UNIVERSITÉ CÔTE D'AZUR

Modélisation de l'incertitude (M2 MIAGE IA²)

Andrea G. B. Tettamanzi
Laboratoire I3S – Équipe SPARKS
andrea.tettamanzi@univ-cotedazur.fr

univ-cotedazur.fr

Séance 2
Raisonnement probabiliste
Réseaux bayésiens

Dans cette séance

- Classification bayésienne
- Indépendance conditionnelle
- Réseaux bayésiens

Classification bayésienne

- **Un classificateur statistique** : il effectue une prédiction probabiliste, c'est-à-dire qu'il prédit les probabilités d'appartenance à une classe
- **Fondation** : Basée sur le théorème de Bayes.
- **Performance** : Un classificateur bayésien simple, le classificateur naïf de Bayes, a des performances comparables à celles des arbres de décision et de certains classificateurs basés sur les réseaux de neurones
- **Incrémentale** : chaque exemple d'entraînement peut augmenter/diminuer progressivement la probabilité qu'une hypothèse soit correcte – les connaissances préalables peuvent être combinées avec les données observées
- **Standard** : Même lorsque les méthodes bayésiennes sont difficiles à calculer, elles peuvent fournir une norme de prise de décision optimale par rapport à laquelle d'autres méthodes peuvent être mesurées

Rappels sur le théorème de Bayes

- Soit X un tuple de données : l’étiquette de classe est inconnue
- Soit H l’hypothèse que X appartienne à la classe C
- La classification consiste à déterminer P(H | X), la probabilité que l'hypothèse soit correcte, étant donné le tuple de données observées X
- P(H) (probabilité *a priori*), la probabilité initiale
  - Par exemple, X achètera un ordinateur, quels que soient son âge, ses revenus, etc.
- P(X) : probabilité que les données du tuple soient observées
- P(X | H) (probabilité *a posteriori*), la probabilité d'observer le tuple X, étant donné que l'hypothèse H est correcte
  - Par exemple, étant donné que X va acheter un ordinateur, la probabilité que X soit de 31..40, revenu moyen, etc.

Théorème de Bayes

- Étant données les données d'entraînement **X**, la probabilité a posteriori de l'hypothèse *H*, P(*H* | **X**), suit la formule de Bayes

$$P(H \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid H)P(H)}{P(\mathbf{X})}$$

Informellement, ceci peut s'écrire comme

a posteriori = vraisemblance x a priori / évidence

- On prédit que **X** appartienne a C_i ssi la probabilité $P(C_i \mid \mathbf{X})$ est la plus haute parmi toutes les $P(C_k \mid \mathbf{X})$ pour toutes le *k* classes
- Difficulté pratique : cela réquiert une connaissance initiale de beaucoup de probabilités, ce qui signifie un coût de calcul

Classification bayesienne naïve

- Soit D un jeu de données d'entraînement de tuples avec leurs étiquettes de classe, et soit chaque tuple représenté par un vecteur de n attributs **X** = (x1, x2, …, xn)
- Supposons qu'il y ait *m* classes C_1, C_2, …, C_m.
- La classification se fait sur la base de la plus grande probabilité a posteriori, c-à-d, du $P(C_i \mid \mathbf{X})$ maximum

  On peut calculer ceci grâce au Théorème de Bayes :

$$P(C_i \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid C_i)P(C_i)}{P(\mathbf{X})}$$

- Puisque P(**X**) est constante pou toute classe, il suffit juste de maximiser

$$P(\mathbf{X} \mid C_i)P(C_i)$$

Classificateur naïf bayésien

- Simplification : on suppose que les attributs soit conditionnellement indépendant :

$$P(\mathbf{X} \mid C_i) = \prod_{k=1}^{n} P(X_k \mid C_i)$$

- Ceci réduit beaucoup la complexité des calculs : il suffit juste de prendre en compte la distribution de chaque classe
- Si A_k est catégorielle, $P(X_k \mid C_i)$ est le nombre de tuples de C_i ayant la valeur X_k pour A_k, divisé par $|C_{i,D}|$ (nombre de tuples de C_i en D)
- Si A_k est à valeurs continues, $P(X_k \mid C_i)$ est généralement calculé sur la base d'une distribution gaussienne avec une moyenne μ et un écart-type σ et $P(x_k \mid C_i)$ est

$$\mathcal{N}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$P(\mathbf{X} \mid C_i) = \mathcal{N}(X_k; \mu_{C_i}, \sigma_{C_i})$$

