe des paris qui vous feront perdre quelque soit l'issue du tirage.

- Conclusion:

  Être cohérent = suivre les règles de « Probability as Logic » [Jaynes]

  Donc P(X=x) représente à la fois

  . La fréquence avec laquelle X prend la valeur x (Loi des grands nb.)

  . Le degré de croyance que X=x (Cox)

Règles de calcul

- x: variable discrète ou continue

$$P(x) \geq 0 \quad f_X(x) \geq 0$$

$$\sum_{i=0}^{k} P(x_i) = 1 \qquad \int f_X(x)dx = \int p(x)dx = 1$$

- Probabilité conjointe $P(x, y) = P(x \; and \; y)$
- Probabilité conditionnelle $P(x|y) = P(x, y) / P(y)$
- Marginalisation $P(x) = \sum_y P(x,y)$
- Règle de Bayes

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Apprentissage bayésien

- Pour apprendre un modèle paramétrique:

$$P(paramètres \mid données) = \frac{P(données \mid paramètres) P(paramètres)}{P(données)}$$

$$\propto P(d \mid \theta) P(\mid \theta)$$

$$\boxed{Posterior \propto Likelihood * prior}$$

- Prédiction

$$P(new\ x \mid données) = \int_{param} P(new\ x \mid param) P(param \mid données)$$

- On trouve aussi

$$P(x \mid d, M) = \int_{\theta} P(x \mid \theta) P(\theta \mid d, M) d\theta$$

Limitations et Avantages

- Avantages
  - Cadre cohérent
  - Prise en compte optimale de l'incertitude
  - Ne présuppose pas l'existence d'une « vraie » distribution suivie par les données
  - Pas d'overfitting si bon prior
  - Beaucoup de méthodes se reformulent en bayésien (ML, Kalman, Markov, PCA, ICA, EM…)
- Limitations
  - Prior est subjectif, mais non arbitraire
  - Difficulté de trouver le bon prior
  - Très lourd en calcul
  - Beaucoup de monde ne fait pas du « vrai » bayésien
    - Pas de distribution sur les modèles
    - Posteriors impropres
    - Priors ridicules, seulement agréable (conjugate prior)
    - Utilisation que du MAP
    - Prior systématiquement uniforme

Terminologie

- Maximum Likelihood (ML) Learning

$$\theta^* = ArgMax\ P(d|\theta)$$

- Maximum a Posteriori (MAP) Learning:

$$\theta^* = ArgMax\ P(\theta|d) = ArgMax\ P(d|\theta)P(\theta)$$

- Bayesian Learning: On garde

$$P(\theta|d) \propto P(d|\theta)P(\theta)$$

Comparaison de modèles

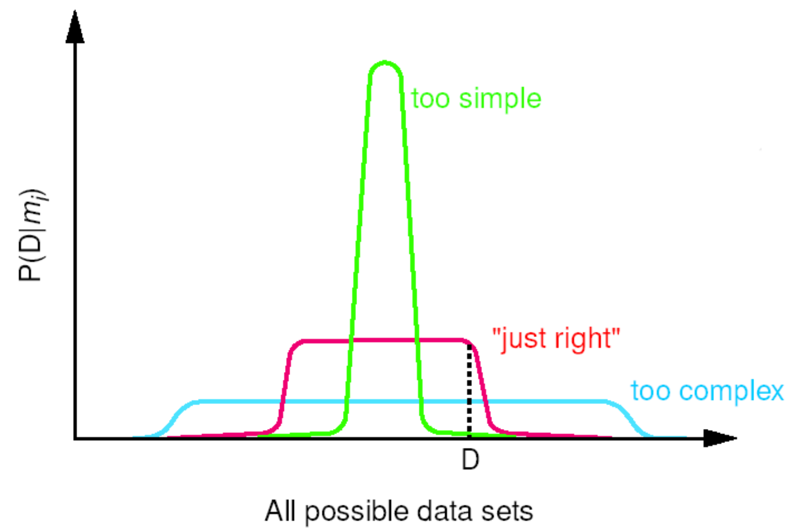
- Choix entre deux modèles

$$\frac{P(M_1|d)}{P(M_2|d)} = \frac{P(d|M_1)P(M_1)}{P(d|M_2)P(M_2)} = \frac{\int P(d|\theta, M_1)P(\theta|M_1)d\theta}{\int P(d|\theta, M_2)P(\theta|M_2)d\theta}\frac{P(M_1)}{P(M_2)}$$

- Rasoir d'Occam automatique

Une classe de modèles trop simple ou trop complexe donnera une faible probabilité au jeu de données

La classe de modèles la plus probable pour un jeu de données aura une complexité adaptée.



