ANALYSE ET VISUALISATION DES DONNÉES

Alexandra-Maria DOBRESCU

| Cours 7 |

Collection et préparation des données

AVD | Cours 7 | AMD

- SOURCES ET ACQUISITION DES DONNÉES
- NETTOYAGE ET PRÉTRAITEMENT DES DONNÉES
- TYPES DE DONNÉES
- LA MESURE

Discrétisation des données

ALGORITHMES

Apprentissage par Réseau Bayésien
Bayesian Network Learning

UN RÉSEAU BAYÉSIEN FAIT PARTIE DE LA CATÉGORIE DES TECHNIQUES DE MODÉLISATION GRAPHIQUE PROBABILISTE (MGP) QUI SONT UTILISÉES POUR CALCULER LES INCERTITUDES À L'AIDE DU CONCEPT DE PROBABILITÉ.

Remarque : Plus connus sous le nom de Réseaux de Croyance (*Belief Networks)*, les réseaux bayésiens sont utilisés pour modéliser les incertitudes à l'aide de graphes acycliques dirigés (*Directed Acyclic Graphs, DAG*).

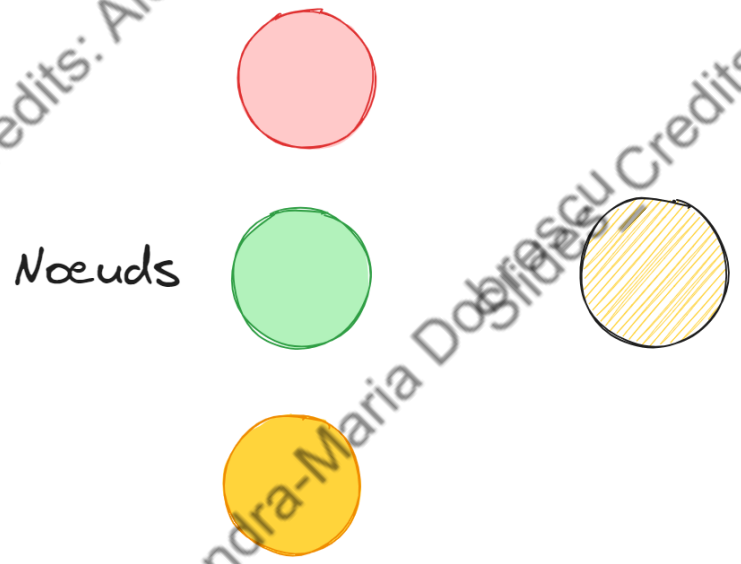
Exemple : Lors de l'apprentissage de réseaux bayésiens, les attributs continus peuvent être discrétisés afin d'estimer plus efficacement les probabilités conditionnelles.

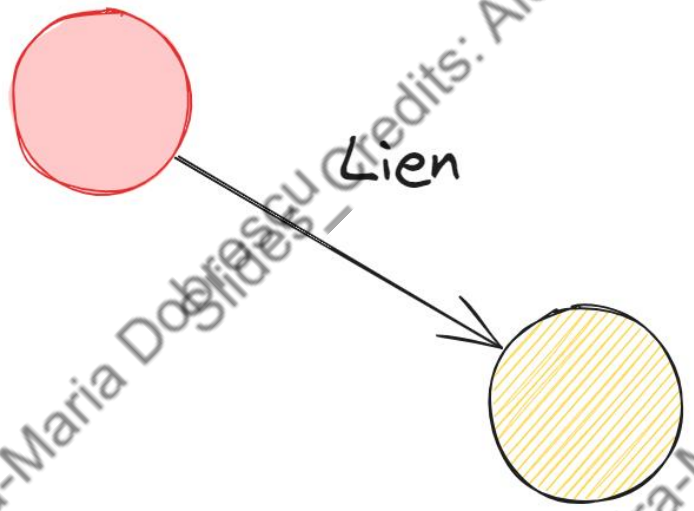
Apprentissage par Réseau Bayésien

ALGORITHMES

Graphique acyclique dirigé
Directed Acyclic Graph (DAG)

UN GRAPHE ACYCLIQUE DIRIGÉ EST UTILISÉ POUR REPRÉSENTER UN RÉSEAU BAYÉSIEN ET, COMME TOUT AUTRE GRAPHE STATISTIQUE, UN DAG CONTIENT UN ENSEMBLE DE NŒUDS ET DE LIENS, LES LIENS INDIQUANT LA RELATION ENTRE LES NŒUDS.

