

# Apprentissage bayésien

Synthèse de lectures

(Séminaire sur l'apprentissage automatique)

de

Benoit Lavoie

[benoit@benoit-lavoie.ca](mailto:benoit@benoit-lavoie.ca)

Programme de Doctorat en Informatique Cognitive

Université du Québec à Montréal

11 avril 2006

# Table des matières

<b>1. INTRODUCTION .....</b>	<b>1</b>
<b>2. THÉORÈME DE BAYES ET CONCEPTS RELIÉS .....</b>	<b>1</b>
THÉORÈME .....	1
HYPOTHÈSE AVEC PROBABILITÉ A POSTERIORI MAXIMUM.....	2
HYPOTHÈSE AVEC LIKELIHOOD MAXIMUM .....	2
ALGORITHME DE FORCE BRUTE .....	2
<b>3. CLASSIFICATEURS BAYÉSIENS .....</b>	<b>2</b>
CLASSIFICATEUR BAYÉSIEN OPTIMAL.....	2
CLASSIFICATEUR BAYÉSIEN NAÏF.....	3
CLASSIFICATEUR À BASE DE RÉSEAU BAYÉSIEN.....	3
EXEMPLES D'IMPLÉMENTATION DE CLASSIFICATEURS BAYÉSIENS .....	3
<b>4. RÉSEAUX BAYÉSIENS.....</b>	<b>3</b>
REPRÉSENTATION .....	3
APPRENTISSAGE D'UN RÉSEAU BAYÉSIEN.....	4
<b>5. AVANTAGES ET INCONVÉNIENTS DES MODÈLES BAYÉSIENS .....</b>	<b>5</b>
AVANTAGES .....	5
INCONVÉNIENTS .....	6
<b>6. RÉFÉRENCES .....</b>	<b>6</b>

## 1. Introduction

Ce document présente une synthèse de travaux portant sur l'apprentissage bayésien, faisant ressortir les grandes tendances dans ce domaine. Ce travail a été réalisé dans le cadre d'un séminaire sur l'apprentissage automatique (DIC9380) du programme de Doctorat en Informatique Cognitive de l'Université du Québec à Montréal (UQAM).

La Section 2 de ce document présente des concepts reliés au théorème de Bayes sur lequel repose l'apprentissage bayésien. La Section 3 décrit des classificateurs bayésiens et la Section 4 décrit les réseaux bayésiens qui peuvent également être utilisés comme classificateurs. La Section 5 présente des avantages et inconvénients des modèles bayésiens.

## 2. Théorème de Bayes et concepts reliés

L'apprentissage bayésien repose sur le raisonnement bayésien (Mitchell, 1997; Lounis, 2006). Ce type de raisonnement trouve son fondement théorique dans le Théorème de Bayes: il permet les inférences probabilistes et il repose sur l'hypothèse que les solutions recherchées peuvent être trouvées à partir de distributions de probabilité dans les données et dans les hypothèses.

### Théorème

Le Théorème de Bayes fut développé au 18<sup>ème</sup> siècle par Thomas Bayes (Wikipedia, 2006; Witten & Frank, 2005, p.141). Le Théorème de Bayes associe la *probabilité a posteriori d'une hypothèse h sachant les données D*,  $P(h | D)$ , à 3 autres probabilités (Lounis, 2006; Mitchell, 1997):

$$P(h | D) = \frac{P(D | h) \times P(h)}{P(D)}$$

où

- $P(h)$  = probabilité que l'hypothèse  $h$  soit vérifiée indépendamment des données  $D$  (ce terme est également appelé *probabilité a priori* ou *prior probability*);
- $P(D)$  = probabilité d'observer les données  $D$  indépendamment de l'hypothèse  $h$  (ce terme est également appelé *évidence*);
- $P(D | h)$  = probabilité d'observer les données  $D$  sachant que l'hypothèse  $h$  est vérifiée (ce terme est également appelé *likelihood*).

## Hypothèse avec probabilité a posteriori maximum

Le Théorème de Bayes peut être utilisé afin de déterminer l'une des hypothèses la plus probable selon les données; i.e. une hypothèse maximisant  $P ( h | D )$ . Cette hypothèse avec probabilité a posteriori maximum ( $h_{MAP}$ ) est définie par:

$$h_{MAP} = \operatorname{argmax}_h P ( h | D )$$

Une méthode de raisonnement (ou d'apprentissage) qui recherche  $h_{MAP}$  est dite méthode de probabilité a posteriori maximum.

## Hypothèse avec *likelihood* maximum

Le Théorème de Bayes peut être utilisé afin de déterminer l'une des hypothèses pour laquelle la probabilité d'observer des données est maximale; i.e. une hypothèse maximisant  $P ( D | h )$ . Cette hypothèse avec *likelihood* maximum ( $h_{ML}$ ) est définie par:

$$h_{ML} = \operatorname{argmax}_h P ( D | h )$$

Une méthode de raisonnement (ou d'apprentissage) qui recherche  $h_{ML}$  est dite méthode de *likelihood* maximum.

