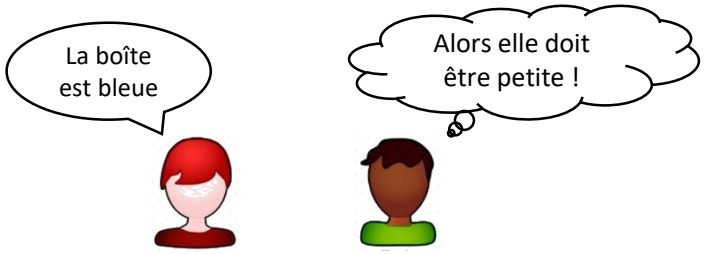
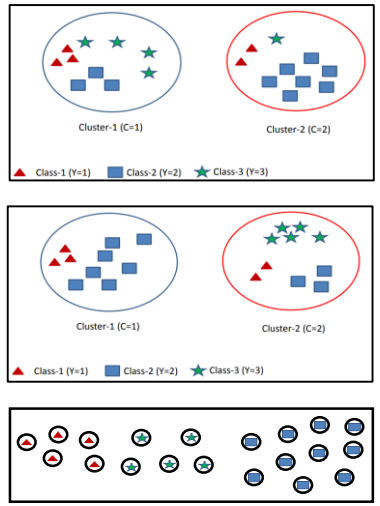
Information mutuelle entre plus de deux variables

Dans l'épisode précédent…

- Dépendance entre variables, et **information** partagée / **mutuelle** :
  - $IM(C,T) = H(C) + H(T) - H(C,T)$
  - $IM(C,T) = H(T) - H(T|C) = H(C) - H(C|T)$



- Calcul pratique de l'information mutuelle
  - Information mutuelle normalisée
- Quelques utilisations de l'information mutuelle
  - Clustering
  - Sélection de features…



Exemple d’utilisation de l’information mutuelle

Sélection de variables / attributs / features :

Nous voulons prédire le risque qu’un patient ait des complications après une crise cardiaque.

Nous avons récolté plein de données, mais nous ne savons pas lesquelles sont pertinentes :

Age, poids, taille, code postal, couleur des cheveux et des yeux, glycémie, saturation du sang en oxygène, rythme cardiaque au repos et pendant l’effort, etc.

Comment faire le tri ?

Sélection de variables

Propriétés désirées :

1. Les variables sélectionnées sont liées au risque de complications

2. Les variables sélectionnées sont le moins redondantes possible

Sélection de variables

Propriétés désirées :

1. Les variables sélectionnées sont liées au risque de complications

Rappel…

1. Les variables sélectionnées sont le moins redondantes possible

Sélection des variables associées au risque de complications

Méthode : Maximiser l'IM entre les variables sélectionnées et le risque de complication

1. Calculez l'IM entre chaque variable (tranches d'âge/de poids…) et le risque de complications
2. Sélectionnez les variables les plus pertinentes

Age	Poids	Taille	Code postal	Couleur cheveux	Couleur yeux	Glycémie	Saturation	Rythme repos	Rythme effort	Complications
78	84	168	83113	Brun	Bleu	1.02	97	73	113	Oui
67	93	170	83271	Blond	Bleu	1.10	92	82	115	Oui
84	110	186	83230	Chatain	Marron	1.30	90	90	144	Oui
70	68	158	83000	Chatain	Marron	0.75	96	78	136	Non
93	73	182	83018	Roux	Bleu	1.12	89	63	147	Non
59	83	171	83620	Brun	Marron	0.86	100	84	128	Non
80	99	143	83330	Chatain	Bleu	0.96	99	77	139	Non

Sélection des variables associées au risque de complications

Age :

$IM(Complications; Age) \approx 0.0202$

$NMI(Complications; Age) \approx 0.0205$

Poids :

$IM(Complications; Poids) \approx 0.3060$

$NMI(Complications; Poids) \approx 0.2408$

Glycémie :

$IM(Complications; Glycémie) \approx 0.1280$

$NMI(Complications; Glycémie) \approx 0.1299$

etc...

