Année universitaire : 2018 - 2019

Ecole Nationale d'Ingénieurs de Sousse

RAPPORT PROJET MODULE INTÉLLIGENCE ARTIFICIELLE DISTRIBUÉE

LES RÉSEAUX BAYÉSIENS

Proposé par : CHAINBI Walid

Réalisé par : SLAMA Hamza - BOUALI Houssem

Groupe : IA3.1 - Ingénieurie des systèmes distribuées

Table des matières

Table des figures

Résumé

Depuis l'émergence de la robotique et de l'informatique, les chercheurs essaient d'injecter des notions d'intelligence humaine dans des machines. Étant conçue et fabriquée par l'homme, on qualifie cette forme d'intelligence comme "L'intelligence artificielle" ou IA.

L'intelligence artificielle est souvent définie comme la science de créer et de programmer des ordinateurs capables d'accomplir des taches qui nécessitent l'intelligence quand elle doit être fait par un être humain : que ce soit faire des mathématiques, jouer aux échecs ou même juste parler.

Depuis les débuts de l'IA dans les années 1950, ce domaine de recherche s'est enrichit avec des nombre incalculables de recherches et d'expériences qui ont était fait dans les derniers 60 ans. Aujourd'hui, il existe des ordinateurs capables de résoudre des équations et des problèmes en quelques dixièmes de secondes, dont l'homme était incapable de résoudre il y a 100 ans ; il existe des ordinateurs capables de mener une discussion avec un humain, qui sont capables de répondre a des questions comme un humain ; il y a même des ordinateurs capables de reconnaître son "propriétaire" juste par la voix.

Effectivement, grâce à l'IA, il existe aujourd'hui des ordinateurs et des systèmes capables de faire des choses incroyables. Mais cette "intelligence" vient d'une source commun : l'homme, qui a transmis à ces ordinateurs ses propres capacités ; c'est sûrement un des développements le plus important dans l'histoire moderne de l'humanité et qui continue à nous révolutionner la façon de vivre.

Chapitre 1

Réseaux Bayesiens

1.1 Introduction

L'émergence de l'intelligence artificielle a déclencher les recherches vers des applications et de nouvelles approches pour mieux améliorer la capacité d'apprentissage ainsi que la prise de décision par les machines. Parmi ces domaines, on trouve les réseaux bayésiens qui sont tout simplement le mariage entre la **théorie des graphes** et la **théorie des probabilités**.

1.2 Définition

Un réseau bayésien est un système représentant la connaissance et permettant de calculer des probabilités conditionnelles apportant des solutions à différentes sortes de problématiques.

La structure de ce type de réseau est simple : un **graphe** dans lequel les noeuds représentent des **variables aléatoires**, et les arcs (le graphe est donc orienté) reliant ces dernières sont rattachées à des **probabilités conditionnelles**. Notons que le graphe est acyclique : il ne contient pas de boucle. Les arcs représentent des relations entre variables qui sont soit déterministes, soit probabilistes. Ainsi, l'observation d'une ou plusieurs causes n'entraîne pas systématiquement l'effet ou les effets qui en dépendent, mais modifie seulement la probabilité de les observer.

L'intérêt particulier des réseaux bayésiens est de tenir compte simultanément de connaissances a priori d'experts (dans le graphe) et de l'expérience contenue dans les données.

Pour résumer, un réseau bayésien est un modèle probabiliste graphique permettant d'acquérir, de capitaliser et d'exploiter des connaissances, né du besoin de créer des systèmes experts à base de probabilités.

1.3 Historique

Dans son essai historique publié à titre posthume (Essay Towards Solving a Problem in the Doctrine of Chances, 1763), **Thomas Bayes** a introduit à la fin du 18e siècle deux notions essentielles dans la théorie de la décision.

D'une part, il définit la **probabilité** comme une notion liée à ce qu'on appellerait aujourd'hui l'utilité. **Le prix P** que je suis prêt à payer pour pouvoir bénéficier du **gain R** que m'offrirait la survenance d'un **événement incertain** définit selon Bayes la probabilité de cet événement, comme **P/R**.