## Algorithme de force brute

Un algorithme de force brute recherche à travers toutes les hypothèses, soit  $h_{MAP}$  (une hypothèse maximisant la probabilité a posteriori), ou soit  $h_{ML}$  (une hypothèse maximisant le *likelihood*).

## 3. Classificateurs bayésiens

Les classificateurs bayésiens utilisent des méthodes basées sur le Théorème de Bayes afin de déterminer les probabilités d'associer certaines classes à certaines instances selon les données d'entraînement (Mitchell, 1997; Lounis, 2006).

### Classificateur bayésien optimal

Un classificateur bayésien optimal permet de déterminer la classification la plus probable d'une instance selon les données d'entraînement. La méthode de calcul est différente de la méthode de probabilité a posteriori maximum (voir Section 2). La classification la plus probable d'une nouvelle instance est obtenue en combinant les prédictions des toutes les hypothèses qui sont pondérées par leurs probabilités a priori. Pour un ensemble de classes  $C$ , la valeur de classification optimale de Bayes,  $c$ , est définie comme suit:

$$c = \underset{c_j \in C}{\operatorname{argmax}} \sum_{h_i \in H} P(c_j | h_i) P(h_i | D)$$

### Classificateur bayésien naïf

Un classificateur bayésien naïf permet de déterminer la classification d'une instance spécifiée en terme d'attributs en présupposant que les attributs sont indépendants. Cette présupposition d'indépendance des attributs ne tient pas compte de la réalité dans beaucoup de domaines, d'où l'épithète *naïf* pour qualifier ce type de classificateur. Pour un ensemble de classes possibles  $C$  et une instance spécifiée par un ensemble d'attributs  $A$ , la valeur de classification bayésienne naïve,  $c$ , correspondant à ces attributs est définie comme suit:

$$c = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{a_i \in A} P(a_i | c_j)$$

Un classificateur bayésien naïf est efficace et, dans certains domaines, ses résultats sont compétitifs à ceux des meilleures méthodes. Cette efficacité s'applique même dans des domaines où la présupposition d'indépendance des attributs ne s'applique pas tout à fait: le domaine de la classification de document en particulier est un domaine pour lequel les classificateurs bayésiens naïfs sont souvent utilisés avec succès malgré qu'il existe une certaine dépendance entre les attributs (mots d'un document) (Mitchell, 1997; Witten & Frank, 2005).

### Classificateur à base de réseau bayésien

Les réseaux bayésiens présentés à la Section 4 peuvent également être utilisés pour la classification.

### Exemples d'implémentation de classificateurs bayésiens

Le logiciel de forage de données Weka (Witten & Frank, 2005) implémente plusieurs classificateurs bayésiens dont différentes variantes de classificateurs bayésiens naïfs et une variante de classificateur à base de réseaux bayésiens.

## 4. Réseaux bayésiens

### Représentation

Un réseau bayésien (ou réseau de croyance bayésien) (Mitchell, 1997; Lounis, 2006) décrit la distribution de probabilités associées à un ensemble de variables (correspondant à des attributs ou à des hypothèses) dont certaines sont directement dépendantes et d'autres sont indépendantes conditionnellement. Un réseau bayésien est représenté comme un graphe orienté acyclique où les nœuds représentent les variables auxquelles sont associées des probabilités conditionnelles locales. Le réseau représente la distribution

jointe de probabilité pour toutes les variables qu'il représente.

Dans un réseau bayésien, un nœud donné est parent d'un autre nœud si une relation orientée part de ce premier nœud pour relier directement le second. La notion de parent est centrale dans le calcul de la distribution des probabilités jointes entre un ensemble de variables. La probabilité jointe d'un ensemble de variables  $x_1, x_2, \dots, x_n$  est définie comme suit:

$$P(x_1, x_2, \dots, x_n) = \prod_i P(x_i | \text{Parents}(x_i))$$

où

$\text{Parents}(x_i)$  = ensemble des variables parents de  $x_i$  (i.e. d'où partent des relations directes avec  $x_i$ ).

Dans le cadre de l'apprentissage, un réseau bayésien peut être utilisé pour la classification si l'une des variables représente la classe cible.

Étant donné sa flexibilité à l'égard des relations entre variables, un réseau bayésien peut être vu comme un modèle intermédiaire entre le modèle bayésien optimal qui présume la dépendance entre toutes les variables et le modèle bayésien naïf qui présume l'indépendance entre toutes les variables.

Selon Pearl et Russell (2000), la caractéristique la plus importante des réseaux bayésiens se rapporte à leur utilisation pour représenter directement les connaissances du domaine et non des procédures de raisonnement. Toutefois, ces connaissances permettent le raisonnement bidirectionnel: une relation de dépendance peut être traitée afin de permettre le raisonnement dans les deux sens de la relation.