Age	Poids	Taille	Code postal	Couleur cheveux	Couleur yeux	Glycémie	Sat.	Rythme repos	Rythme effort	Complications
78	84	168	83113	Brun	Bleu	1.02	97	73	113	Oui
67	93	170	83271	Blond	Bleu	1.10	92	82	115	Oui
84	110	186	83230	Chatain	Marron	1.30	90	90	144	Oui
70	68	158	83000	Chatain	Marron	0.75	96	78	136	Non
93	73	182	83018	Roux	Bleu	1.12	89	63	147	Non
59	83	171	83620	Brun	Marron	0.86	100	84	128	Non
80	99	143	83330	Chatain	Bleu	0.96	99	77	139	Non

Sélection des variables associées au risque de complications

Age :

$IM(Complications; Age) \approx 0.0202$

$NMI(Complications; Age) \approx 0.0205$

Poids :

$IM(Complications; Poids) \approx 0.3060$

$NMI(Complications; Poids) \approx 0.2408$

Glycémie :

$IM(Complications; Glycémie) \approx 0.1280$

$NMI(Complications; Glycémie) \approx 0.1299$

etc...

Méthode dite « greedy » :

1. On sélectionne la meilleure variable
2. On sélectionne la deuxième meilleure variable
3. etc.
4. ... jusqu'à ce qu'on ait assez de variables :
   - gain d'information négligeable

   ou
   - amélioration de la prédiction négligeable

Remarque : il existe d'autres algorithmes pour optimiser la sélection

Sélection des variables associées au risque de complications

Age :

$IM(Complications; Age) \approx 0.0202$

$NMI(Complications; Age) \approx 0.0205$

Poids :

$IM(Complications; Poids) \approx 0.3060$

$NMI(Complications; Poids) \approx 0.2408$

Glycémie :

$IM(Complications; Glycémie) \approx 0.1280$

$NMI(Complications; Glycémie) \approx 0.1299$

etc...

Méthode un peu moins « greedy » :

1. On sélectionne la variable qui maximise l'information mutuelle entre l'**ensemble** des n variables sélectionnées et la variable *Complications* : $IM(X^n; Complications)$ avec $X^n = (X_1, \dots, X_n)$
2. ... jusqu'à ce qu'on ait assez de variables :
   - gain d'information négligeable

   ou
   - amélioration de la prédiction négligeable

Remarque : il existe d'autres algorithmes pour optimiser la sélection

Sélection de variables

Propriétés désirées :

1. Les variables sélectionnées sont liées au risque de complications

2. Les variables sélectionnées sont le moins redondantes possible

Comment faire?

Rejet des variables redondantes

Sachant que des variables ont déjà été sélectionnées,

la nouvelle variable apporte-t-elle de l'information supplémentaire ?

Rejet des variables redondantes

Information mutuelle *conditionnelle*

Sachant que des variables ont déjà été sélectionnées,

la nouvelle variable apporte-t-elle de l'information supplémentaire ?

$$IM(X;Y) = H(X) + H(Y) - H(X,Y)$$

Information mutuelle conditionnelle

Information mutuelle entre X et Y **sachant une troisième variable Z** :

$$IM(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z)$$

Information mutuelle conditionnelle

Information mutuelle entre X et Y **sachant une troisième variable Z** :

$$IM(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z)$$

$$IM(X;Y|Z) = H(X,Z) - H(Z) + H(Y,Z) - H(Z) - H(X,Y,Z) + H(Z)$$

Information mutuelle conditionnelle

Information mutuelle entre X et Y **sachant une troisième variable Z** :

$$IM(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z)$$

$$IM(X;Y|Z) = H(X,Z) - H(Z) + H(Y,Z) - H(Z) - H(X,Y,Z) + H(Z)$$

$$IM(X;Y|Z) = H(Y,Z) + H(X) - H(X,Y,Z) - H(Z) - H(X) + H(X,Z)$$

Information mutuelle conditionnelle

Information mutuelle entre X et Y **sachant une troisième variable Z** :