D'autre part, il définit la notion de **probabilité conditionnelle**, mettant en évidence le fait que la probabilité est attribuée à un événement à venir et incertain dépend du niveau d'information dont on dispose avant sa survenue. Cette notion est fondamentale car elle exprime le fait que **l'incertitude** est spécifique à chacun, selon son **niveau de connaissance**, et est donc plus proche d'une « **croyance** » que d'une fréquence.

En formalisant le lien intuitif entre « **information** » et « **probabilité** », ou autrement dit entre « **connaissance** » et « **non-connaissance** », Bayes pose les bases de toute **théorie de la décision**. La **décision rationnelle** est celle qui recherche toute l'information disponible - décider en connaissance de cause.

Pendant les années 70 et jusqu'au début des années 80, l'intelligence artificielle est essentiellement représentée par les **systèmes experts**, caractérisés par une approche formelle de la connaissance. La connaissance est vue comme une donnée, manipulée par un outil de déduction logique, le moteur d'**inférence**. Les systèmes experts s'appuient sur la logique formelle, d'ordre 0 ou d'ordre 1, mais toujours caractérisés par une **causalité déterministe**. À partir de données (faits connus P) et de règles (Si P alors Q), ils peuvent déduire de nouveaux faits (Q) en utilisant pour l'essentiel le syllogisme comme règle d'inférence : si P est vrai (fait ou prémisse) et si l'on sait que P implique Q (règle) alors, Q est vrai (nouveau fait ou conclusion).

Au début des années 90, l'utilisation des systèmes experts a commencé à **décliner**; ce déclin tenant, selon nous, à la difficulté du recueil d'expertise sous contrainte déterministe. Autrement dit, et pour l'avoir longuement pratiqué à l'époque, il est très difficile, sinon illusoire, de contraindre un expert à formuler une règle déterministe. En effet, les règles exprimées par des experts sont souvent valides dans un domaine limité. Autrement dit, elles tolèrent les exceptions. Les experts se trouvent alors dans une situation fort inconfortable. En effet, soit ils expriment des règles fausses car partielles - mais le système expert ne sait pas en tenir compte - soit ils doivent identifier toutes les exceptions pour que leurs règles soient effectivement exactes, ce qui est matériellement impossible.

À la fin des années 80, **Judea Pearl**, chercheur américain de l'UCLA, a proposé une approche probabiliste de l'intelligence artificielle, appelée « **réseaux bayesiens** », qui visait précisément à dépasser les limites des systèmes experts et leur incapacité à prendre en compte l'**incertitude** dans le raisonnement. Cette approche intègre en un formalisme très simple l'approche bayesienne des probabilités - la probabilité d'un événement dépend du niveau de connaissance de son contexte et d'une représentation de la causalité.

En 1996, Steve Ballmer indiqua, dans Los Angeles Times, que selon Bill Gates, « **les réseaux bayesiens étaient l'avantage concurrentiel de Microsoft** »

1.4 Domaines d'applications

Les applications des réseaux bayésiens peuvent être classées en trois grandes catégories correspondant, finalement, à des utilisations de la théorie bayesienne élémentaire.

- **Le diagnostic** : on utilise le modèle en cherchant à remonter des effets vers les causes.
- **Simulation** : on utilise le modèle en cherchant à évaluer les conséquences de certaines hypothèses.
- **Prise de décision** : on utilise le modèle pour fonder une décision en contexte incertain, c'est-à-dire que l'on cherche à maximiser une utilité espérée.

À ce premier axe de classification, on peut en ajouter un autre selon que le domaine est caractérisé par des observations rares et/ou coûteuses, ou si au contraire les données sont abondantes. Dans le premier cas, le modèle sera construit par un expert et l'application considérée sera son utilisation. Dans le deuxième cas, le modèle sera construit à partir de données expérimentales. L'application pourra alors se limiter à observer le modèle obtenu, qui est finalement une excellente synthèse des données. Un modèle obtenu par apprentissage peut bien sûr être utilisé sur de nouvelles données.