Classificateur naïf bayésien : données entraînement

Classes :
C1:buys_computer = ‘yes’
C2:buys_computer = ‘no’

Tuple donné :
X = (age <=30,
Income = medium,
Student = yes
Credit_rating = Fair)

age	income	student	credit_rating	_comp
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Classificateur naïf bayésien : exemple

- $P(C_i)$: P(buys_computer = “yes”) = 9/14 = 0.643
  P(buys_computer = “no”) = 5/14= 0.357

- Compute $P(X|C_i)$ for each class
  P(age = “<=30” | buys_computer = “yes”) = 2/9 = 0.222
  P(age = “<= 30” | buys_computer = “no”) = 3/5 = 0.6
  P(income = “medium” | buys_computer = “yes”) = 4/9 = 0.444
  P(income = “medium” | buys_computer = “no”) = 2/5 = 0.4
  P(student = “yes” | buys_computer = “yes) = 6/9 = 0.667
  P(student = “yes” | buys_computer = “no”) = 1/5 = 0.2
  P(credit_rating = “fair” | buys_computer = “yes”) = 6/9 = 0.667
  P(credit_rating = “fair” | buys_computer = “no”) = 2/5 = 0.4

- **X = (age <= 30 , income = medium, student = yes, credit_rating = fair)**

$P(X|C_i)$: P(X|buys_computer = “yes”) = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
P(X|buys_computer = “no”) = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

$P(X|C_i)*P(C_i)$: P(X|buys_computer = “yes”) * P(buys_computer = “yes”) = 0.028
P(X|buys_computer = “no”) * P(buys_computer = “no”) = 0.007

Therefore, X belongs to class (“buys_computer = yes”)

Problème de la probabilité nulle

- La classification naïve bayésienne requiert que chaque probabilité conditionnelle soit non-nulle. Sinon, la probabilité prédite sera nulle !

$$P(\mathbf{X} \mid C_i) = \prod_{k=1}^{n} P(X_k \mid C_i)$$

- Exemple : soit un jeu de données de 1000 tuples, avec income=low (0), income= medium (990), et income = high (10),
- On utilise la correction laplacienne
  - On ajoute 1 à chaque cas
    - Pr(income = low) = 1/1003
    - Pr(income = medium) = 991/1003
    - Pr(income = high) = 11/1003
  - Ainsi, les estimations de probabilité « corrigées » sont proches de leur contreparties « brutes » ; en outre, elle ne sont jamais nulles

Remarques sur Naïve Bayes

- Avantages
  - Facile à coder
  - De bons résultats dans la plupart des cas
- Inconvénients
  - Hypothèse de fond : indépendance statistique des classes, ce qui rarement est vrai, et qui donc entraîne une perte de précision
  - Dans la réalité, il existe des dépendances entre les variables
    - Exemples : patients d’un hôpital : âge, familiarité, etc.
    - Symptômes: fièvre, tous, etc. ; Maladie : tumeur, diabète, etc.
    - Ces dépendances ne peuvent pas être modélisées par un classificateur naïf bayésien
- Comment modéliser ces dépendances ?
  - Avec les **réseaux bayésiens**