$$IM(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z)$$

$$IM(X;Y|Z) = H(X,Z) - H(Z) + H(Y,Z) - H(Z) - H(X,Y,Z) + H(Z)$$

$$IM(X;Y|Z) = H(Y,Z) + H(X) - H(X,Y,Z) - H(Z) - H(X) + H(X,Z)$$

$$IM(X;Y|Z) = IM(X\ ;\ Y,Z) - IM(X;Z)$$

$$IM(X;Y|Z) = IM\big(X\ ;(Y,Z)\big) - IM(X;Z)$$

avec $IM(X\ ;\ Y,Z)$ ou $IM\big(X\ ;\ (Y,Z)\big)$ l'information mutuelle entre X et la **variable jointe** $(\boldsymbol{Y},\boldsymbol{Z})$.

Information mutuelle conditionnelle

Information mutuelle entre X et Y **sachant une troisième variable Z** :

$$IM(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z)$$

$$IM(X;Y|Z) = H(X,Z) - H(Z) + H(Y,Z) - H(Z) - H(X,Y,Z) + H(Z)$$

$$IM(X;Y|Z) = H(Y,Z) + H(X) - H(X,Y,Z) - H(Z) - H(X) + H(X,Z)$$

$$IM(X;Y|Z) = IM(X\ ;\ Y,Z) - IM(X;Z)$$

$$IM(X;Y|Z) = IM\big(X\ ;\ (Y,Z)\big) - IM(X;Z)$$

avec $IM(X\ ;\ Y,Z)$ ou $IM\big(X\ ;\ (Y,Z)\big)$ l'information mutuelle entre X et la **variable jointe** $(\boldsymbol{Y},\boldsymbol{Z})$.

L'IMC mesure donc la différence entre l'information partagée par X et la variable jointe (Y,Z) , et l'information partagée par X et Z.

Information mutuelle conditionnelle

Information mutuelle entre $Glycémie$ et $Complications$ **sachant que la variable $\boldsymbol{Poids}$ est déjà sélectionnée** :

$$IM(Glycémie; Complications|Poids) = IM\big(Glycémie;\ (Complications, Poids)\big) - IM(Glycémie; Poids)$$

Information partagée entre
la nouvelle variable considérée
et la paire $(Complications, Poids)$ déjà obtenue

Redondance de l'information

Information mutuelle conditionnelle

Et si on a déjà choisi plus d'une variable ?

Pareil :

$$IM(X;Y \mid Z_1, \dots, Z_n) = IM(X\ ;\ Y, Z_1, \dots, Z_n) - IM(X\ ; Z_1, \dots, Z_n)$$

$$IM(X;Y \mid Z^n) = IM(X\ ;\ Y, Z^n) - IM(X\ ; Z^n) \text{ avec } Z^n = (Z_1, \dots, Z_n)$$

et même démonstration en partant de $IM(X;Y|Z^n) = H(X|Z^n) + H(Y|Z^n) - H(X,Y|Z^n)$.

Information mutuelle conditionnelle

Et si on a déjà choisi plus d'une variable ?

Pareil :

$$IM(X;Y \mid Z_1,\dots,Z_n) = IM(X\,;\,Y,Z_1,\dots,Z_n) - IM(X\,;Z_1,\dots,Z_n)$$

$$IM(X;Y \mid Z^n) = IM(X\,;\,Y,Z^n) - IM(X\,;Z^n) \text{ avec } Z^n = (Z_1,\dots,Z_n)$$

et même démonstration en partant de $IM(X;Y|Z^n) = H(X|Z^n) + H(Y|Z^n) - H(X,Y|Z^n)$.