Les situations mixtes sont évidemment possibles : domaines où de nombreuses données expérimentales cohabitent avec un vaste corpus théorique. L'une des applications les plus banales est l'**antispam** qui décide de la probabilité qu'un email soit ou non indésirable, selon un certain nombre de critères. Même si la technologie utilisée est en général plus rudimentaire que celle des réseaux bayesiens (on parle de filtre bayesien ou de modèle bayesien naïf), c'est sans doute grâce à cette application qu'un ordinateur personnel ou serveur applique tous les jours le théorème de Bayes pour une décision que l'on peut qualifier d'**« intelligente »**.

À l'autre extrême, ce que l'on peut considérer comme une évolution de la **doctrine militaire américaine**, les EBO (Effect Based Operations ou « opérations basées sur les effets ») utilisent les réseaux bayesiens pour modéliser l'efficacité des actions menées, les possibles réactions de l'ennemi et envisager de façon globale les « effets » directs et indirects des opérations.

Les applications à la génétique sont nombreuses : de ses utilisations en contexte judiciaire à l'identification de chaînes causales d'expression génétique, les réseaux bayesiens sont l'un des outils privilégiés de ce domaine dont la causalité et les probabilités sont les fondements.

- **Outils informatiques qui implémentent les réseaux bayésiens**
  Les outils de réseaux sont relativement peu nombreux.

  - Hugin (www.hugin.com) : Le plus ancien et le plus performant créé par des anciens de l'université d'Aalborg au Danemark, parmi les fondateurs des réseaux bayesiens.
  - Netica (www.norsys.com) : créé et commercialisé par la société canadienne Norsys est un outil convivial et simple, particulièrement adapté aux développeurs d'applications.
  - BayesiaLab : (www.bayesia.com) créé par une société française implantée à Laval.
  - Elvira (http ://leo.ugr. es/elvira/) :logiciel gratuit disponible à des fins universitaires et Agena (www. agena.co.uk), dont l'offre est orientée sur la gestion des risques, mais qui reste un outil relativement général de réseaux bayesiens.

1.5 Exemple illustratif

Soit un opérateur travaillant sur une machine risque de se blesser, s'il l'utilise mal. Ce risque dépend de l'expérience de l'opérateur et de la complexité de la machine. **«Expérience»** et **«Complexité»** sont deux facteurs déterminants de ce risque (FIGURE 1.1). Bien sûr, ces facteurs ne permettent pas de créer un modèle déterministe.

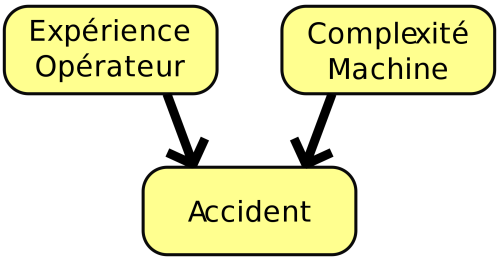

FIGURE 1.1 – Stucture de causalité

Ce graphe peut être complété par des tables de probabilité (FIGURE 1.2). La table la plus importante est celle qui exprime la dépendance de l'accident à l'expérience de l'opérateur et à la complexité de la machine.

Complexité	Faible			Moyenne			Elevée		
Expérience	Faible	Moyenne	Elevée	Faible	Moyenne	Elevée	Faible	Moyenne	Elevée
Accident	1.0	0.5	0.1	1.5	1.0	0.6	2.0	1.6	1.1
Pas d'accident	99.0	99.5	99.9	98.5	99.0	99.4	98.0	98.4	98.9

FIGURE 1.2 – table de probabilités

On comprend bien ici que la connaissance est partielle. Si l'utilisateur est expérimenté et la machine simple, il n'y aura sans doute pas d'accident (0,1% de chances dans l'exemple ci-dessus). Mais le fait de conserver une probabilité résiduelle reconnaît le fait que tous les facteurs de risque ne sont pas pris en compte dans ce modèle.

Le réseau bayesien représente donc à la fois la connaissance et la non connaissance du domaine. Ce qui est connu est représenté par la structure de causalité (le graphe). Ce qui n'est pas connu est matérialisé par des probabilités.