### **Apprentissage d'un réseau bayésien**

Un réseau bayésien peut être appris dans différents contextes (Mitchell, 1997; Lounis, 2006): (i) la structure du réseau peut être a priori connue (totalement ou partiellement) ou inconnue; et (ii) toutes les variables peuvent être observables à partir des données ou certaines peuvent être manquantes.

Dans le cas où la structure est connue totalement a priori et que toutes les variables sont observables à partir des données, l'apprentissage des probabilités conditionnelles associées aux variables (nœuds du réseau) peut se faire à partir des fréquences des valeurs dans les données d'entraînement.

Dans le cas où la structure est connue totalement a priori mais que certaines variables ne sont pas observables dans les données, on peut utiliser la méthode du gradient ascendant afin de déterminer les probabilités conditionnelles associées aux variables: la fonction maximisée par cette méthode est  $P(D | h)$  qui correspond à chercher les hypothèses à *likelihood* maximum.

Dans le cas où la structure n'est pas connue a priori, on peut tenter de déterminer des réseaux alternatifs à l'aide d'heuristiques de recherche (e.g. recherche gloutonne) et sélectionner un réseau particulier à l'aide d'une métrique d'évaluation (essayant de minimiser la complexité du réseau et/ou maximiser les probabilités jointes du réseau).

## 5. Avantages et inconvénients des modèles bayésiens

### Avantages

Le raisonnement (et l'apprentissage) bayésien offre plusieurs avantages (Lounis, 2006; Witten & Frank, 2005; Mitchell, 1997):

- il permet de supporter des données bruitées;
- il permet de supporter des données manquantes;
- il associe des probabilités aux prédictions, ce qui est utile dans les nombreux domaines où les connaissances sont incertaines;
- il est utilisé par certaines techniques de classification qui sont parmi les plus pratiques ou les plus performantes dans certains domaines;
- il permet le support de connaissances a priori;
- il fournit une approche théorique quantitative permettant l'analyse de d'autres approches qui ne sont pas nécessairement basées sur des modèles probabilistes; et
- il permet le traitement incrémentale des données.

Les réseaux bayésiens ont des avantages supplémentaires qui sont liés à leurs représentations (Pearl & Russell, 2000; *Unknown*<sup>-1</sup>):

- leurs représentations permettent le raisonnement de façon bidirectionnelle (i.e. en suivant les relations de dépendances entre variables dans les deux directions);

---

<sup>1</sup> *Unknown* fait référence à une publication disponible en ligne mais dont le nom de l'auteur et le titre ne sont pas indiqués: [http://www.info2.uqam.ca/~lounis/dic9380-30/bayes/complement\\_bayes.pdf](http://www.info2.uqam.ca/~lounis/dic9380-30/bayes/complement_bayes.pdf)

- leurs représentations facilitent la compréhensibilité dans un domaine de connaissance (elles représentent directement les connaissances du domaine et non des procédures de raisonnement); et
- leurs représentations modélisent explicitement tous les liens de dépendances entre variables.

## Inconvénients

Le raisonnement (et l'apprentissage) bayésien est associé à certains désavantages (Mitchell, 1997):

- son application nécessite des probabilités dont la détermination requière typiquement de grandes quantités de données ou plusieurs connaissances a priori;
- il nécessite un coût de calcul relativement élevé pour déterminer l'hypothèse optimale dans un cas général; et
- un modèle de probabilité n'est parfois pas un concept intuitif pour un expert du domaine.

Les réseaux bayésiens ont des désavantages supplémentaires qui sont liés à leurs représentations (*Unknown*):

- la compréhensibilité des réseaux peut devenir difficile avec plusieurs variables et/ou plusieurs liens de dépendances; et
- les données continues doivent être discrétisées.

## 6. Références

Lounis, Hakim (2006). *Apprentissage Bayésien*. Notes de cours (Séminaire sur l'apprentissage automatique), Programme de Doctorat en Informatique Cognitive, Université du Québec à Montréal.

Mitchell, Tom (1997). Bayesian Learning. Chapter 6 of *Machine Learning*, McGraw-Hill, pp. 274–306.

Pearl, Judea; & Stuart Russell (2000). Bayesian Networks. UCLA Cognitive Systems Laboratory, Technical Report (R-277), November.

*Unknown*. Sections d'un mémoire de thèse sur les réseaux Bayésiens. Université du Québec à Montréal.

En ligne: [http://www.info2.uqam.ca/~lounis/dic9380-30/bayes/complement\\_bayes.pdf](http://www.info2.uqam.ca/~lounis/dic9380-30/bayes/complement_bayes.pdf)

Wikipedia (Editor, 2006). Thomas Bayes. *Wikipedia*. Wikipedia Foundation, Inc. (Editor). (Online encyclopedia).

*En ligne:* [http://en.wikipedia.org/wiki/Thomas\\_Bayes](http://en.wikipedia.org/wiki/Thomas_Bayes)

Witten, Ian H.; & Eibe Frank (2005, 2<sup>nd</sup> edition). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.