Donc : $IM(Glycémie;Complications|Z^n) = IM(Glycémie\,;\,Complications,Z^n) - IM(Glycémie;Z^n)$

Information partagée entre
la nouvelle variable considérée
et le système $(Complications, Z^n)$ déjà obtenu

Redondance de l'information

Calcul de l'information mutuelle conditionnelle

Information mutuelle entre X et Y **sachant une troisième variable Z** :

$$IM(X;Y \mid Z) = H(X|Z) + H(Y|Z) - H(X,Y \mid Z)$$

En utilisant : $H(Y|X) = -\sum_{x,y} p_{x,y} \log_2 \frac{p_{x,y}}{p_x}$ (voir slide 17 de la dernière fois)

$$IM(X;Y \mid Z) = -\sum_{x,z} p_{x,z} \log_2 \frac{p_{x,z}}{p_z} - \sum_{y,z} p_{y,z} \log_2 \frac{p_{y,z}}{p_z} + \sum_{x,y,z} p_{x,y,z} \log_2 \frac{p_{x,y,z}}{p_z}$$

Calcul de l'information mutuelle conditionnelle

Information mutuelle entre X et Y **sachant une troisième variable Z** :

$$IM(X;Y \mid Z) = H(X|Z) + H(Y|Z) - H(X,Y \mid Z)$$

En utilisant : $H(Y|X) = -\sum_{x,y} p_{x,y} \log_2 \frac{p_{x,y}}{p_x}$ (voir slide 17 de la dernière fois)

$$IM(X;Y \mid Z) = -\sum_{x,z} p_{x,z} \log_2 \frac{p_{x,z}}{p_z} - \sum_{y,z} p_{y,z} \log_2 \frac{p_{y,z}}{p_z} + \sum_{x,y,z} p_{x,y,z} \log_2 \frac{p_{x,y,z}}{p_z}$$

$$IM(X;Y \mid Z) = \sum_{x,y,z} p_{x,y,z} \log_2 \frac{p_{x,y,z} p_z}{p_{x,z} p_{y,z}}$$

Calcul de l'information mutuelle conditionnelle

Information mutuelle entre X et Y **sachant une troisième variable Z** :

$$IM(X;Y \mid Z) = H(X|Z) + H(Y|Z) - H(X,Y \mid Z)$$

En utilisant : $H(Y|X) = -\sum_{x,y} p_{x,y} \log_2 \frac{p_{x,y}}{p_x}$ (voir slide 17 de la dernière fois)

$$IM(X;Y \mid Z) = -\sum_{x,z} p_{x,z} \log_2 \frac{p_{x,z}}{p_z} - \sum_{y,z} p_{y,z} \log_2 \frac{p_{y,z}}{p_z} + \sum_{x,y,z} p_{x,y,z} \log_2 \frac{p_{x,y,z}}{p_z}$$

$$IM(X;Y \mid Z) = \sum_{x,y,z} p_{x,y,z} \log_2 \frac{p_{x,y,z} p_z}{p_{x,z} p_{y,z}}$$

$$IM(X;Y \mid Z) = \sum_{z} p_z \sum_{x,y} p_{x,y|z} \log_2 \frac{p_{x,y|z}}{p_{x|z} p_{y|z}}$$

Rejet des variables redondantes

Quelle est l'information mutuelle de G*lycémie* et *Complications* sachant que *Poids* est déjà sélectionnée ?

Age	Poids	Taille	Code postal	Couleur cheveux	Couleur yeux	Glycémie	Sat.	Rythme repos	Rythme effort	Complications
78	84	168	83113	Brun	Bleu	1.02	97	73	113	Oui
67	93	170	83271	Blond	Bleu	1.10	92	82	115	Oui
84	110	186	83230	Chatain	Marron	1.30	90	90	144	Oui
70	68	158	83000	Chatain	Marron	0.75	96	78	136	Non
93	73	182	83018	Roux	Bleu	1.12	89	63	147	Non
59	83	171	83620	Brun	Marron	0.86	100	84	128	Non
80	99	143	83330	Chatain	Bleu	0.96	99	77	139	Non

Rejet des variables redondantes

Quelle est l'information mutuelle de *Glycémie* et *Complications* sachant que *Poids* est déjà sélectionnée ?