Indépendance

- Deux variables aléatoires sont (absolument) indépendantes ssi

$$P(A \mid B) = P(A)$$

ou $$P(A, B) = P(A \mid B)P(B) = P(A)P(B)$$

  - Exemple : deux lances d'une pièce

- Si *n* v.a. booléennes sont indépendantes la distribution conjointe complète est

$$P(X_1, \ldots, X_n) = \prod_i P(X_i)$$

- Elle a donc *n* degrés de liberté

Indépendance conditionnelle

- Considérons les trois v.a. (booléennes)
  - Covid, PCR, Dyspnée
- La distribution conjointe a $2^3 = 8$ évts élémentaires
- Si un patient a la Covid, la probabilité qu'il ait une dyspnée ne dépend pas du résultat du test PCR :

$$P(Dyspnée \mid Covid, PCR) = P(Dyspnée \mid Covid)$$

  - Dyspnée est conditionnellement indépendante de PCR
- De même, s'il n'a pas la Covid,

$$P(Dyspnée \mid \neg Covid, PCR) = P(Dyspnée \mid \neg Covid)$$

Indépendance conditionnelle

- Des propositions équivalentes sont

$$P(PCR \mid Covid, Dyspnée) = P(PCR \mid Covid)$$

$$P(PCR, Dyspnée \mid Covid,) = P(PCR \mid Covid)P(Dyspnée \mid Covid)$$

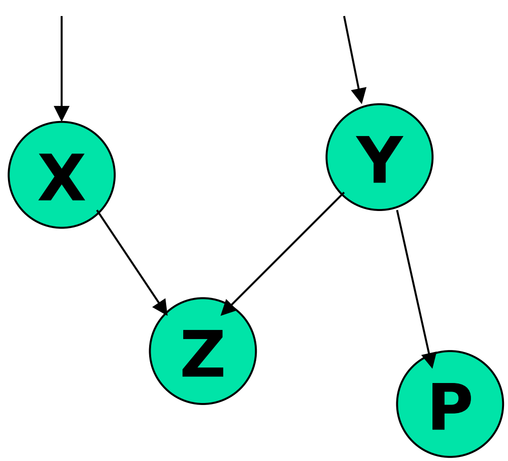
- La distribution conjointe peut donc s'écrire

$$\begin{aligned} &P(PCR, Dyspnée, Covid) = \\ &\quad = P(PCR, Dyspnée \mid Covid)P(Covid) \\ &\quad = P(PCR \mid Covid)P(Dyspnée \mid Covid)P(Covid) \end{aligned}$$