Choix des priors

- Non informatifs: « objectifs »
  - Invariance par re-paramétrisation (Jeffrey's prior)
  - Impropres (uniformes sur R)
- Informatifs: capture une connaissance
  - Aussi bien que possible
  - Maximum d'entropie
- Prior hiérarchiques:
  - Distribution sur la distribution sur les paramètres : hyper paramètres, hyper-hyper-hyper…

$$P(\theta) = \int P(\theta|\alpha)P(\alpha)d\alpha = \int P(\theta|\alpha)\int P(\alpha|\beta)P(\beta)d\beta d\alpha = \ldots$$

Choix des priors (cont)

- Priors empiriques
  - Ex: paramètre et hyper paramètre
    $$P(d|\alpha) = \int P(d|\theta)P(\theta|\alpha)d\theta$$
  - Alpha estimé a partir des données
    $$\alpha_{ML2} = ArgMax\ P(d|\alpha)$$
  - Prédiction
    $$P(x|d,\alpha^*) = \int P(x|\theta)P(\theta|d,\alpha^*)d\theta$$
  - Robuste mais on compte 2 fois les données, overfitting

Choix des priors (Fin)

- Priors conjugués de la vraisemblance
  - Pour simplifier les calculs analytiques
  - Tels que le posterior ait la même forme que la vraisemblance
  - Ex: Dirichlet si vraisemblance multinomiale

Inférence générale

$$\mathrm{P}(Search \mid Known \otimes \delta \otimes \pi) = \sum_{Unknown} \mathrm{P}(Search \otimes Unknown \mid Known \otimes \delta \otimes \pi)$$

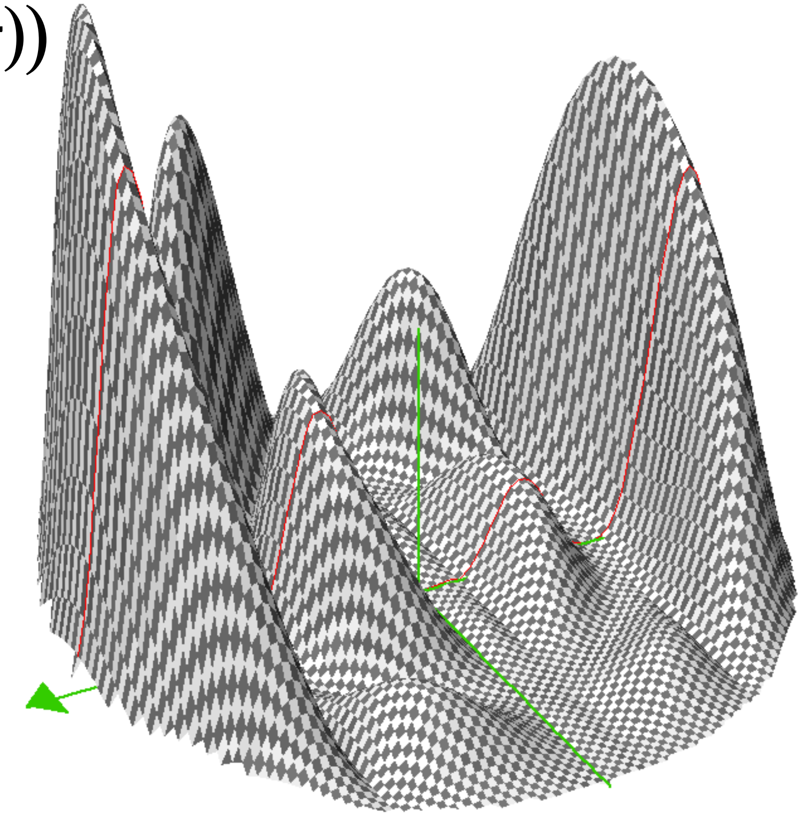
$$= \frac{\sum_{Unknown} \mathrm{P}(Search \otimes Unknown \otimes Known \mid \delta \otimes \pi)}{\mathrm{P}(Known \mid \delta \otimes \pi)}$$

