TP 2 : Réseaux bayésiens

Objectifs

Le but de ce TP est passer en revue les capacités des réseaux bayésiens sur un problème relativement simple.

Afin de gagner du temps dans l'implémentation, il vous est demandé d'utiliser la bibliothèque python pgmpy. Vous pourrez notamment vous appuyer sur la documentation fournie dans les pages suivantes :

— `https://github.com/pgmpy/pgmpy_notebook/blob/master/notebooks/2.%20Bayesian%20Networks.ipynb`
— `https://github.com/pgmpy/pgmpy/tree/dev/examples`

1 Construction du modèle

On vous demande de développer pour une banque un prédicteur indiquant si un client est suffisamment fiable pour que la banque lui octroie un prêt. La banque est capable d'observer les revenus du client (Income), son capital disponible (Assets), son historique de paiement (PaymentHistory) ou encore son âge (Age). La banque estime que la capacité d'une personne à rembourser un prêt (BankLoan) dépend au final de la confiance que l'on a en elle (Reliability), son revenu futur estimé (FutureIncome) et le rapport entre ses dettes et ses revenus (DebtIncomeRatio). À partir de ces observations, a été défini un fichier squelette bank.py contenant les 8 variables aléatoires mentionnées ci-dessus, avec leurs valeurs possibles. On vous demande les connexions entre ces variables et leur distribution de probabilité conditionnelle, en vous basant sur les observations suivantes :

1. Meilleur est l'historique de paiement d'une personne, plus celle-ci est digne de confiance ;
2. Plus une personne est âgée, plus celle-ci est probablement digne de confiance ;
3. Meilleur est l'historique de paiement d'une personne, plus celle-ci est probablement âgée ;
4. Plus une personne a un rapport élevé entre leur dette, plus le risque est grand qu'elle ait un mauvais historique de paiement ;
5. Plus hauts sont les revenus d'une personne, plus elle risque d'avoir un capital élevé ;
6. Plus une personne a un capital élevé et dispose de hauts revenus, plus il est probable que ses revenus futurs soient élevés ;
7. Les personnes dignes de confiance sont plus susceptibles de rembourser leur prêt que les autres. De même celles qui ont des revenus futurs prometteurs et qui ont des rapports faibles entre leur dette et leur revenu ont une probabilité plus élevée de rembourser leur prêt.

Dans un premier temps, on vous demande de dessiner le graphe tenant compte de ces observations sur les 8 variables aléatoires. Vous pourrez dessiner le graphe avec l'outil de votre choix, la bibliothèque pgmpy ne permettant pas de construire de figure.

Dans un second temps, vous définirez une série de probabilités conditionnelles permettant de respecter le comportement défini ci-dessus. Par exemple, pour être cohérent avec l'observation 1,

votre réseau devra vérifier

$$\begin{aligned} &P(\text{Reliability} = \text{Reliable}|\text{PaymentHistory} = \text{Excellent}) \\ >&P(\text{Reliability} = \text{Reliable}|\text{PaymentHistory} = \text{Acceptable}) \\ >&P(\text{Reliability} = \text{Reliable}|\text{PaymentHistory} = \text{Unacceptable}) \end{aligned}$$

Plusieurs distributions sont bien sûr possibles, il vous est demandé d'en définir une seule qui vérifie toutes les observations. Des distributions respectant les observations 1 à 4, ainsi que la 7e, vous sont fournies ; vous n'avez plus qu'à compléter pour satisfaire les observations 5 et 6.

Vérifiez que les distributions de probabilités conditionnelles sont bien des lois de probabilité en appliquant check_model() sur votre modèle bayésien.

2 Indépendance entre les variables

En utilisant les capacités offertes par pgmpy, donnez les indépendances locales entre les variables.

En calculant toutes les chaînes actives accessibles depuis un nœud, indiquez toutes les variables aléatoires qui sont indépendantes de ;

— Income,

— Income étant donné BankLoan.

3 Apprentissage par maximum de vraisemblance

Les informations sur 50 000 clients d'une banque vous sont fournies dans le fichier 50000-cases.csv. En utilisant le critère du maximum de vraisemblance et en reprenant la même architecture que précédemment, apprenez sur ce fichier les distributions de probabilités conditionnelles. Affichez ces distributions apprises et comparez-les à celles que vous aviez définies manuellement.

Dans les sections qui suivent, vous ne considérerez dorénavant que le modèle entraîné sur le fichier.

4 Inférence exacte

Le modèle bayésien étudié est de taille modeste, ce qui permet de faire de l'inférence exacte à l'aide de la méthode par élimination de variables. Donnez les probabilités suivantes calculées par votre modèle :

— $P(\text{BankLoan})$;

— $P(\text{BankLoan}|\text{Income} = \text{Low}, \text{Age} = \text{Between16and25}, \text{PaymentHistory} = \text{Excellent}, \text{Assets} = \text{Low})$;

— $P(\text{BankLoan}|\text{Income} = \text{High}, \text{Age} = \text{Between16and25}, \text{PaymentHistory} = \text{Excellent}, \text{Assets} = \text{High})$;

— $P(\text{BankLoan}|\text{Income} = \text{High}, \text{Age} = \text{Over65}, \text{PaymentHistory} = \text{Excellent}, \text{Assets} = \text{High})$.

5 Inférence par échantillonnage en avant

On s'intéresse toujours à l'inférence des quatre mêmes requêtes faites au modèle. Utilisez la bibliothèque pgmpy pour avoir une estimation approchée par échantillonnage en avant. Comparez les valeurs obtenues et le temps de calcul avec celles déterminées par élimination de variables.