LE DATA MINING

La fouille de données

Dr. Nedioui Med Abdelhamid
Université Echahid Hama Lakhder Eloued - Algérie
nedioui3904@gmail.com

Chapitre 4

Classification probabiliste

1. Introduction

Les classificateurs Bayesians (CB) sont couramment utilisé dans l'apprentissage automatique, est une collection d'algorithmes de classification basés sur le théorème de Bayes. Ce n'est pas un algorithme unique, mais une famille d'algorithmes. Tous ces algorithmes partagent tous un principe commun, à savoir que chaque caractéristique classée est indépendante de la valeur des autres.

2. Les classificateurs bayésiens

1. Définition :

Les CBs sont les classificateurs les plus simples en apprentissage supervisé basée sur le théorème de Bayes. Ils peuvent prédire la classe probabilités d'appartenance, telles que la probabilité qu'un échantillon donné appartient à une classe particulière. Les classificateurs supposent que l'effet d'une valeur d'attribut sur une classe donnée est **indépendant** des autres attributs. Cette hypothèse est appelé **indépendance conditionnelle** de classe. Il est fait pour simplifier la calcul impliqué et, en ce sens, est considéré comme «naïf».

2. Les classificateurs bayésiens

2. Théorème de Bayes:

Soit $X = \{x_1, x_2, \ldots, x_n\}$ être un échantillon, dont les composants représentent valeurs faites sur un ensemble de n attributs. En termes bayésiens, X est considéré l'échantillon de données observé . Soit H une hypothèse telle que les données X appartient à une classe spécifique C. Pour les problèmes de classification, l'objectif est de déterminer P(H|X), la probabilité que l'hypothèse H soit vérifiée étant donné les l'échantillon de données observé X.

2. Les classificateurs bayésiens

2.1 Théorème de Bayes:

Selon le théorème de Bayes, la probabilité que nous voulons calculer P (H | X) peut être exprimé en termes de probabilités P(H), P(X|H) et P (X) comme:

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

et ces probabilités peuvent être estimées à partir des données fournies.

2. Les classificateurs bayésiens

3. Exemple :

Notre jeu de données se présente comme suit :

Type	Long (L)	Petit	Sucré (S)	Non sucré	Jaune (J)	Non Jaune	Total
Banane (B)	400	100	350	150	450	50	500
Orange (O)	0	300	150	150	300	0	300
Autre fruit (F)	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

L'idée du jeu est de prédire le type d'un fruit (orange, banane ou autre) qu'on n'a pas encore vu. Ceci en se basant sur ses caractéristiques.

2. Les classificateurs bayésiens

3. Exemple :

Supposons que quelqu'un nous demande de lui donner le type d'un fruit qu'il a. Ses caractéristiques sont les suivantes : Il est jaune , est long , est sucré

Pour savoir s'il s'agit d'une banane, une orange ou d'un autre fruit, il faut calculé les 3 probabilités suivantes:

P(banane|long,jaune,sucre): La probabilité qu'il s'agisse d'une banane sachant que le fruit est long, jaune et sucré.

P(Orange|long,jaune,sucre): La probabilité qu'il s'agisse d'une orange sachant que le fruit est long, jaune et sucré.

P(Autre|long,jaune,sucre):La probabilité qu'il s'agisse d'un autre fruit sachant que qu'il est long, jaune et sucré

2. Les classificateurs bayésiens

3. Exemple :

Le type du fruit “inconnu” qu’on cherche à classifier sera celui où on a **la plus grande probabilité.**

Selon la formule de Bayes, on a :

P(B|L,J,S) = P(S|B)*P(J|B)*P(L|B)*P(B) / P(L)*P(S)*P(J)

P(Banane)= |Banae| / |Tous les fruits| = 50/100=0.5

P(Orange)= 0.3

P(Autre fruits)= 0.2

P(Long)= 0.5

P(Sure)= 0.65

P(Jaune)= 0.8

2. Les classificateurs bayésiens

3. Exemple :

Calculons maintenant le terme P(Long|Banane) la Probabilité que le fruit est long sachant qu'il s'agit d'une banane.