Commencer par construire les tables de proba jointes :

$poids \in \{[0,75] ;]75,85] ;]85, \infty[\}$

$glycémie \in \{[0,1.10[; [1.10, \infty[\}$

Pour $p \leq 75$:

g \ c	Oui	Non
< 1.10		
≥ 1.10		

Age	Poids	Taille	Code postal	Couleur cheveux	Couleur yeux	Glycémie	Sat.	Rythme repos	Rythme effort	Complications
78	84	168	83113	Brun	Bleu	1.02	97	73	113	Oui
67	93	170	83271	Blond	Bleu	1.10	92	82	115	Oui
84	110	186	83230	Chatain	Marron	1.30	90	90	144	Oui
70	68	158	83000	Chatain	Marron	0.75	96	78	136	Non
93	73	182	83018	Roux	Bleu	1.12	89	63	147	Non
59	83	171	83620	Brun	Marron	0.86	100	84	128	Non
80	99	143	83330	Chatain	Bleu	0.96	99	77	139	Non

Rejet des variables redondantes

Quelle est l'information mutuelle de *Glycémie* et *Complications* sachant que *Poids* est déjà sélectionnée ?

Commencer par construire les tables de proba jointes :

$poids \in \{[0,75] ;]75,85] ;]85, \infty[\}$

$glycémie \in \{[0,1.10[; [1.10, \infty[\}$

Pour $p \leq 75$:

g \ c	Oui	Non
< 1.10	0	1/2
≥ 1.10	0	1/2

Age	Poids	Taille	Code postal	Couleur cheveux	Couleur yeux	Glycémie	Sat.	Rythme repos	Rythme effort	Complications
78	84	168	83113	Brun	Bleu	1.02	97	73	113	Oui
67	93	170	83271	Blond	Bleu	1.10	92	82	115	Oui
84	110	186	83230	Chatain	Marron	1.30	90	90	144	Oui
70	68	158	83000	Chatain	Marron	0.75	96	78	136	Non
93	73	182	83018	Roux	Bleu	1.12	89	63	147	Non
59	83	171	83620	Brun	Marron	0.86	100	84	128	Non
80	99	143	83330	Chatain	Bleu	0.96	99	77	139	Non

Rejet des variables redondantes

Quelle est l'information mutuelle de *Glycémie* et *Complications* sachant que *Poids* est déjà sélectionnée ?

Commencer par construire les tables de proba jointes :

$poids \in \{[0,75] ;]75,85] ;]85, \infty[\}$

$glycémie \in \{[0,1.10[; [1.10, \infty[\}$

Pour $75 < p \leq 85$:

g \ c	Oui	Non
< 1.10		
≥ 1.10		

Age	Poids	Taille	Code postal	Couleur cheveux	Couleur yeux	Glycémie	Sat.	Rythme repos	Rythme effort	Complications
78	84	168	83113	Brun	Bleu	1.02	97	73	113	Oui
67	93	170	83271	Blond	Bleu	1.10	92	82	115	Oui
84	110	186	83230	Chatain	Marron	1.30	90	90	144	Oui
70	68	158	83000	Chatain	Marron	0.75	96	78	136	Non
93	73	182	83018	Roux	Bleu	1.12	89	63	147	Non
59	83	171	83620	Brun	Marron	0.86	100	84	128	Non
80	99	143	83330	Chatain	Bleu	0.96	99	77	139	Non

Rejet des variables redondantes

Quelle est l'information mutuelle de *Glycémie* et *Complications* sachant que *Poids* est déjà sélectionnée ?

Commencer par construire les tables de proba jointes :

$poids \in \{[0,75] ;]75,85] ;]85, \infty[\}$

$glycémie \in \{[0,1.10[; [1.10, \infty[\}$

Pour $75 < p \leq 85$:

g \ c	Oui	Non
< 1.10	1/2	1/2
≥ 1.10	0	0

Age	Poids	Taille	Code postal	Couleur cheveux	Couleur yeux	Glycémie	Sat.	Rythme repos	Rythme effort	Complications
78	84	168	83113	Brun	Bleu	1.02	97	73	113	Oui
67	93	170	83271	Blond	Bleu	1.10	92	82	115	Oui
84	110	186	83230	Chatain	Marron	1.30	90	90	144	Oui
70	68	158	83000	Chatain	Marron	0.75	96	78	136	Non
93	73	182	83018	Roux	Bleu	1.12	89	63	147	Non
59	83	171	83620	Brun	Marron	0.86	100	84	128	Non
80	99	143	83330	Chatain	Bleu	0.96	99	77	139	Non

Rejet des variables redondantes

Quelle est l'information mutuelle de *Glycémie* et *Complications* sachant que *Poids* est déjà sélectionnée ?