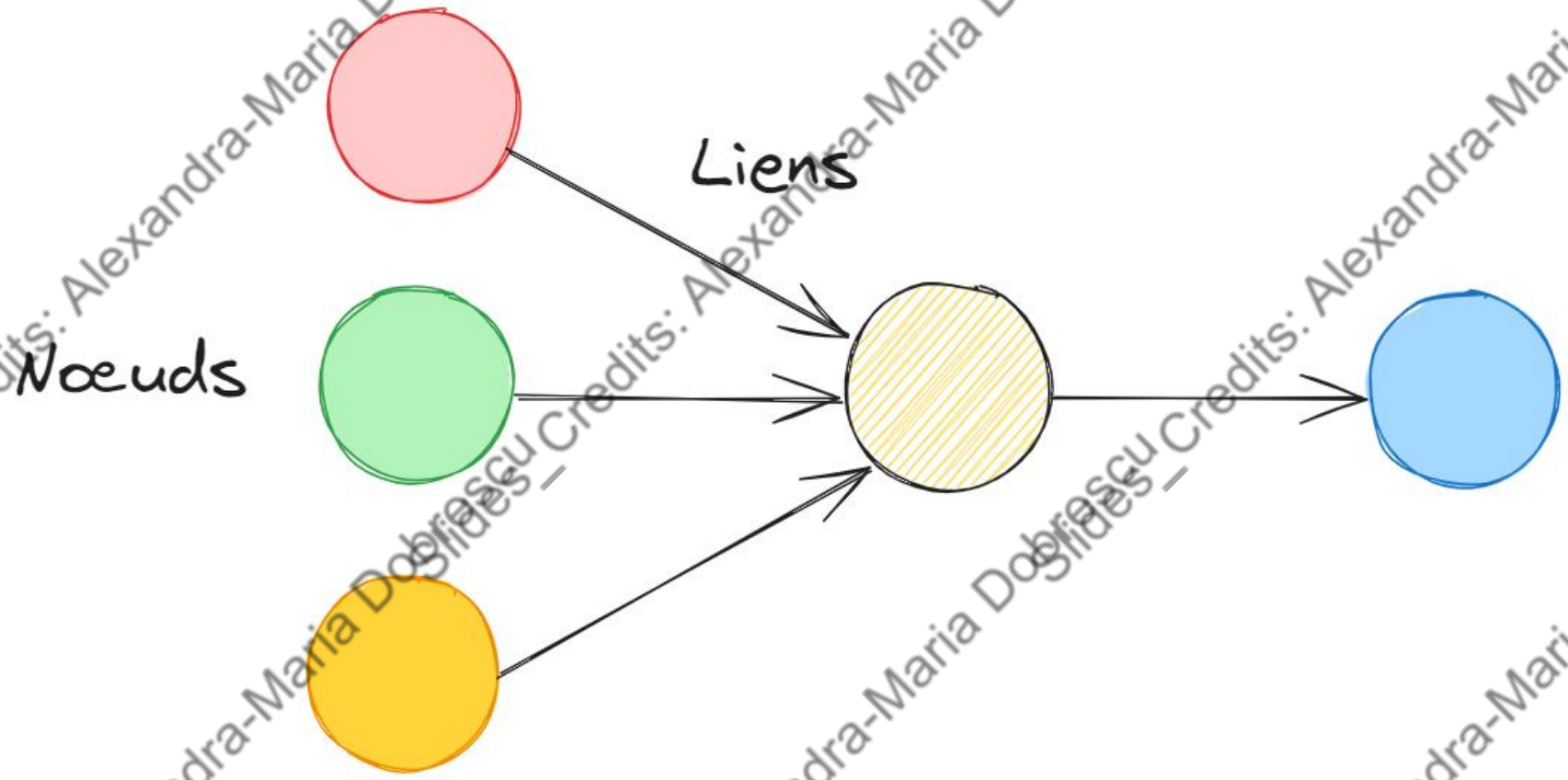




Apprentissage par Réseau Bayésien

ALGORITHMES

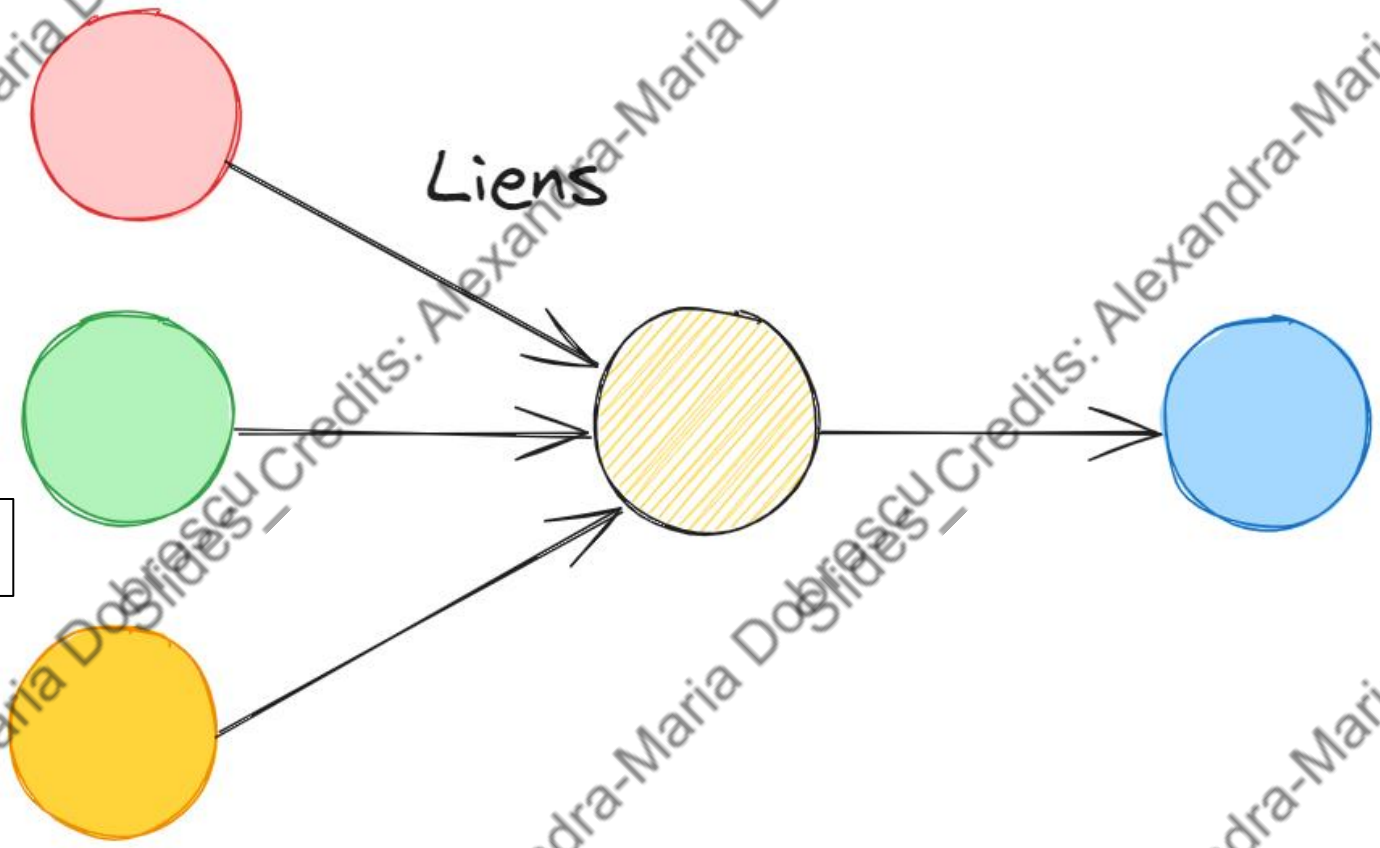
Graphique acyclique dirigé
Directed Acyclic Graph (DAG)



Apprentissage par Réseau Bayésien

ALGORITHMES

Graphique acyclique dirigé
Directed Acyclic Graph (DAG)



les variables aléatoires

Apprentissage par Réseau Bayésien

ALGORITHMES

Variables aléatoires
Random Variables

Exemple : Lançons trois pièces de monnaie et comptons le nombre de têtes. On a:

{TTT, THH, HTH, HHT, HTT, THT, TTH, HHH}, où H correspond à Heads (Face) et T à Tails (Pile).

Remarque 1:

- En général, la variable aléatoire est notée ***X*** en majuscule.
- Avec la lettre minuscule ***x***, nous notons les valeurs possibles.

Remarque 2: ***X*** est une variable aléatoire discrète car ses valeurs possibles peuvent être comptées comme des nombres entiers et les résultats sont aléatoires.

Apprentissage par Réseau Bayésien

ALGORITHMES

Types de variables aléatoires

CES VARIABLES SONT CLASSÉES EN FONCTION DE LEUR DISTRIBUTION DE PROBABILITÉ.

Une variable aléatoire est liée soit à une **Distribution de Probabilité** (variable aléatoire discrète), soit à une fonction de **Densité de Probabilité** (variable aléatoire continue).

variable aléatoire discrète

Comment cela se présente-t-il ?

variable aléatoire continue

Comment cela se présente-t-il ?