Chapitre 2

Etude des réseaux bayésiens

2.1 Notions fondamentales

2.1.1 Construction d'un graphe

Construire un réseau bayésien revient à :
— Définir le graphe du modèle.
— Définir les tables de probabilités de chaque variable, conditionnellement à ses causes.
Le graphe est aussi appelé la « **structure** » du modèle, et les tables de probabilités ses « **paramétres** ».
Généralement, la structure est définie par des **experts** et les tables de probabilités calculées à partir de **données expérimentales**.
Il est possible d'utiliser des algorithmes tels que les algorithmes génétiques pour construire le réseau.

2.1.2 Espace de recherche

L'espace de recherche est relatif aux nombre de variables bien sûr, mais aussi au nombre d'arcs et de valeurs. Au pire des cas, cet espace peut mesurer $\mathbf{2^{nbvariables}}$, lorsque toutes les variables sont binaires. De plus, il augmente si les variables ont de multiples valeurs (exemple : une variable fumeur peut avoir en valeurs : non, léger, gros). Cependant, cela signifie que toutes les variables sont dépendantes les unes des autres. Or, couramment, les variables ne sont pas dépendantes de toutes les autres, ce qui réduit considérablement la taille de l'espace de recherche.

2.1.3 Table de probabilités

Les tables de probabilités sont définies par des statistiques relatives au problème à résoudre (peuvent aussi être déterminées par des experts). Chacune des variables dispose d'une table de probabilités conditionnelles relatives aux variables causales dont elle dépend. Par exemple, l'alarme (FIGURE 2.1) peut se déclencher soit à cause d'un cambriolage, soit à cause d'un séisme. Les probabilités conditionnelles alarme sachant cambriolage ou/et séisme sont déduites en fonction des probabilités que tel ou tel événement survienne. Lorsqu'une variable possède plusieurs valeurs, pour chacune d'elles est calculé les probabilités conditionnelles en fonction des événements causaux.