P(Long|Banane)= |Banane **Et** Long | / | Banane |= 400/500=0.8

P(Sucre|Banane)= 0.7

P(Jaune|Banane)= 0.9

P(banane|long,jaune,sucre)=0.8*0.7*0.9*0.5/0.5*0.65*0.8 ≈ 0.969

P(Orange|long,jaune,sucre)= 0

P(Autre|long,jaune,sucre)= 0.072

On remarque que la probabilité que notre fruit soit une banane est largement plus grande que celle des autres. **On classifie notre fruit inconnu comme étant une banane.**

2. Les classificateurs bayésiens

4. Avantages

- Les classifieurs naïfs bayésiens, malgré leurs simplicité, ont des points forts:
- Ils ont besoin d'une petite quantité de données d'entrainement.
- Ils sont très rapides par rapport aux autres classifieurs.
- Ils donnent de bonnes résultats dans le cas de filtrage du courrier indésirable et de classification de documents.

2. Les classificateurs bayésiens

5. Limites

- Les classifieurs naïfs bayésiens certes sont populaires à cause de leur simplicité. Mais, une telle simplicité vient avec un cout.
- Les probabilités obtenues en utilisant ces classifieurs ne doivent pas être prises au sérieux.
- S'il existe une grande corrélation entre les caractéristiques, ils vont donner une mauvaise performance.
- Dans le cas des caractéristiques continues (prix, surface..), les données doivent suivre la loi normale.

3. Les réseaux bayésiens

1.Introduction

On retrouve les réseaux Bayésiens dans beaucoup d'applications, sans même le savoir. Microsoft par exemple est un fervent utilisateur de cette structure , aussi Google et Mozilla via leurs filtres anti-spam. De nombreux travaux dans le domaine sont réalisés, preuve de l'intérêt et de la puissance de ces réseaux.

Les réseaux Bayésiens (RB) constituent un ensemble de méthodes statistiques utilisées pour modéliser des problèmes, extraire de l'information et prendre des décisions. Ils sont un formalisme de raisonnement probabiliste utilisé dans plusieurs domaines tels que l'industrie, la santé, finance et le traitement d'images.

3. Les réseaux bayésiens

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.

Un réseau Bayésien est un graphe acyclique compose de:

Sommets : Un ensemble de variables aléatoires ayant chacune un ensemble d'états fini, représentent les variable des systèmes.

Arcs : Un arc A relie une et une seule variable B signifie que A constitue une cause de B (A influence B).

3. Les réseaux bayésiens

3. Construction d’un graphe

Construire un réseau bayésien c'est donc :

- 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 "paramètres".

Généralement, la structure est définie par des experts et les tables de probabilités calculées à partir de données expérimentales.

3. Les réseaux bayésiens

4. Apprentissage

4.1 Definition

L'apprentissage consiste à trouver un réseau Bayésien modélisant les données disponibles en s'appuyant sur les connaissances à priori disponibles.

L'apprentissage d'un réseau bayésien doit répondre aux deux questions suivantes :

- Comment estimer les lois de probabilités conditionnelles ?
- Comment trouver la structure du réseau bayésien ?

3. Les réseaux bayésiens

4. Apprentissage

5.2 Les types d'apprentissage

Apprentissage avec structure connue et données complètes :

Dans le cas où toutes les variables sont observées, la méthode la plus simple et la plus utilisée est l'estimation statistique qui consiste à estimer la probabilité d'un événement par la fréquence d'apparition de l'événement dans la base de données. Cette approche, appelée maximum de vraisemblance (MV).

3. Les réseaux bayésiens

4. Apprentissage

5.2 Les types d'apprentissage

Apprentissage avec structure connue et de données incomplètes :

Dans les applications pratiques, les bases de données sont très souvent incomplètes. Certaines variables ne sont observées que partiellement ou même jamais. La méthode d'estimation de paramètres avec des données incomplètes la plus couramment utilisée est fondée sur l'algorithme itératif EM (Expectation Maximisation).

3. Les réseaux bayésiens

5. Exemple