Apprentissage par Réseau Bayésien

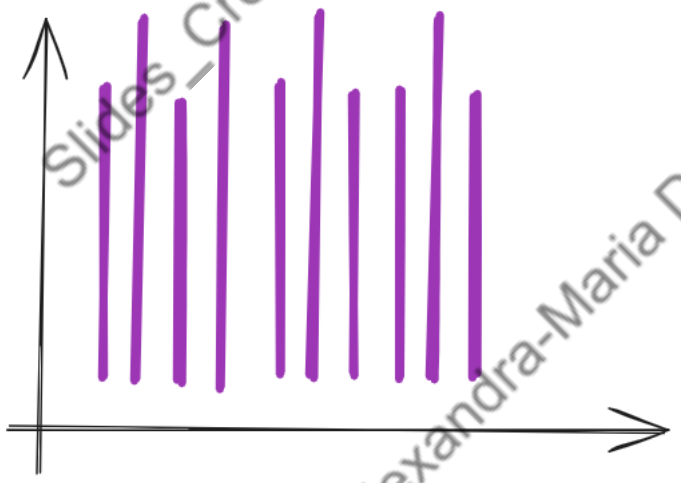
ALGORITHMES

Types de variables aléatoires

CES VARIABLES SONT CLASSÉES EN FONCTION DE LEUR DISTRIBUTION DE PROBABILITÉ.

Une variable aléatoire est liée soit à une **Distribution de Probabilité** (variable aléatoire discrète), soit à une fonction de **Densité de Probabilité** (variable aléatoire continue).

variable aléatoire discrète

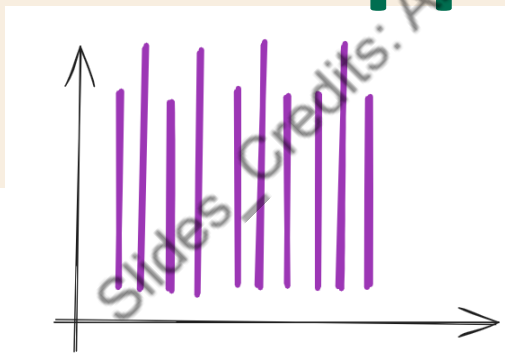


variable aléatoire continue

Apprentissage par Réseau Bayésien

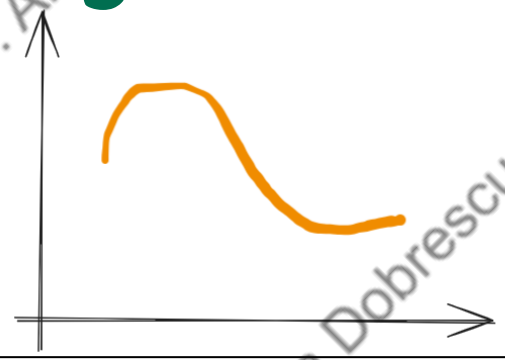
ALGORITHMES

Types de variables aléatoires



variable aléatoire discrète

- Une variable aléatoire discrète est une variable aléatoire dont **les résultats sont comptés**.
- X ne peut prendre qu'un ensemble de nombres distincts (0,1,2,...), c'est-à-dire un nombre fini de valeurs différentes.
- Les variables aléatoires discrètes sont impliquées dans les Distributions Binomiale et de Poisson.



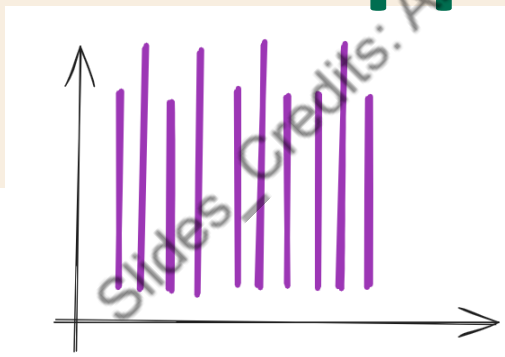
variable aléatoire continue

- Une variable aléatoire continue est une variable aléatoire dont **les résultats sont mesurés**.
- Elle peut prendre n'importe quelle valeur dans un intervalle. Lorsque les valeurs possibles d'une variable aléatoire forment un intervalle, on parle de variable aléatoire continue.
- Étant donné [a, z) , nous appelons une variable aléatoire X une variable aléatoire continue, si elle peut prendre n'importe quelle valeur dans cet intervalle.

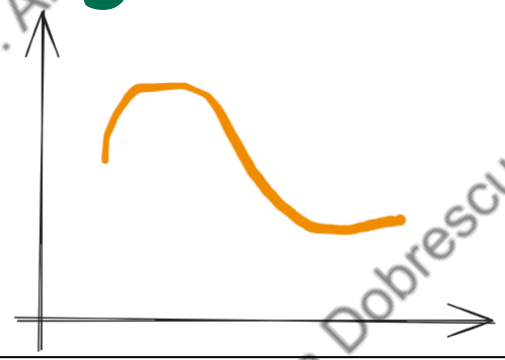
Apprentissage par Réseau Bayésien

ALGORITHMES

Types de variables aléatoires



? Distribution ?



variable aléatoire discrète

- Une variable aléatoire discrète est une variable aléatoire dont **les résultats sont comptés**.
- X ne peut prendre qu'un ensemble de nombres distincts (0,1,2,...), c'est-à-dire un nombre fini de valeurs différentes.
- Les variables aléatoires discrètes sont impliquées dans les Distributions Binomiale et de Poisson.

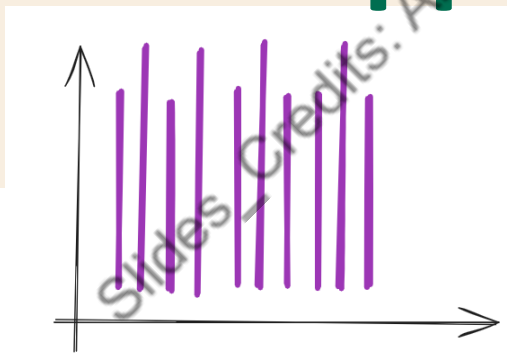
variable aléatoire continue

- Une variable aléatoire continue est une variable aléatoire dont **les résultats sont mesurés**.
- Elle peut prendre n'importe quelle valeur dans un intervalle. Lorsque les valeurs possibles d'une variable aléatoire forment un intervalle, on parle de variable aléatoire continue.
- Étant donné [a, z) , nous appelons une variable aléatoire X une variable aléatoire continue, si elle peut prendre n'importe quelle valeur dans cet intervalle.

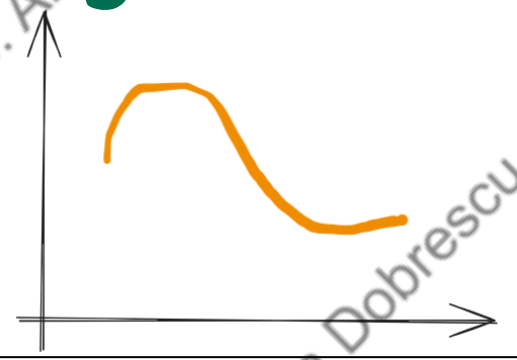
Apprentissage par Réseau Bayésien

ALGORITHMES

Types de variables aléatoires



variable aléatoire discrète



variable aléatoire continue

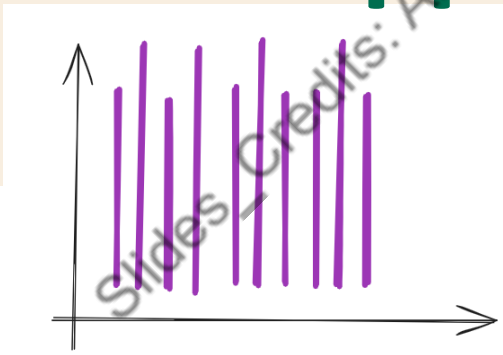
Plus précis (exemples) ?