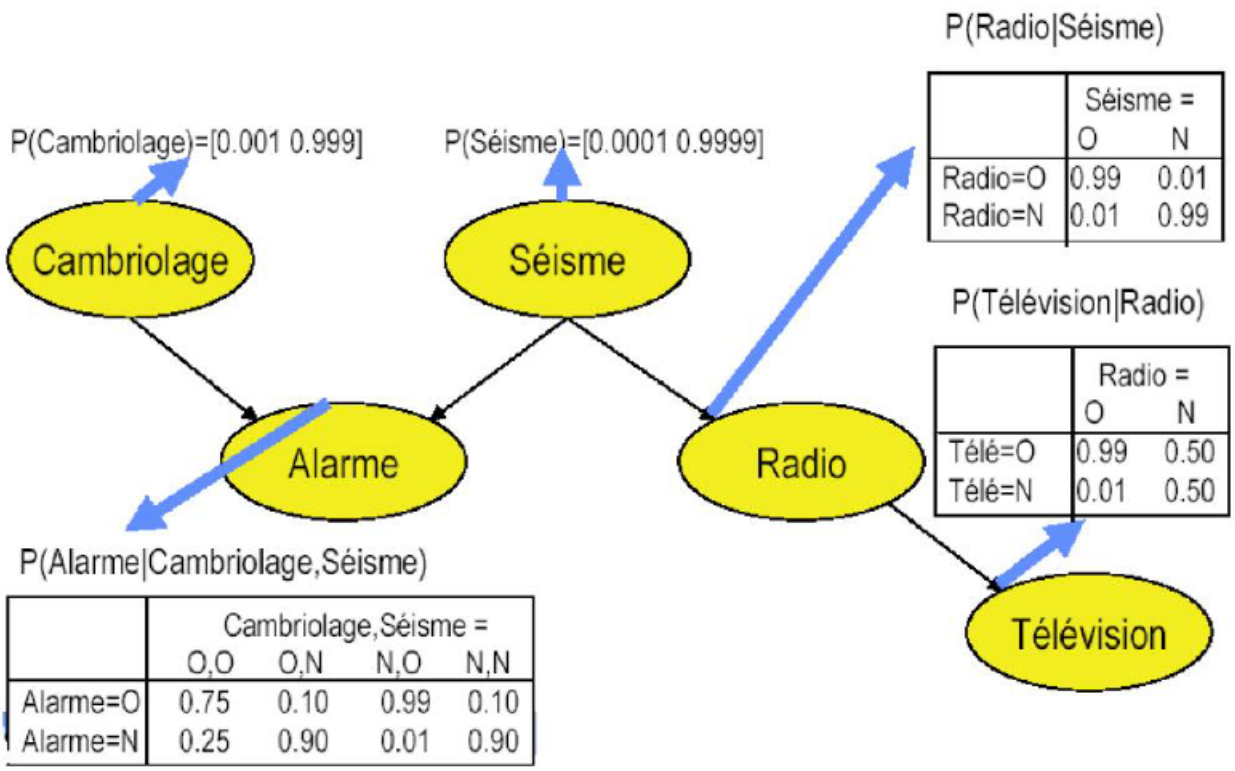

FIGURE 2.1 – exemple illustratif

P(Alarme|Cambriolage,Séisme)

	Cambriolage,Séisme =			
	O,O	O,N	N,O	N,N
Alarme=O	0.75	0.10	0.99	0.10
Alarme=N	0.25	0.90	0.01	0.90

FIGURE 2.2 – table des probabiltés correspondante

2.1.4 Inférence bayésienne

L'inférence bayésienne est basée sur l'utilisation d'énoncés probabilistes, qui dans le cas général sont trouvés par des experts étudiant un système qui leurs ait connu. Ces énoncés doivent être clairs et précis afin d'éviter toute confusion dans les relations de dépendance qui en découleront. L'inférence bayésienne est particulièrement utile dans les problèmes d'induction, car se basant sur des cas particuliers et n'a de validité qu'en terme probabiliste. Les méthodes bayésiennes se distinguent des méthodes dites standard par l'application systématique de règles formelles de transformation des probabilités. On cherche à induire sur un système bayésien aussi bien par le haut que par le bas, aussi bien les conséquences que les causes, du graphe de dépendance. Les règles de la logique des probabilités utilisées sont les suivantes :

— La règle d'addition :
$$p(A \cup B|C) = p(A|C) + p(B|C) - p(A \cap B|C)$$

— La règle de multiplication :
$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A)$$

Le théorème de Bayes peut être dérivé simplement en mettant à profit la symétrie de la règle de multiplication

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

Le théorème de Bayes permet d'inverser les probabilités. C'est-à-dire que si l'on connaît les conséquences d'une cause, l'observation des effets permet de remonter aux causes, c'est l'effet d'induction bf « bottom-up ». Sachant aussi qu'une lecture littéral du théorème de Bayes permet une induction **« top-down »**, c'est à dire à partir des causes en déduire les conséquences. Mais il existe aussi un troisième type d'induction dit **« explaining away »** ou comment réfuter une cause en en constant une autre, autrement dit partir d'une conséquence pour remonter aux causes, constater laquelle est vrai et réfuter les conséquences sous jacentes des autres causes.

En résumé pour induire sur un réseau bayésien, il faut en premier lieu trouver les probabilités conditionnelles de chaque variable aléatoire avec lesquelles elles sont directement dépendantes. Ce que les experts font à partir des statistiques étudiées sur le système voulu. Puis partir de faits produits auxquels on appliquera une probabilité de 1 ou 0 sur le réseau bayésien suivant qu'ils identifient une variable à vrai ou faux dans celui ci. Et enfin par le biais de calcul respectant la règle d'addition et ou de multiplication précédemment décrite, on modifie les probabilités causales et ou conséquentes. La nouvelle probabilité obtenue est l'induction que l'on peut faire sur un réseau bayésien.

2.2 Méthodes de résolution

Les réseaux bayésiens ont été développés au début des années 1980 pour tenter de résoudre certains problèmes de prédiction et d'abduction, courants en intelligence artificielle. Dans ce type de tâche, il est nécessaire de trouver une interprétation cohérente des observations avec les données connues a priori.