$$= \frac{\sum_{Unknown} \mathrm{P}(Search \otimes Unknown \otimes Known \mid \delta \otimes \pi)}{\sum_{Search, Unknown} \mathrm{P}(Search \otimes Unknown \otimes Known \mid \delta \otimes \pi)}$$

$$= \frac{1}{Z} \times \sum_{Unknown} \mathrm{P}(Search \otimes Unknown \otimes Known \mid \delta \otimes \pi)$$

Coût calculatoire: 2 pb

$$\mathrm{Draw}\big(\mathrm{P}(Search \mid Known \otimes \delta \otimes \pi)\big)$$



$$\mathrm{P}(Search \mid Known \otimes \delta \otimes \pi)$$

$$= \frac{1}{Z} \times \sum_{Unknwon} \mathrm{P}(Search \otimes Known \otimes Unknown \mid \delta \otimes \pi)$$

Approximations

- Laplace
  - Approx gaussienne du posterior autour du MAP
- Bayesian information Criterion
- Variational approximation
  - Minorer la vraisemblance marginale (comme EM)
- Makov Chain Monte Carlo MCMC
  - Simuler une chaine de markov convergeant vers le posterior
- Exact sampling

Exemple 1

- Jeu de Pile ou Face (T ou H)
  - Paramètre q: $P(H)=q$ et $P(T)=1-q$
- 2 modèles de la pièces
  - Équilibrée $P(H)=q=0.5$
  - Truquée P(H)=totalement inconnu
- Il faut une distribution sur p pour formaliser
  - Équilibrée P(q)=dirac en 0.5
  - Truquée P(q)=contant=1 pour tout q dans 0..1
- Données: 10 lancés: THTHTTTTTT
- Question: quel est le modèle le plus probable ?

Exemple 2

- **Réponse: Prior nécessaire**
  - P(Équilibrée) = 0.8 et P(Truquée) = 0.2
- **Et les vraisemblances sont**
  - P(d | Équilibrée)
    = P(THTHTTTTTT|Eq)
    = P(T|Eq) P(H|eq) P(T|Eq) P(H|eq)....
    = (1/2)^10 = 0.001

$$P(d|Truquée) = \int_{q=0}^{1} P(d|Truquée, q) P(q|Truquée) dq$$

$$= \int_{q=0}^{1} P(H|Truquée, q)^2 P(T|Truquée, q)^8 * 1 * dq$$

$$= \int_{q=0}^{1} q^2 (1-q)^8 dq = ... = 0.002$$

Exemple fin

- Les rapport des posteriors est

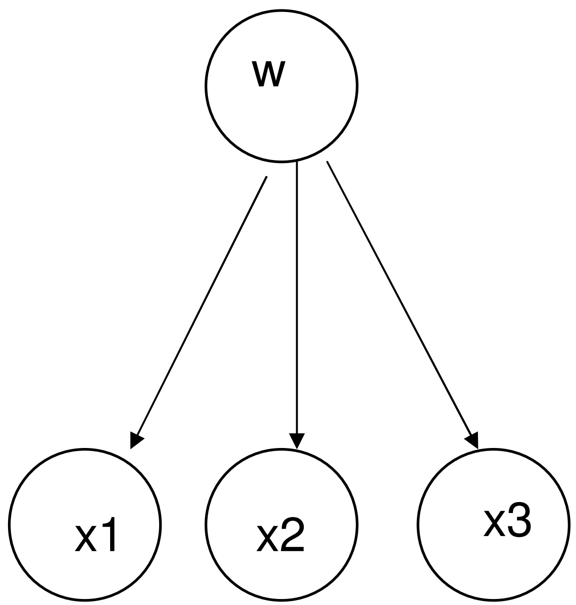
$$\frac{P(Eq|d)}{P(Tru|d)} = \frac{P(d|Eq)P(Eq)}{P(d|Tru)P(Tru)} = \frac{0.001 * 0.8}{0.002 * 0.2} = 2$$