- 2 + 2 + 1 = 5 degrés de liberté

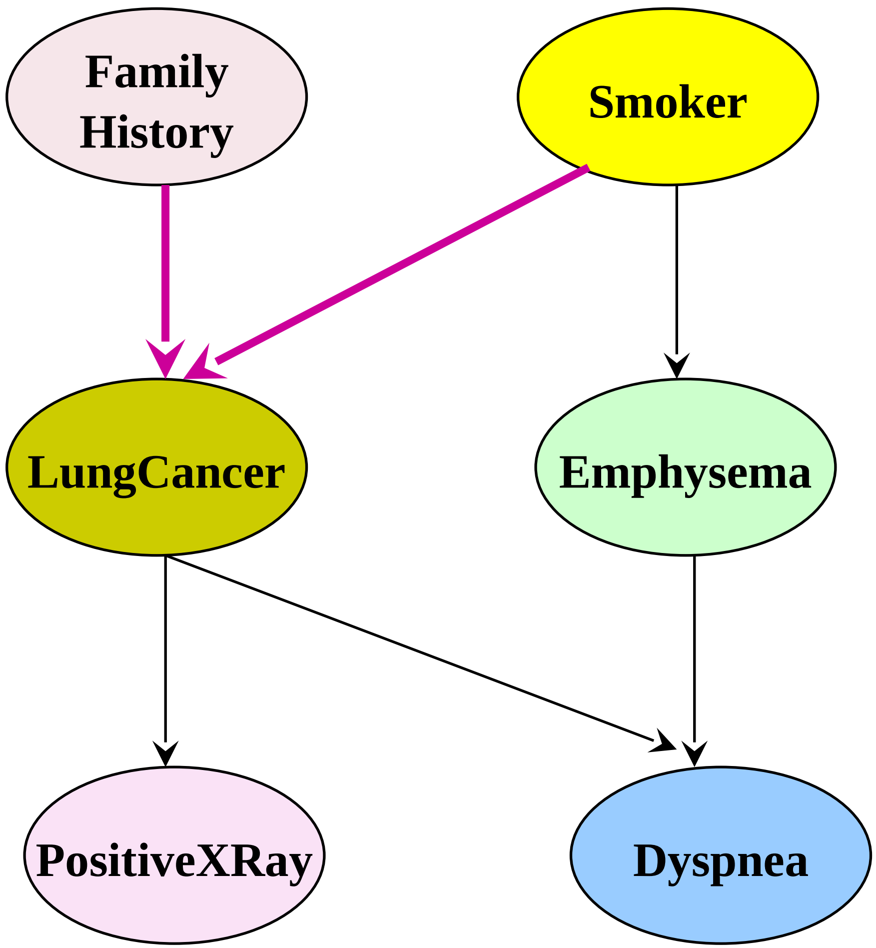
Réseaux bayésiens

- Le réseau bayésiens (ou de croyance) permettent une spécification concise d'une distribution de probabilité conjointe en prenant en compte l'indépendance conditionnelle
- Un modèle graphique des relations causales
  - Les arcs représentent les <u>dépendances</u> entre variables



- ❑ Nœuds: variables aléatoires
- ❑ Arcs: dépendance
- ❑ X et Y sont les parents de Z, Y est le parent de P
- ❑ Z et P sont indépendants
- ❑ Graphe acyclique

Exemple



La **table de probabilité conditionnelle** (**TPC**) de la v.a. LungCancer :

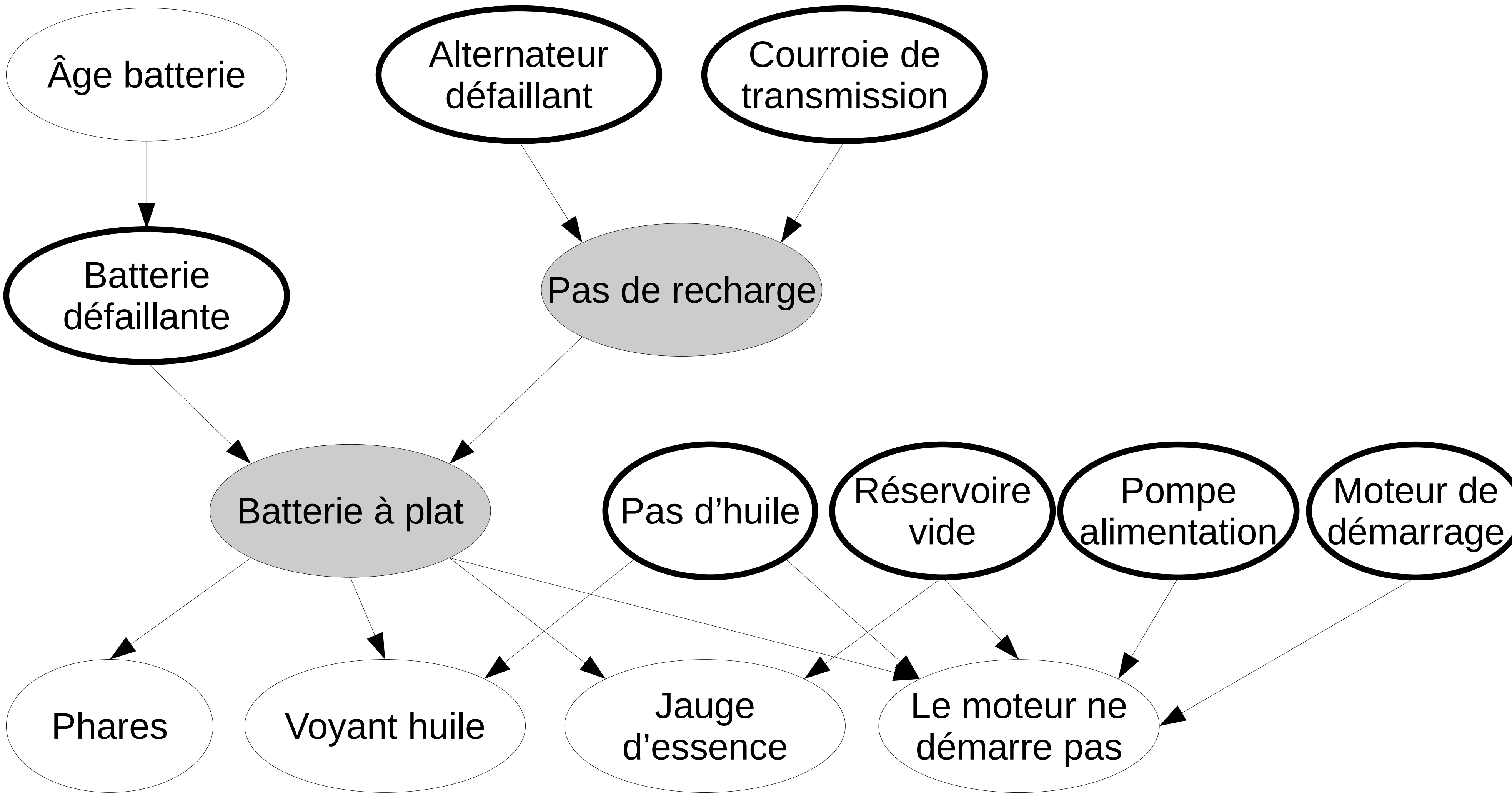
	(FH, S)	(FH, ~S)	(~FH, S)	(~FH, ~S)
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