Plus précis (exemples) ?

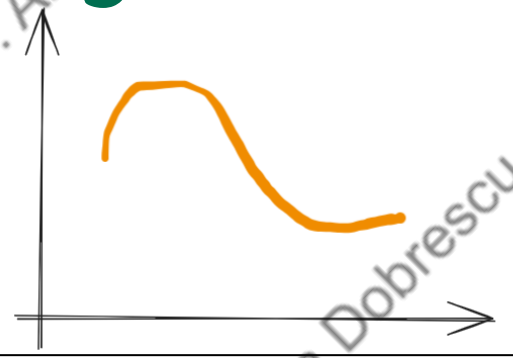
Apprentissage par Réseau Bayésien

ALGORITHMES

Types de variables aléatoires



variable aléatoire discrète



variable aléatoire continue

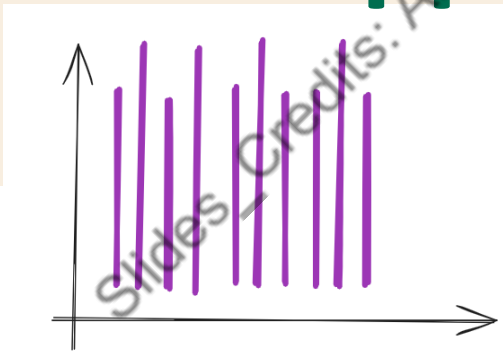
- Père Noël secret : nombre de collègues dans une équipe. 0.256 ?
- Nombre de têtes en lançant trois pièces
- Nombre d'élèves dans une classe, dans une famille, à la salle de sport, etc...

- Durée des trajets domicile-travail, mesurée sous la forme d'un ensemble de nombres réels
- Les scores de IQ
- La taille d'une personne est une variable aléatoire continue.
- Somme d'argent

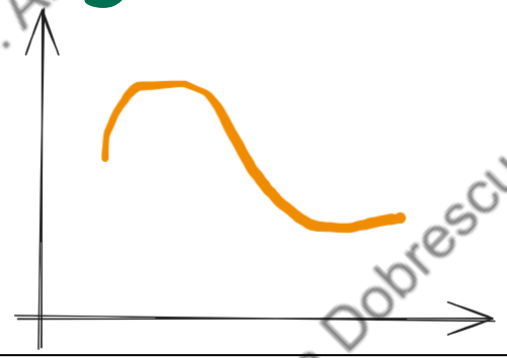
Apprentissage par Réseau Bayésien

ALGORITHMES

Types de variables aléatoires



variable aléatoire discrète



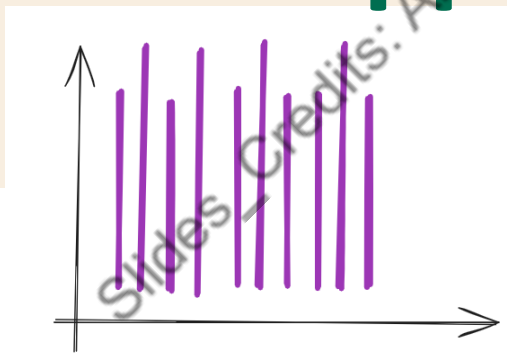
variable aléatoire continue

Distribution de probabilité: La distribution de probabilité d'une variable aléatoire discrète est décrite par une fonction de masse de probabilité (PMF), qui attribue des probabilités à chaque valeur possible.

Distribution de probabilité: La distribution de probabilité d'une variable aléatoire continue est décrite par une fonction de densité de probabilité (PDF). Contrairement aux variables aléatoires discrètes, la probabilité en un point donné est généralement nulle et les probabilités sont définies sur des intervalles.

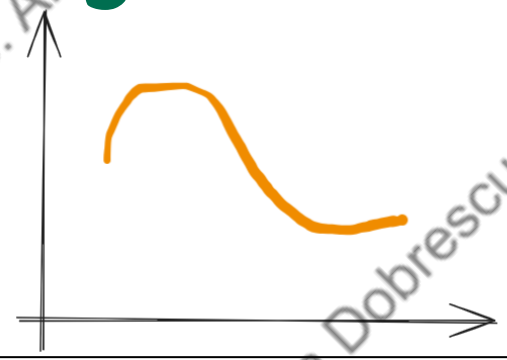
Apprentissage par Réseau Bayésien

ALGORITHMES



variable aléatoire discrète

Types de variables aléatoires



variable aléatoire continue

Les variables aléatoires mixtes, où une variable aléatoire peut être une combinaison de composantes discrètes et continues.
Il est essentiel de comprendre la nature de la variable aléatoire pour effectuer des analyses statistiques et des prédictions probabilistes dans diverses applications.

Apprentissage par Réseau Bayésien

ALGORITHMES

Types de variables aléatoires

variable aléatoire mixte

Exemple : Notes d'examen avec points bonus.

Considérons un scénario dans lequel les étudiants passent un examen et dont les notes sont composées de deux éléments :

1) Composante discrète (examen principal):

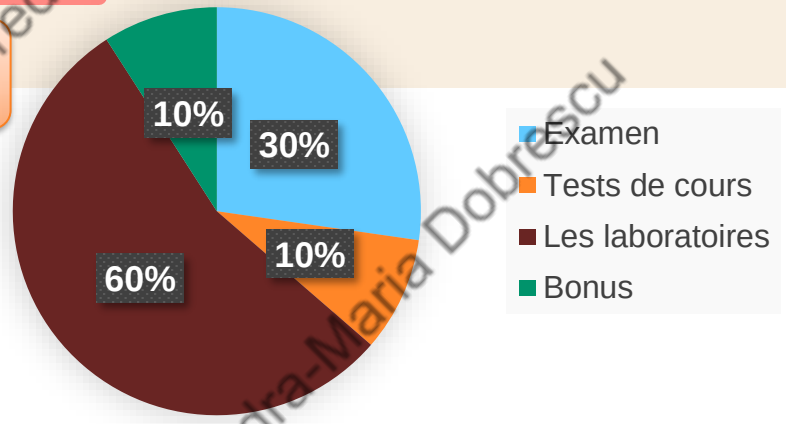
L'examen principal est noté sur une échelle discrète, telle que des nombres entiers de 0 à 10. Chaque nombre entier représente une note possible et la probabilité d'obtenir une note spécifique est donnée par une fonction de masse de probabilité (FMP).

Apprentissage par Réseau Bayésien

ALGORITHMES

Types de variables aléatoires

variable aléatoire mixte



Exemple : Notes d'examen avec points bonus.

Considérons un scénario dans lequel les étudiants passent un examen et dont les notes sont composées de deux éléments :

1) Composante discrète (examen principal):

L'examen principal est noté sur une échelle discrète, telle que des nombres entiers de 0 à 10. Chaque nombre entier représente une note possible et la probabilité d'obtenir une note spécifique est donnée par une fonction de masse de probabilité (FMP).