- Donc P(Eq|d)=2/3
- Prédiction: faire une face au prochain coup
  - P(H | d) = P(H | Tru d) P(Tru | d) + P(H | Eq d) P(Eq | d)

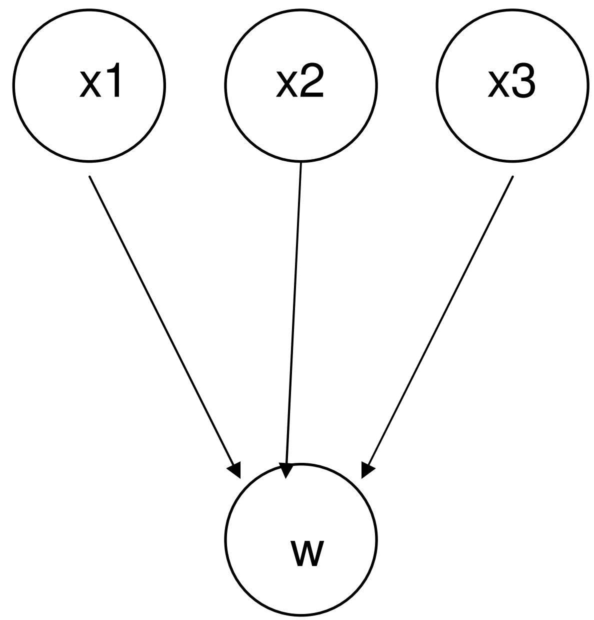
    = 3/12 * 1/3 + ½ * 2/3 = 5/12

Modèle génératif / discriminatif

- Une classification binaire peut s'exprimer
  - Génératif
  - Discriminatif



$$P(x,w) = P(w)P(x|w)$$



$$P(x,w) = P(x)P(w|x)$$

Modèle génératif

$$P(\vec{x}|w_i) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\vec{x} - \eta_i)^T \Sigma^{-1} (\vec{x} - \eta_i) \right)$$

$$P(w_0|\vec{x}) = \frac{P(\vec{x}|w_0)P(w_0)}{P(\vec{x})} = \frac{P(\vec{x}|w_0)P(w_0)}{P(\vec{x}|w_0)P(w_0) + P(\vec{x}|w_1)P(w_1)}$$