La TPC donne la probabilité conditionnelle pour chaque combinaison des valeurs de ses parents

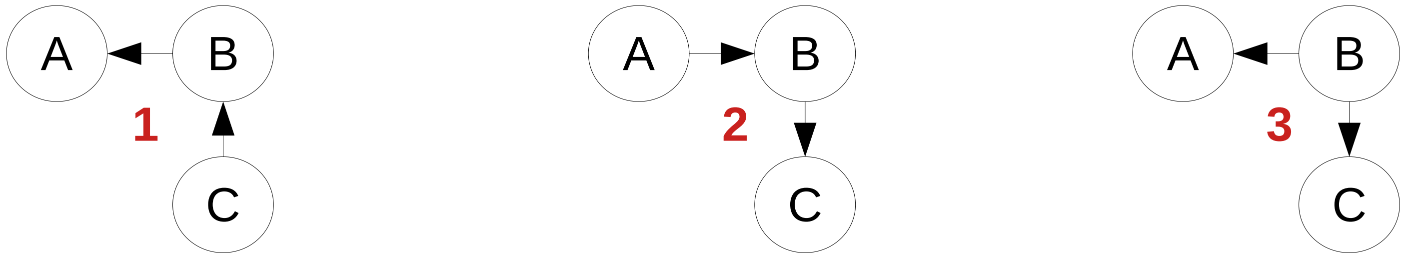
Calcul de la probabilité conjointe de **X**, à partir des TPC :

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid Parents(Y_i))$$

Exemple avec variables cachées



Dépendance ≠ causalité



$$P_1(A,B,C) = P(A \mid B)P(B \mid C)P(C)$$

$$\begin{aligned} P_2(A,B,C) &= P(A) \cdot P(B \mid A) \cdot P(C \mid B) \\ &= P(A) \cdot \frac{(P(A|B)P(B)}{P(A)} \cdot \frac{P(B|C)P(C)}{P(B)} \\ &= P(A \mid B) \cdot P(B \mid C)P(C) = P_1(A,B,C) \end{aligned}$$

$$\begin{aligned} P_3(A,B,C) &= P(A \mid B) \cdot P(B) \cdot P(C \mid B) \\ &= P(A \mid B) \cdot P(B) \cdot \frac{P(B|C)P(C)}{P(B)} \\ &= P(A \mid B) \cdot P(B \mid C)P(C) = P_1(A,B,C) \end{aligned}$$

Apprentissage de RB

- Plusieurs scénarios :
  - Structure donnée, toutes les variables observables: on apprend juste les TPCs
  - Structure connue, mais quelques variables cachées : descente du gradient, comme pour les réseaux de neurones
  - Structure non connue, variables observables : recherche dans l’espace des structures
  - Structure non connue, variables cachées : pas d’algorithmes efficaces connus