Apprentissage par Réseau Bayésien

ALGORITHMES

Types de variables aléatoires

variable aléatoire mixte

2) Composante continue (points bonus):

Les étudiants peuvent gagner des points bonus en effectuant des devoirs supplémentaires facultatifs, en prenant des cours de tests aléatoires ou en corrigeant des problèmes de laboratoire.
Les points bonus se situent sur une échelle continue, ce qui permet de prendre en compte n'importe quel nombre réel. La probabilité d'obtenir une valeur spécifique de points bonus est donnée par une fonction de densité de probabilité (PDF).

Apprentissage par Réseau Bayésien

ALGORITHMES

Types de variables aléatoires

variable aléatoire mixte

3) Variable aléatoire mixte: Dénotons le score total par ***X***, où ***X*** est la somme de la note de l'examen principal et des points bonus.
La variable aléatoire ***X*** est mixte car elle comporte à la fois des composantes discrètes et continues.

$$X = Examen\ Principal + Points\ Bonus$$

Takeaway:

- La note obtenue à l'examen principal est une variable aléatoire discrète dotée d'une fonction de masse de probabilité (PMF).
- Les points bonus sont une variable aléatoire continue avec une fonction de densité de probabilité (PDF).

Apprentissage par Réseau Bayésien

ALGORITHMES

Types de variables aléatoires

variable aléatoire mixte

Comment cela se présente-t-il ?

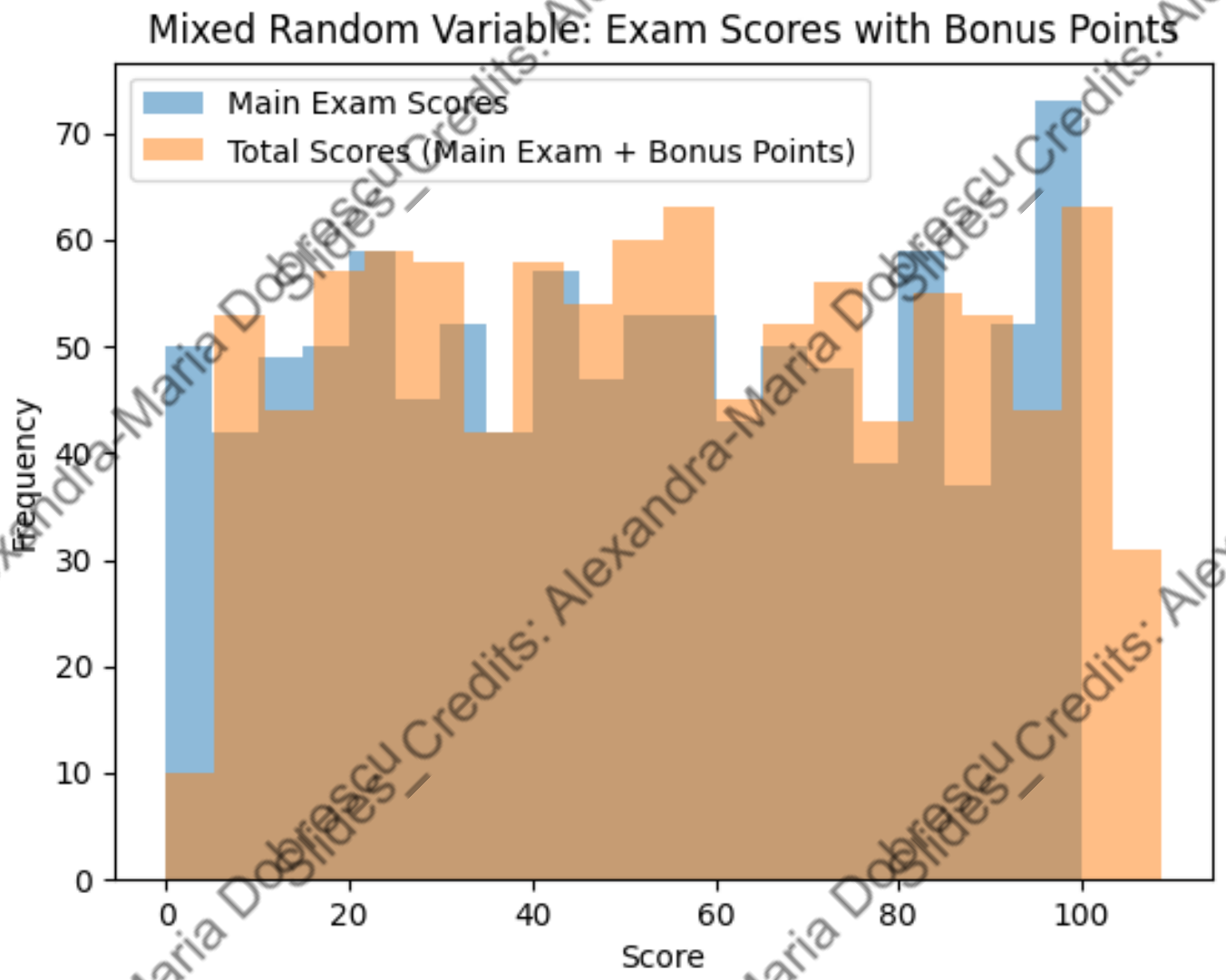
3) Variable aléatoire mixte: Dénotons le score total par ***X***, où ***X*** est la somme de la note de l'examen principal et des points bonus.
La variable aléatoire ***X*** est mixte car elle comporte à la fois des composantes discrètes et continues.

$$X = Examen\ Principal + Points\ Bonus$$

Takeaway:

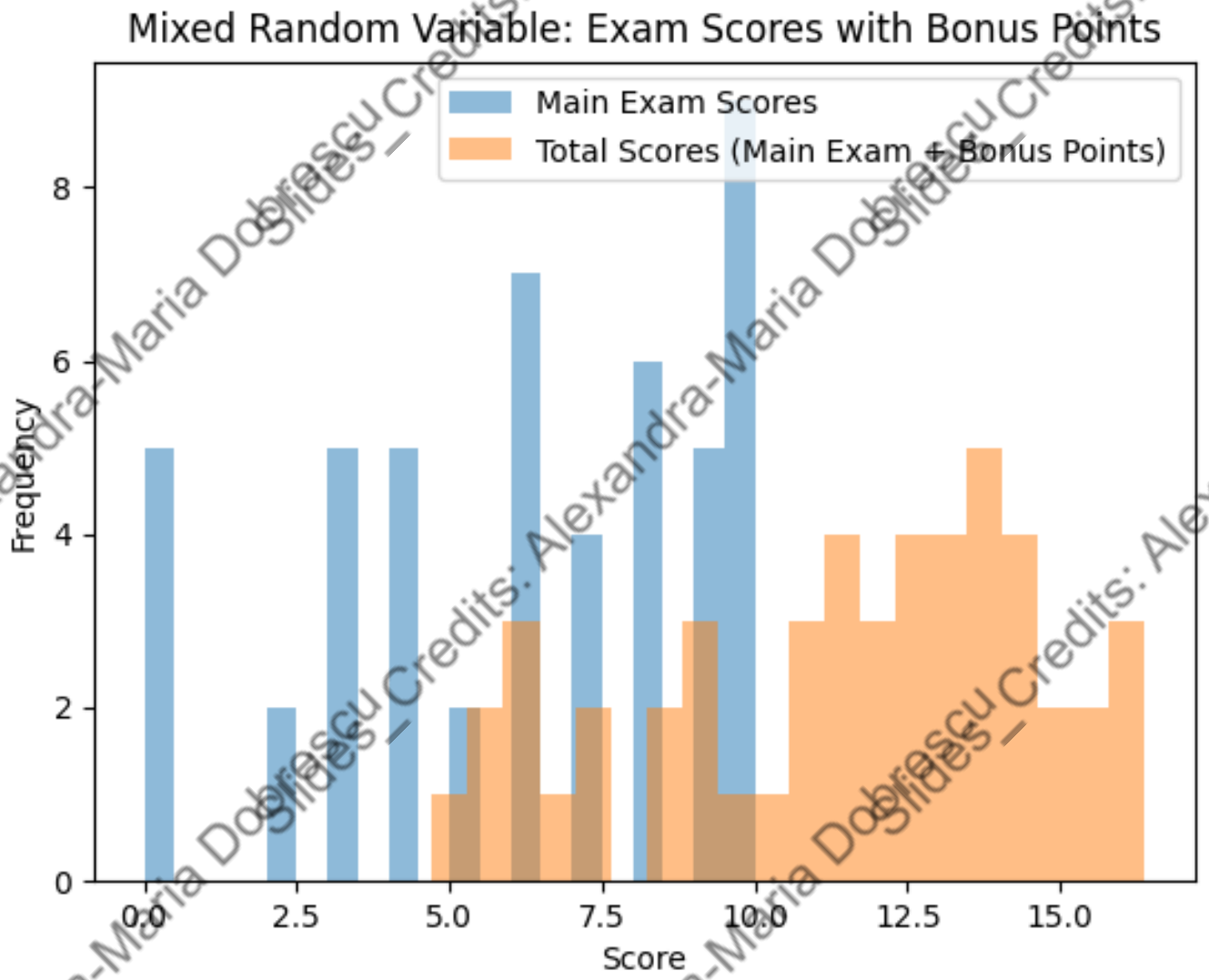
- La note obtenue à l'examen principal est une variable aléatoire discrète dotée d'une fonction de masse de probabilité (PMF).
- Les points bonus sont une variable aléatoire continue avec une fonction de densité de probabilité (PDF).