2.2.1 Résolution par calcul des inférences

L'inférence probabiliste signifie donc le calcul de $p(Y \setminus X)$ où X est un ensemble d'observations et Y un ensemble de variables décrivant le problème et qui sont jugées importantes pour la prédiction ou le diagnostic.

Le calcul d'inférence probabiliste étant en général **NP-difficile**, deux types de méthodes ont été développés, il y a les approches **complète** et **approximative**.

Il existe une classe de réseaux pour lesquels le calcul est réalisable : les polyarbres (Pour tout noeud x du graphe, x d-sépare ses voisins, et le fait qu'un noeud x d-séparé deux autres noeuds y et z implique que y et z sont indépendant si x est connu).

La D-Séparation

La d-séparation est un critère important qui permet de caractériser graphiquement toutes les contraintes d'indépendance des lois p qui peuvent être représentées par une même **DAG(Direct Acyclic Graph)**. Donc il faut introduire la notion de graphes :

— Chaînes et chemins, simples ou composés.
— Parents, enfants, descendants, ancêtres etc.

Par une chaine de X vers Y transite une information bruitée : les sommets sont des vannes ouverts ou fermés. Une chaîne est dite ouverte si toutes les vannes sont ouvertes auquel cas la chaîne laisse passer l'information, inversement ou si l'une des vannes est bloquér, la chaîne est dite fermée.

$\Rightarrow$ L'information qu'apporte X sue Y peut se voir comme la somme des flots d'information sur tous les chaînes ouvertes reliant X à Y.

Les méthodes complétes

Les premiers algorithmes d'inférence pour les réseaux bayésiens sont basés sur une architecture à passage de messages et ils étaient limités aux arbres. Dans cette technique, à chaque noeud est associé un processeur qui peut envoyer des messages de façon asynchrone à ses voisins jusqu'à ce qu'un équilibre soit atteint, en un nombre fini d'étapes. Cette méthode a été depuis étendue aux réseaux quelconques pour donner l'algorithme de l'**arbre de jonction (JT)**. Cet algorithme s'applique en 4 étapes de transformation du graphe : Moralisation, Absorbé les faits mesuré, Triangulation, Construire l'arbre de Jonction. Et une étape d'induction : Transmettre les messages pour réaliser la cohérence.

Le coût de l'algorithme JT est déterminé par la taille de la plus grande clique et est **exponentiel** en espace, pour les graphes **densément connectés** l'inférence peut être impraticable.

Une autre méthode s'appelle le **cut-set conditionning** : elle consiste à instancier un certain nombre de variables de manière à ce que le graphe restant forme un arbre. On procède à une propagation par messages sur cet arbre. Puis une nouvelle instanciation est choisie. On réitère ce processus jusqu'à ce que toutes les instanciations possibles aient été utilisées. On fait alors la moyenne des résultats. L'avantage de cette méthode est une complexité en temps linéaire sur la taille du réseau, mais le calcul des probabilités conditionnelles est en général impraticable pour les réseaux assez grands car ayant une complexité dans le pire des cas exponentielle dans le nombre de variables et aussi à cause du problème des boucles dans le réseau.

Les méthodes approximatives

Il existe trois approches pour réaliser des inférence approchées, faire comme si le graphe était un arbre : « loopy belief propagation », Markov chain Monte Carlo (e.g . échantillonnage de Gibbs), Inférence variationnelle.

Markov chain Monte Carlo exploite la topologie du réseau et effectue un échantillonnage de Gibbs sur des sous-ensembles locaux de variables de façon séquentielle et concurrente. L'inférence variationnelle est une méthode de plus en plus utilisée, elle est une sorte d'adaptation de l'algorithme EM (Expectation-Maximization).

2.2.2 Résolution par apprentissage

Un deuxième axe de la recherche porte sur la **construction automatique des modèles**. Il s'agit d'un sujet fascinant. En effet, si l'on y réfléchit, la notion de probabilité conditionnelle s'applique aussi aux modèles. De ce point de vue, la théorie bayesienne offre une réponse à l'un des aspects les plus critiques de la modélisation empirique qui est la dialectique entre observations et modèle. Dans la démarche empirique, un modèle est inféré par un scientifique à partir d'un a priori théorique validé par des observations, nombreuses ou bien choisies. Si, dans une situation particulière, une observation vient à contredire le modèle, on est conduit invariablement à l'une ou l'autre des conclusions : soit, le plus souvent, l'observation est rejetée comme insuffisamment fiable, soit le modèle est remis en question.