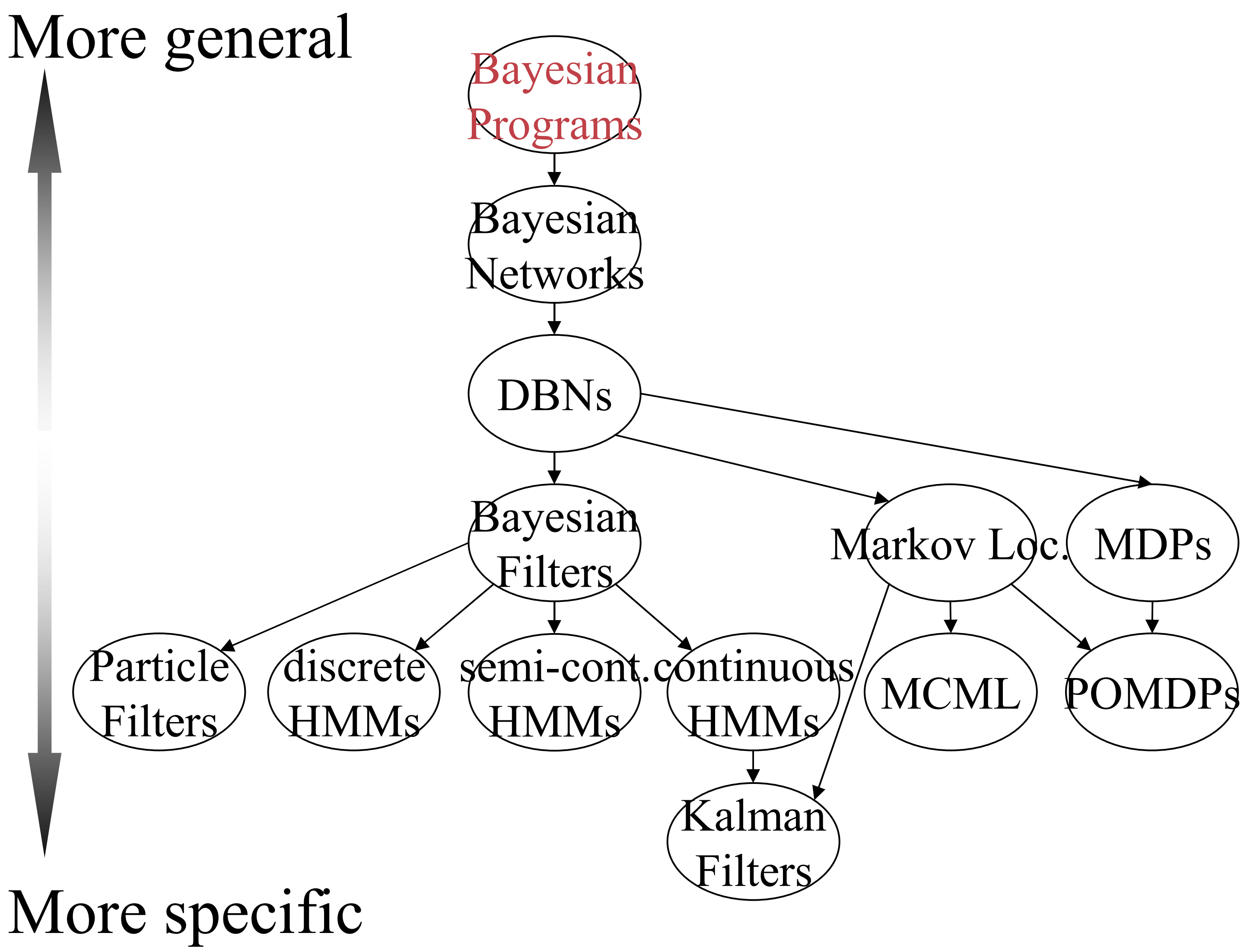
$$= \frac{1}{1 + \exp\left(-\log\left[\frac{P(\vec{x}|w_0)}{P(\vec{x}|w_1)} \right] - \log\left[\frac{P(w_0)}{P(w_1)} \right] \right)}$$

$$= \frac{1}{1 + e^{-(w^T \vec{x} + b)}}$$

Le Bayésien chez E-Motion

- **Modèles bayésiens** :
  - variété de techniques de modélisation prenant en compte l'incomplétude et les incertitudes,
- **Programmation bayésienne** :
  - un formalisme générique pour implanter une variété de modèles
- **ProBT** :
  - une bibliothèque disponible pour réaliser l'inférence efficacement