1000 notes pour l'examen principal



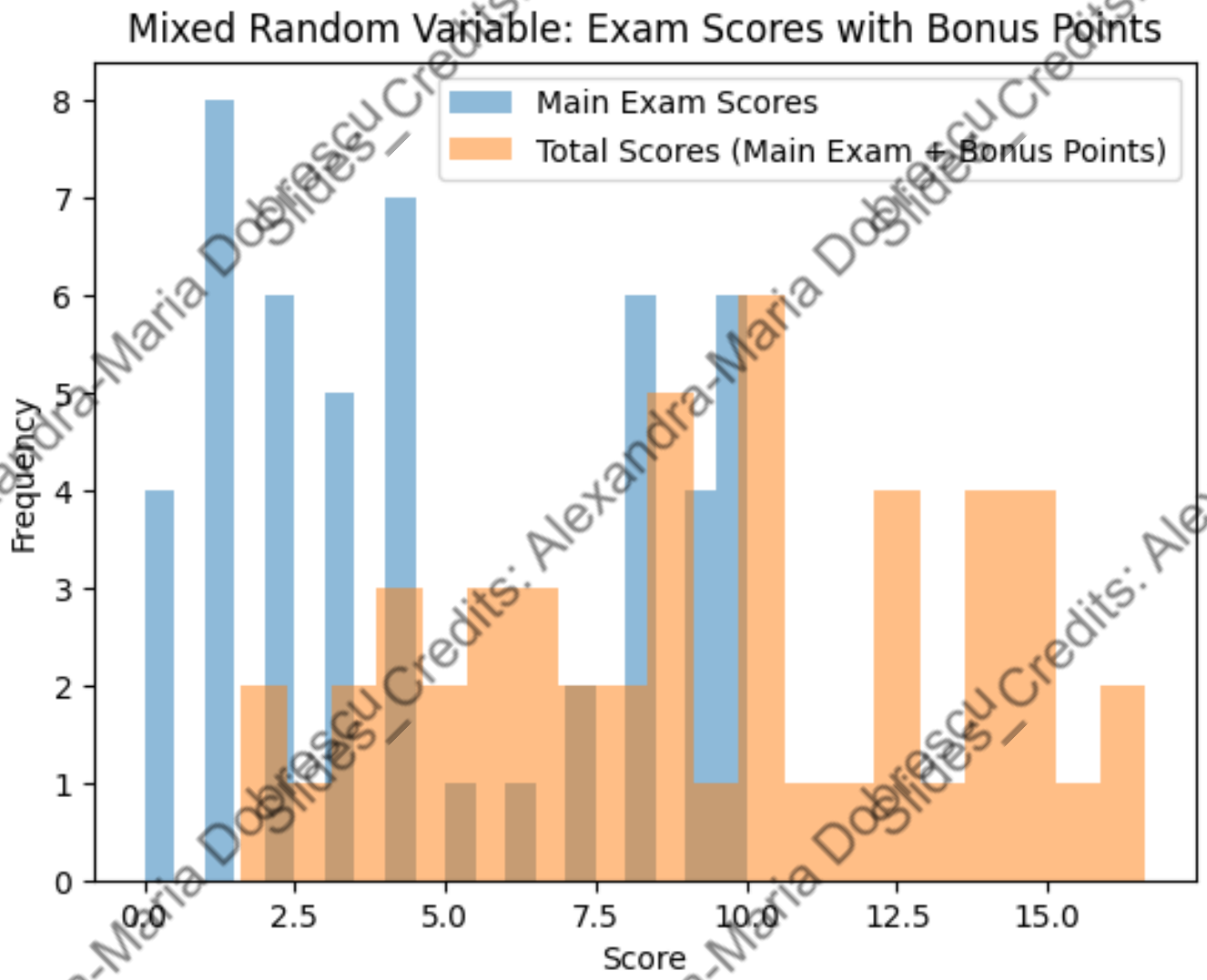
1000 points bonus

50 notes pour l'examen principal



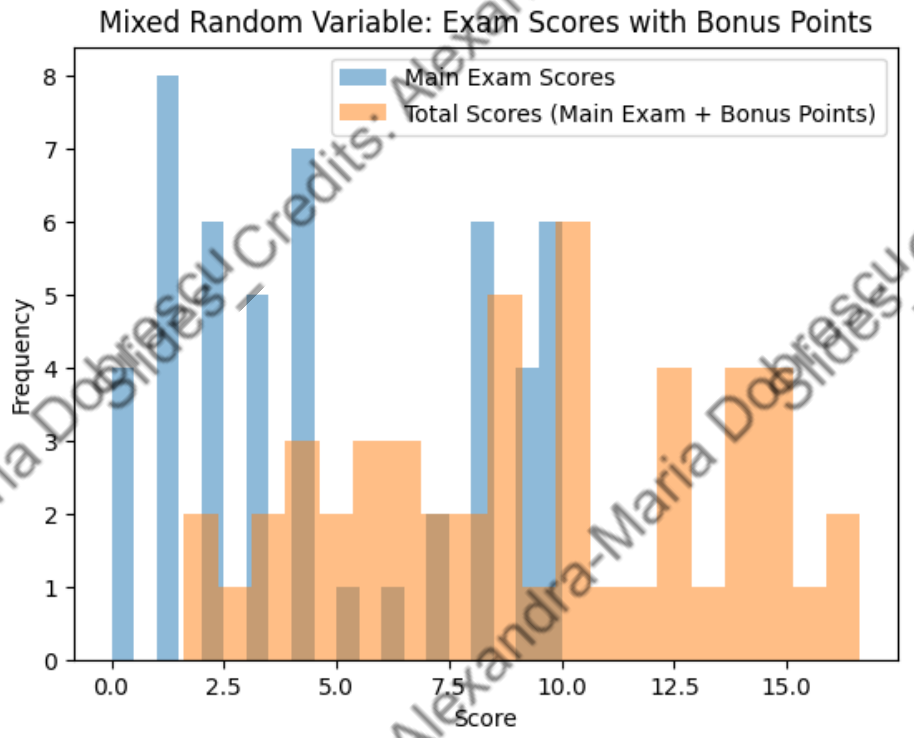
50 points bonus

50 notes pour l'examen principal



50 points bonus

50 notes pour l'examen principal



50 points bonus

Takeaway:

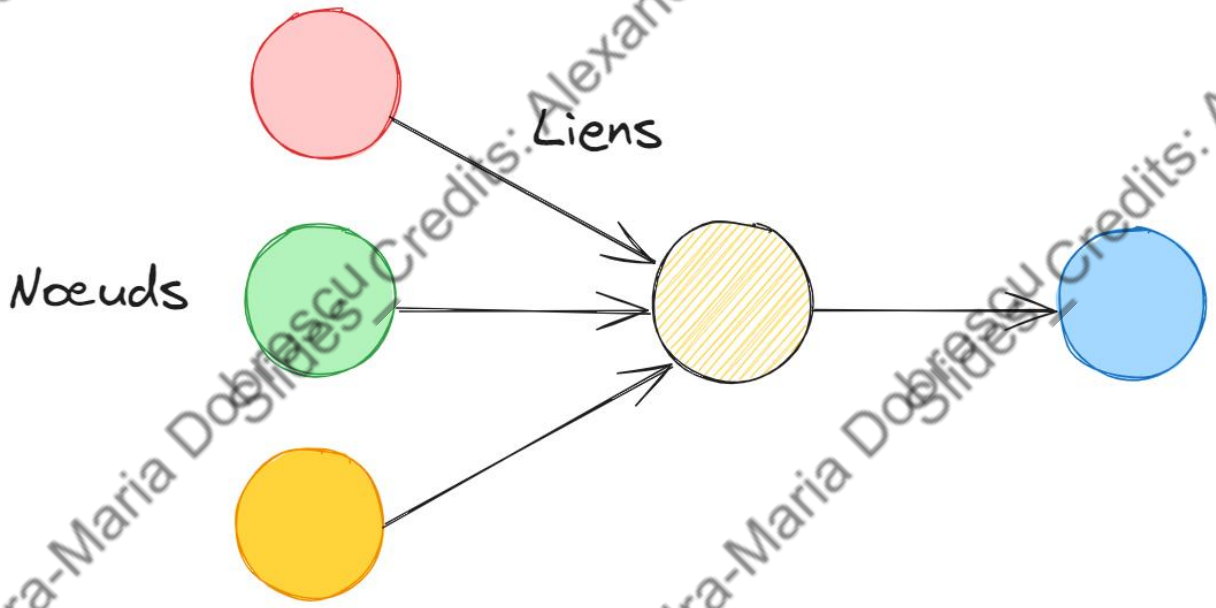
- Quelle était l'échelle de fréquence ?
- Pourquoi le graphique change-t-il ?
- Comment s'appelle ce type de graphique ?

Apprentissage par Réseau Bayésien

ALGORITHMES

Graphique acyclique dirigé
Directed Acyclic Graph (DAG)

MAIS QUE MODÉLISENT CES GRAPHES ?
QUEL RÉSULTAT PEUT-ON OBTENIR À PARTIR D'UN DAG ?



Apprentissage par Réseau Bayésien

ALGORITHMES

Graphique acyclique dirigé
Directed Acyclic Graph (DAG)

UN DAG MODÉLISE L'INCERTITUDE D'UN ÉVÉNEMENT SUR LA BASE DE LA DISTRIBUTION DE PROBABILITÉ CONDITIONNELLE DE CHAQUE VARIABLE ALÉATOIRE. UNE TABLE DE PROBABILITÉ CONDITIONNELLE EST UTILISÉE POUR REPRÉSENTER LA CDP DE CHAQUE VARIABLE DU RÉSEAU.

Probabilité conjointe / Joint Probability :

- mesure statistique de deux ou plusieurs événements se produisant en même temps;
- la probabilité que l'intersection de deux événements ou plus se produise.

Exemple : P(A, B, C) ⇔ La probabilité que les événements A, B et C se produisent.

Apprentissage par Réseau Bayésien

ALGORITHMES

Graphique acyclique dirigé
Directed Acyclic Graph (DAG)

Probabilité conditionnelle / Conditional Probability : La probabilité conditionnelle d'un événement ***X*** est la probabilité que l'événement se produise si un événement ***Y*** s'est déjà produit.

Exemple : ***p(X|Y)*** est la probabilité que l'événement ***X*** se produise, étant donné que l'événement ***Y*** se produit.

- Si ***X*** et ***Y*** sont des événements dépendants, l'expression de la probabilité conditionnelle est donnée par:

$$P(X|Y) = P(X)$$

Apprentissage par Réseau Bayésien

ALGORITHMES

Graphique acyclique dirigé
Directed Acyclic Graph (DAG)

Probabilité conditionnelle / Conditional Probability : La probabilité conditionnelle d'un événement ***X*** est la probabilité que l'événement se produise si un événement ***Y*** s'est déjà produit.

Exemple : ***p(X|Y)*** est la probabilité que l'événement ***X*** se produise, étant donné que l'événement ***Y*** se produit.