Dans la mesure où les réseaux bayesiens ne sont pas déterministes mais probabilistes, ils tolèrent « l'erreur ». Un réseau bayesien ne fournit généralement pas de décision, mais seulement une conclusion probable. Si les faits le contredisent, il n'y a pas lieu de rejeter le modèle puisque la conclusion inverse était également possible, mais simplement moins probable. Cependant, si le modèle en vient à se tromper régulièrement, c'est-à-dire à donner souvent pour le plus probable un résultat qui n'est pas celui observé, on peut s'interroger sur sa validité. La théorie bayesienne s'applique parfaitement ici, en disant simplement que le modèle devient moins probable, par rapport à des modèles concurrents

Chapitre 3

Application du réseau bayésien

3.1 Contexte de l'application

Il s'agit d'une prédiction à partir d'une collection des données des cas d'une étude qui a été menée entre 1958 et 1970 à l'Université de l'hôpital Billings de Chicago sur la survie de patients opérés d'un cancer.

Les données contenues dans le dataset sont :

- L'âge du patient au moment de l'opération (numérique).
- Année d'intervention du patient (numérique).
- Nombre de noeuds axillaires positifs détectés (numérique).
- Statut de survie (attribut de classe) :
  → 1 = le patient a survécu 5 ans ou plus.
  → 2 = le patient est décédé dans les 5 ans.

3.2 Réalisation

3.2.1 Outils et librairies utilisés

librairies

- **NumPy**
  est une extension du langage de programmation Python, destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux. Plus précisément, cette bibliothèque logicielle libre et open source fournit de multiples fonctions permettant notamment de créer directement un tableau depuis un fichier ou au contraire de sauvegarder un tableau dans un fichier, et manipuler des vecteurs, matrices et polynômes.

- **Pandas**
  est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles.

— **Matplotlib**
est une bibliothèque du langage de programmation Python destinée à tracer et visualiser des données sous formes de graphiques4. Elle peut être combinée avec les bibliothèques python de calcul scientifique NumPy et SciPy.

— **Scikit-learn**
est une bibliothèque libre Python destinée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs2 notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche comme Inria3 et Télécom ParisTech. Elle comprend notamment des fonctions pour estimer des forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support. Elle est conçue pour s'harmoniser avec d'autres bibliothèques libres Python, notamment NumPy et SciPy.

Outils

— **Jupyter Notebook**
est une application web utilisée pour programmer dans plus de 40 langages de programmation, dont Julia, Python, R, Ruby ou encore Scala2. Jupyter est une évolution du projet IPython. Jupyter permet de réaliser des calepins ou notebooks, c'est-à-dire des programmes contenant à la fois du texte en markdown et du code en Julia, Python, R... Ces notebooks sont utilisés en science des données pour explorer et analyser des données.

3.2.2 Code source

Input

In [20]:
```
print(os.listdir("../input"))

names=['age', 'year_of_treatment', 'positive_lymph_nodes', 'survival_status_after_5_years']
c1,c2,c3,c4= np.loadtxt('../input/data.csv',unpack=True,delimiter = ',')
cancer_df = pd.read_csv('../input/data.csv', header=None, names=names)
print(cancer_df.head())

cancer_df.head()
cancer_df.tail()
```

```
['data.csv']
   age             ...              survival_status_after_5_years
0   30             ...                                         1
1   30             ...                                         1
2   30             ...                                         1
3   31             ...                                         1
4   31             ...                                         1

[5 rows x 4 columns]
```

Out[20]:

	age	year_of_treatment	positive_lymph_nodes	survival_status_after_5_years
301	75	62	1	1
302	76	67	0	1
303	77	65	3	1
304	78	65	1	2
305	83	58	2	2

FIGURE 3.1 – Affichage des données lues à partir du dataset

```
In [12]: print(cancer_df.describe())

                age  year_of_treatment  positive_lymph_nodes
count  306.000000         306.000000            306.000000
mean    52.457516          62.852941              4.026144
std     10.803452           3.249405              7.189654
min     30.000000          58.000000              0.000000
25%     44.000000          60.000000              0.000000
50%     52.000000          63.000000              1.000000
75%     60.750000          65.750000              4.000000
max     83.000000          69.000000             52.000000
```

FIGURE 3.2 – Statistiques et constatations sur l'input

— mean : la moyenne des valeurs.