Commencer par construire les tables de proba jointes :

$poids \in \{[0,75] ;]75,85] ;]85, \infty[\}$

$glycémie \in \{[0,1.10[; [1.10, \infty[\}$

Pour $p > 85$:

g \ c	Oui	Non
< 1.10		
≥ 1.10		

Age	Poids	Taille	Code postal	Couleur cheveux	Couleur yeux	Glycémie	Sat.	Rythme repos	Rythme effort	Complications
78	84	168	83113	Brun	Bleu	1.02	97	73	113	Oui
67	93	170	83271	Blond	Bleu	1.10	92	82	115	Oui
84	110	186	83230	Chatain	Marron	1.30	90	90	144	Oui
70	68	158	83000	Chatain	Marron	0.75	96	78	136	Non
93	73	182	83018	Roux	Bleu	1.12	89	63	147	Non
59	83	171	83620	Brun	Marron	0.86	100	84	128	Non
80	99	143	83330	Chatain	Bleu	0.96	99	77	139	Non

Rejet des variables redondantes

Quelle est l'information mutuelle de *Glycémie* et *Complications* sachant que *Poids* est déjà sélectionnée ?

Commencer par construire les tables de proba jointes :

$poids \in \{[0,75] ;]75,85] ;]85, \infty[\}$

$glycémie \in \{[0,1.10[; [1.10, \infty[\}$

Pour $p > 85$:

g \ c	Oui	Non
< 1.10	0	1/3
≥ 1.10	2/3	0

Age	Poids	Taille	Code postal	Couleur cheveux	Couleur yeux	Glycémie	Sat.	Rythme repos	Rythme effort	Complications
78	84	168	83113	Brun	Bleu	1.02	97	73	113	Oui
67	93	170	83271	Blond	Bleu	1.10	92	82	115	Oui
84	110	186	83230	Chatain	Marron	1.30	90	90	144	Oui
70	68	158	83000	Chatain	Marron	0.75	96	78	136	Non
93	73	182	83018	Roux	Bleu	1.12	89	63	147	Non
59	83	171	83620	Brun	Marron	0.86	100	84	128	Non
80	99	143	83330	Chatain	Bleu	0.96	99	77	139	Non

$$IM(X;Y \mid Z) = \sum_z p_z \sum_{x,y} p_{x,y|z} \log_2 \frac{p_{x,y|z}}{p_{x|z} p_{y|z}}$$

Rejet des variables redondantes

Quelle est l'information mutuelle de *Glycémie* et *Complications* sachant que *Poids* est déjà sélectionnée ?

$poids \in \{[0,75] ;]75,85] ;]85, \infty[\}$

g \ c	Oui	Non
< 1.10	0	1/2
≥ 1.10	0	1/2

g \ c	Oui	Non
< 1.10	1/2	1/2
≥ 1.10	0	0

g \ c	Oui	Non
< 1.10	0	1/3
≥ 1.10	2/3	0

$$IM(X;Y \mid Z) = \sum_z p_z \sum_{x,y} p_{x,y|z} \log_2 \frac{p_{x,y|z}}{p_{x|z} p_{y|z}}$$

Rejet des variables redondantes

Quelle est l'information mutuelle de *Glycémie* et *Complications* sachant que *Poids* est déjà sélectionnée ?