Bayesian Models

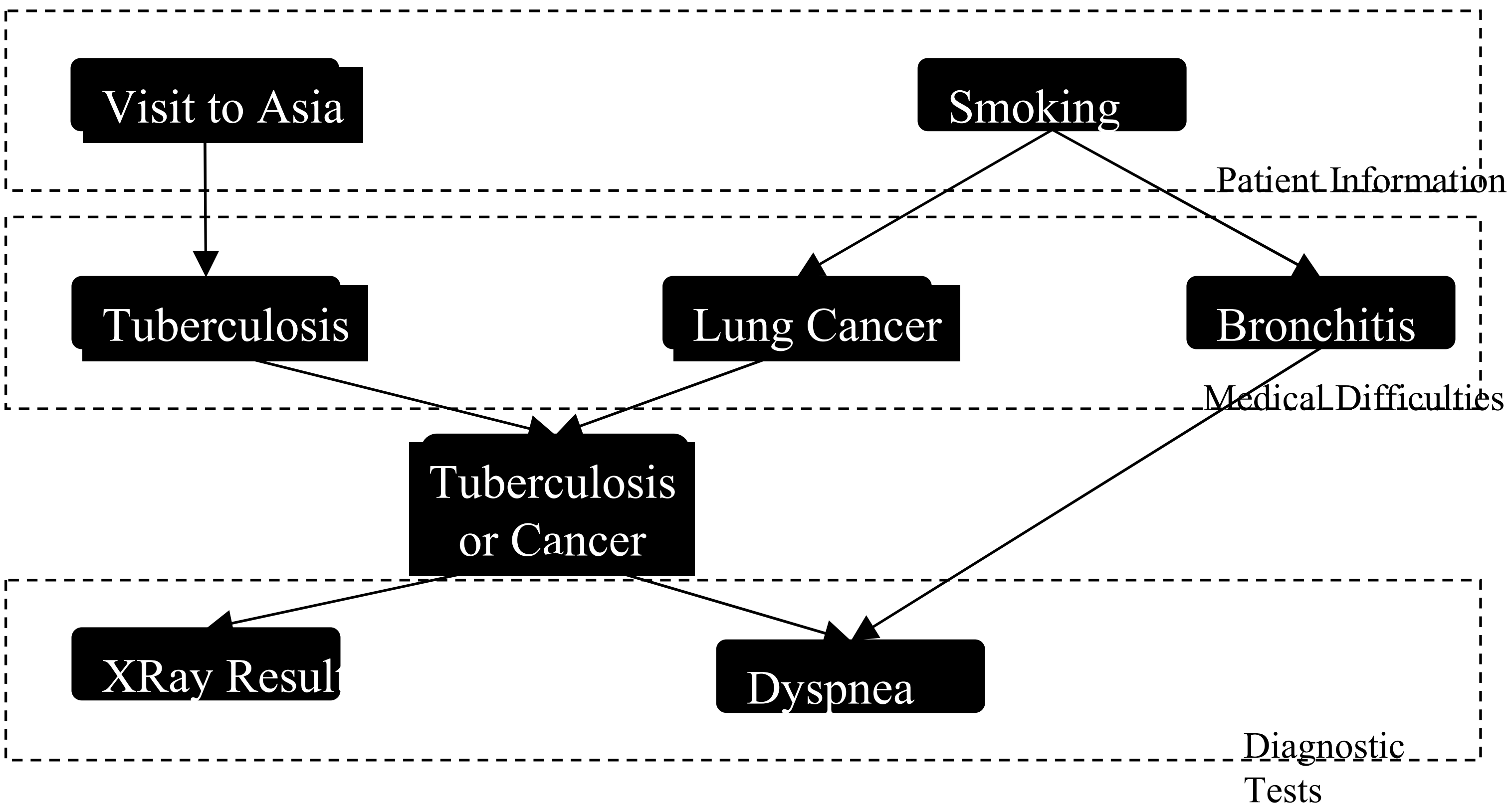


Graphical models
definition

- graphical model = graph
  - A node represents a random variable;
  - An arc between two nodes represents a conditional dependence between these two nodes;
  - A conditional probability distribution associated with each node.

  - Used to represent conditional independence between variables to perform inference.

Réseaux Bayésiens



Network represents a knowledge structure that models the relationship between medical difficulties, their causes and effects, patient information and diagnostic tests

Programme Bayésien

Programme

Description

Spécification **(priors)**

Relevant variables

θ et $X_1 \ldots X_n$

Décomposition

Modèle Capteur ou Fusion Naïve

$$P(X_1 \ldots X_n, \theta) = P(\theta) \prod_{i=1}^{n} P(X_i \mid \theta)$$

Forme paramétriques

Histogrammes

Identification

Apprendre les n histogramme à partir des données

Question

$$ArgMax_{\theta}\ P(\theta \mid X_9 \ldots X_{24})$$

Applications

- Robot Programming
  - Reactive Behaviors
  - Sensor Fusion
  - Combining Descriptions
  - Markov Localization
  - HMM - POMDP - MDP
  - Hunting
  - Smelling
  - Object Recognition
  - Nightwatchman task
  - Bot Inverse Programming
  - Teaching Bot how to play
  - Proscriptive Programming
  - Bayesian Occupation Filter
  - Bayesian maps
  - Pick and Place
- CAD Modelling
  - Bayesian CAD system
  - Knee prosthesis
- Industrial Application
  - SPAM detection
  - Product classification
  - Troubleshooting
  - Containers cost transport
  - Stock Picking
  - Sale prevision and Stock managing
  - Preventive maintenance

Références

- David McKay
- Micheal Jordan
- Radford Neal
- Zoubin Ghahramani
- Judea Pearl
- Edwin T. Jaynes
- Tom Minka
- David Heckerman