— std : l'écart type de l'échantillon.

— min : minimum des valeurs dans le dataset.

— max : maximum des valeurs dans le dataset.

— 25%, 50%, 75% : pourcentage des données analysés à partir du dataset.

```
In [13]: print("Target variable distribution")
         print(cancer_df.iloc[:,-1].value_counts())
         print("-"*50)
         print(cancer_df.iloc[:,-1].value_counts(normalize = True))

         Target variable distribution
         yes    225
         no      81
         Name: survival_status_after_5_years, dtype: int64
         --------------------------------------------------
         yes    0.735294
         no     0.264706
         Name: survival_status_after_5_years, dtype: float64
```

FIGURE 3.3 – Statistiques sur les patients

L'âge des patients varie de 30 à 83 ans avec une médiane de 52 ans. Bien que le nombre maximum de ganglions lymphatiques positifs observés soit de 52, près de 75% des patients ont moins de 5 ganglions lymphatiques positifs et près de 25% des patients n'ont pas de ganglions lymphatiques positifs.
Le jeu de données ne contient qu'un petit nombre d'enregistrements (306). La colonne cible est déséquilibrée avec 73% des valeurs sont «oui».

```
In [26]: from sklearn.naive_bayes import GaussianNB
         clf = GaussianNB()
         clf.fit(x,y)
         predictions=clf.predict(x)

In [27]: from sklearn.metrics import accuracy_score
         accuracy_score(y,predictions)

Out[27]: 0.7483660130718954

 In [ ]: count = 0
         for i in range(0,len(predictions)):
             if predictions[i]==y[i]:
                 count+=1
             else:
                 pass
         accuracy = count/len(predictions)
         print(accuracy)
```

FIGURE 3.4 – Apprentissage

Chapitre 4

Conclusion

L'approche bayesienne propose une définition de la probabilité, ainsi que des méthodes statistiques, alternatives à l'approche fréquentiste, encore universellement répandue. C'est une théorie bien fondée donnant un sens à des usages communs de la notion de probabilité et ne se heurtant pas aux difficultés conceptuelles de l'approche fréquentiste.

Les réseaux bayésiens sont donc un outil de choix dans la représentation de connaissances et dans l'exploitation de celles-ci. Nous l'avons vu, beaucoup de domaines sont intéressés par ce type de représentation. Comme on a pu le voir, l'inférence sur les réseaux bayésiens est un problème NP-difficile, c'est pourquoi il était convenable de le voir de façon complète pour des instances réalisables et incomplète dans les autres cas.

Pour aller plus loin, il pourrait être intéressant de se pencher sur les réseaux bayésiens dynamiques. Ceux ci sont une répétition du réseau classique dans lesquels on rajoute un lien causal d'un pas de temps à l'autre. Ils contiennent chacun un certain nombre de variables aléatoires représentant les observations et les états cachés du processus. Le temps ici est discret et chaque unité de temps représente une nouvelle observation, l ?unité de temps n ?a donc pas toujours la même valeur en temps réel, la complexité inférencielle des réseaux bayésiens dynamiques est évidement bien plus élevée que celle vu précédemment.

Références

https://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html

https://perso.liris.cnrs.fr/amille/enseignements/Master_PRO/TIA/RBayesiens/Intro_RB.pdf

https://www.lri.fr/~antoine/Courses/IIE/WEB-ISX/Tr-SE-3-4.pdf

http://www-clips.imag.fr/geod/User/jean.caelen/Publis_fichiers/Dossier%20Centraliens%20n%C2%B0%205931.pdf