$poids \in \{[0,75] ;]75,85] ;]85, \infty[\}$

g \ c	Oui	Non
< 1.10	0	1/2
≥ 1.10	0	1/2

g \ c	Oui	Non
< 1.10	1/2	1/2
≥ 1.10	0	0

g \ c	Oui	Non
< 1.10	0	1/3
≥ 1.10	2/3	0

$$IM(G;C|P) = \frac{2}{7} * \left(\frac{1}{2}\log_2 \frac{\frac{1}{2}}{\frac{1}{2}*1} + \frac{1}{2}\log_2 \frac{\frac{1}{2}}{\frac{1}{2}*1}\right) + \frac{2}{7} * \left(\frac{1}{2}\log_2 \frac{\frac{1}{2}}{1*\frac{1}{2}} + \frac{1}{2}\log_2 \frac{\frac{1}{2}}{1*\frac{1}{2}}\right) + \frac{3}{7} * \left(\frac{1}{3}\log_2 \frac{\frac{1}{3}}{\frac{1}{3}*\frac{1}{3}} + \frac{2}{3}\log_2 \frac{\frac{2}{3}}{\frac{2}{3}*\frac{2}{3}}\right)$$

$$IM(X;Y \mid Z) = \sum_z p_z \sum_{x,y} p_{x,y|z} \log_2 \frac{p_{x,y|z}}{p_{x|z} p_{y|z}}$$

Rejet des variables redondantes

Quelle est l'information mutuelle de *Glycémie* et *Complications* sachant que *Poids* est déjà sélectionnée ?

$poids \in \{[0,75] ;]75,85] ;]85, \infty[\}$

g \ *c*	Oui	Non
< 1.10	0	1/2
≥ 1.10	0	1/2

g \ *c*	Oui	Non
< 1.10	1/2	1/2
≥ 1.10	0	0

g \ *c*	Oui	Non
< 1.10	0	1/3
≥ 1.10	2/3	0

$$IM(G;C|P) = \frac{2}{7} * \left(\frac{1}{2}\log_2 \frac{\frac{1}{2}}{\frac{1}{2}*1} + \frac{1}{2}\log_2 \frac{\frac{1}{2}}{\frac{1}{2}*1}\right) + \frac{2}{7} * \left(\frac{1}{2}\log_2 \frac{\frac{1}{2}}{1*\frac{1}{2}} + \frac{1}{2}\log_2 \frac{\frac{1}{2}}{1*\frac{1}{2}}\right) + \frac{3}{7} * \left(\frac{1}{3}\log_2 \frac{\frac{1}{3}}{\frac{1}{3}*\frac{1}{3}} + \frac{2}{3}\log_2 \frac{\frac{2}{3}}{\frac{2}{3}*\frac{2}{3}}\right)$$

$$IM(G;C|P) \approx 0.3936$$

Rejet des variables redondantes

Rappel :

$IM(Complications; Poids) \approx 0.3060$

et

$IM(Complications; Glycémie) \approx 0.1280$

$IM(Glycémie\ ;\ Complications|Poids) \approx 0.3936$

Glycémie est-elle une variable intéressante ?

Rejet des variables redondantes

Rappel :

$IM(Complications; Poids) \approx 0.3060$

et

$IM(Complications; Glycémie) \approx 0.1280$

Glycémie partage un peu d'information avec *Complications*, les deux sont un peu liées, mais pas tant que ça comparé à *Poids* et *Complications*

$IM(Glycémie\ ;\ Complications|Poids) \approx 0.3936$

Mais si le poids est connu, alors le lien entre *Glycémie* et *Complications* est renforcé : connaître *Glycémie* apporte plus d'information sur *Complications*

Glycémie est-elle une variable intéressante ?

Petit retour en arrière : Sélection des variables utiles

Age :

$IM(Complications; Age) \approx 0.0202$

$NMI(Complications; Age) \approx 0.0205$

Poids :

$IM(Complications; Poids) \approx 0.3060$

$NMI(Complications; Poids) \approx 0.2408$

Glycémie :

$IM(Complications; Glycémie) \approx 0.1280$

$NMI(Complications; Glycémie) \approx 0.1299$

etc...