- Si ***X*** et ***Y*** sont des événements indépendants, l'expression de la probabilité conditionnelle est donnée par:

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$$

Apprentissage par Réseau Bayésien

ALGORITHMES

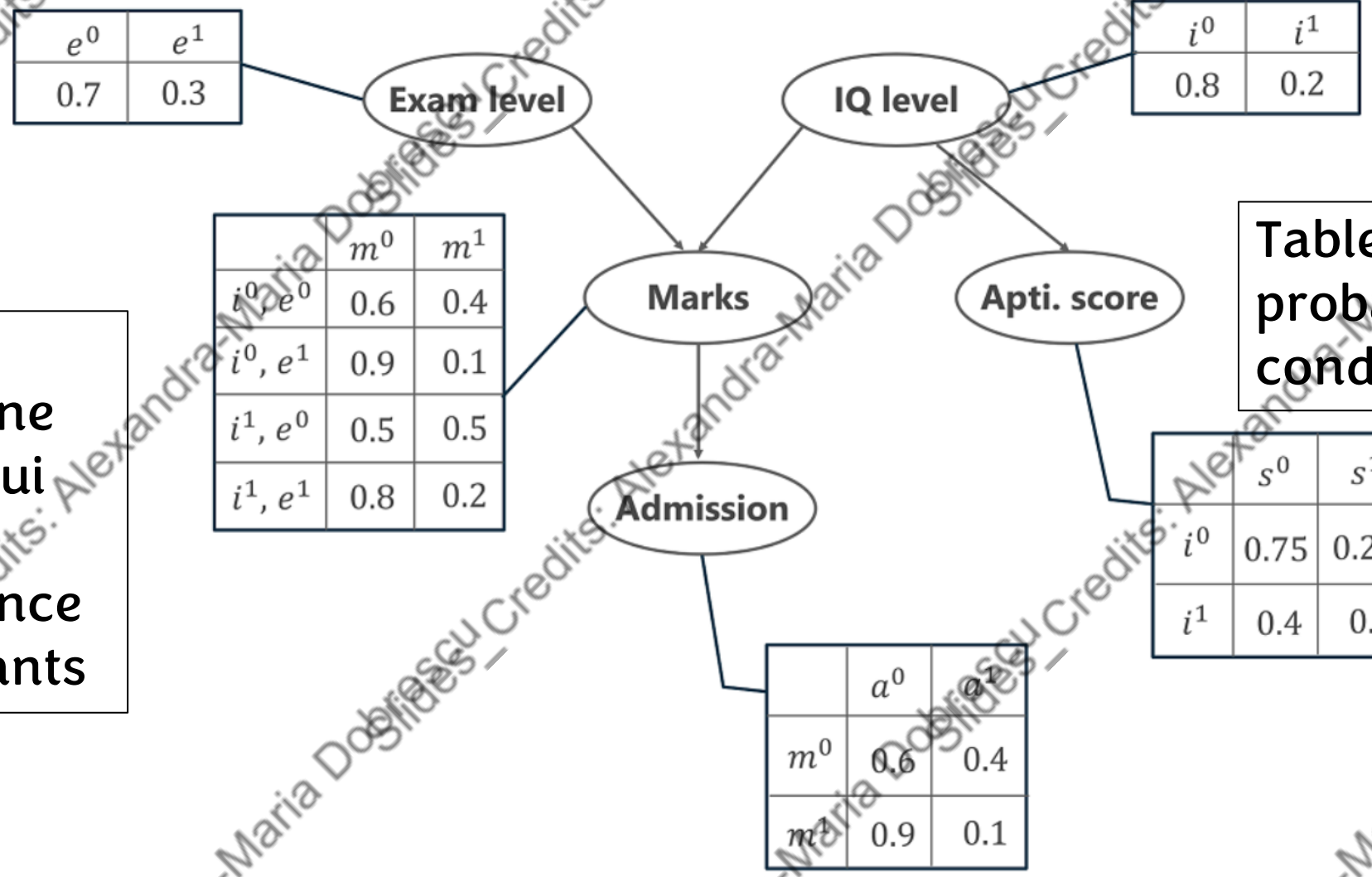
Exemple de Réseaux Bayésien

Exemple : Réseau bayésien qui modélise les notes (m) d'un étudiant lors de son examen. Les notes dépendront de:

- *Niveau d'examen (e):* Il s'agit d'une variable discrète qui peut prendre deux valeurs (difficile, facile);
- *IQ de l'étudiant (i):* Variable discrète pouvant prendre deux valeurs (élevé, faible).

Objectif :

- Les notes permettent de savoir s'il sera admis ou non à l'Université;
- Le IQ permet également de prédire les scores d'aptitude.



e^0	e^1
0.7	0.3

i^0	i^1
0.8	0.2

	m^0	m^1
i^0, e^0	0.6	0.4
i^0, e^1	0.9	0.1
i^1, e^0	0.5	0.5
i^1, e^1	0.8	0.2

	s^0	s^1
i^0	0.75	0.25
i^1	0.4	0.6

	a^0	a^1
m^0	0.6	0.4
m^1	0.9	0.1

Le nœud désigne une variable qui prédit la performance des étudiants

Tableau des probabilités conditionnelles

Apprentissage par Réseau Bayésien

ALGORITHMES

Exemple de Réseaux Bayésien

Maths : La distribution de probabilité conjointe devient:

$$P(a, m, i, e, s) = P(a|m)P(m|i, e)P(i)P(e)P(s|i)$$

- *p(a | m)* représente la probabilité conditionnelle qu'un étudiant soit admis en fonction de ses notes;
- *p(m | i, e)* représente la probabilité conditionnelle des notes de l'étudiant, compte tenu de son niveau de IQ et de son niveau d'examen;
- *p(i)* indique la probabilité de son niveau de IQ (élevé ou faible);
- *p(e)* représente la probabilité du niveau de l'examen (difficile ou facile);
- *p(s | i)* représente la probabilité conditionnelle de ses notes d'aptitude, compte tenu de son niveau de IQ.

Apprentissage par Réseau Bayésien

ALGORITHMES

Exemple de Réseaux Bayésien

Takeaway: Le DAG montre clairement comment chaque variable (nœud) dépend de son nœud parent.

$$P(X_1, .., X_n) = \prod_{i=1}^{n} P(X_i | Parents(X_i))$$

- les notes de l'étudiant dépendent du niveau de l'examen (nœud parent) et niveau d'IQ (nœud parent);
- le score d'aptitude dépend du niveau d'IQ (nœud parent);
- son admission dans une université dépend de ses notes (nœud parent).

Apprentissage par Réseau Bayésien

ALGORITHMES

Exemple de Réseaux Bayésien

Takeaway: Le DAG montre clairement comment chaque variable (nœud) dépend de son nœud parent.

variable aléatoire

$$P(X_1, .., X_n) = \prod_{i=1}^{n} P(X_i | Parents(X_i))$$

dont la probabilité dépend de la probabilité des nœuds parents

Apprentissage par Réseau Bayésien

ALGORITHMES

Diagnostic des maladies: utilisés pour modéliser les symptômes possibles et prédire si une personne est malade ou pas.

Optimisation de la recherche sur le web: utilisés pour améliorer la précision des recherches en comprenant l'intention d'une recherche et en fournissant les résultats les plus pertinents

Application des Réseaux Bayésiens

Le filtrage du spam; également utilisés dans d'autres applications de classification de documents.

Réseaux de régulation génétique : nombreux segments d'ADN; utilisés efficacement pour communiquer avec d'autres segments d'une cellule, directement ou indirectement.

Bibliographie

[1] Subramaniam, A. (2020). What Is Big Data Analytics| Big Data Analytics Tools and Trends| Edureka.