Méthode un peu moins « greedy » :

1. On sélectionne la variable qui maximise l'information mutuelle entre l'ensemble des n variables sélectionnées et la variable *Complications* : $\boldsymbol{IM(X^n; Complications)}$ avec $X^n = (X_1, \dots, X_n)$
2. ... jusqu'à ce qu'on ait assez de variables :
   - gain d'information négligeable

   ou
   - amélioration de la prédiction négligeable

Remarque : il existe d'autres algorithmes pour optimiser la sélection

Petit retour en arrière : Sélection des variables utiles

Comment calculer $IM(X^n; Complications)$ en pratique ?

Nous avons maintenant les outils nécessaires pour montrer que $IM(X^n; Y) = \sum_i IM(X_i; Y|X^{i-1})$. En effet, nous avons appris (slide 20) que :

$$IM(A; B \mid C^n) = IM\big(A\,; (B, C^n)\big) - IM(A\,; C^n)$$

donc

$$\sum_i IM(Y; X_i|X^{i-1}) = \sum_{i=1}^{n} IM\big(Y; X^i\big) - IM(Y; X^{i-1}) = IM(Y; X^n) - IM(Y; X^0)$$

où par convention $X^0 = \emptyset$, d'où :

$$\sum_i IM(Y; X_i|X^{i-1}) = IM(Y; X^n)$$

ou encore, par symétrie de l'IM :

$$\boldsymbol{IM(X^n; Y) = \sum_i IM(X_i; Y|X^{i-1})}$$

Petit retour en arrière : Sélection des variables utiles

Age :

$IM(Complications; Age) \approx 0.0202$

$NMI(Complications; Age) \approx 0.0205$

Poids :

$IM(Complications; Poids) \approx 0.3060$

$NMI(Complications; Poids) \approx 0.2408$

Glycémie :

$IM(Complications; Glycémie) \approx 0.1280$

$NMI(Complications; Glycémie) \approx 0.1299$

etc...

Méthode un peu moins « greedy » :

1. On sélectionne la variable qui maximise l'information mutuelle entre l'ensemble des n variables sélectionnées et la variable *Complications* : $\boldsymbol{IM(X^n; Complications)}$ avec $X^n = (X_1, \dots, X_n)$, calculée comme : $\boldsymbol{IM(X^n; C) = \sum_i IM(X_i; C|X^{i-1})}$
2. ... jusqu'à ce qu'on ait assez de variables :
   - gain d'information négligeable

   ou
   - amélioration de la prédiction négligeable

Remarque : il existe d'autres algorithmes pour optimiser la sélection

Sélection de variables

Propriétés désirées :

1. Les variables sélectionnées sont liées au risque de complications

2. Les variables sélectionnées sont le moins redondantes possible

➢ Information mutuelle conditionnelle

Petit retour en arrière : Sélection des variables utiles

Age :

$IM(Complications; Age) \approx 0.0202$

$NMI(Complications; Age) \approx 0.0205$

Poids :

$IM(Complications; Poids) \approx 0.3060$

$NMI(Complications; Poids) \approx 0.2408$

Glycémie :

$IM(Complications; Glycémie) \approx 0.1280$

$NMI(Complications; Glycémie) \approx 0.1299$

etc...

$IM(Complications;\ Glycémie|Poids) \approx 0.3936$

Méthode moins « greedy » et qui limite la redondance :

1. On sélectionne la variable qui maximise l'information mutuelle **conditionnelle sachant** l'**ensemble** des $n-1$ variables déjà sélectionnées :
   $\boldsymbol{IM(X_n; Complications|X^{n-1})}$
2. ... jusqu'à ce qu'on ait assez de variables :
   - gain d'information négligeable

   ou
   - amélioration de la prédiction négligeable

Remarque : il existe toujours d'autres algorithmes pour optimiser la sélection

Zoom sur une propriété de l'information mutuelle conditionnelle

$$IM(X;Y \mid Z) = \sum_{z} p_z \sum_{x,y} p_{(x,y)|z} \log_2 \frac{p_{(x,y)|z}}{p_{x|z} p_{y|z}}$$

Zoom sur une propriété de l'information mutuelle conditionnelle

$$IM(X;Y \mid Z) = \sum_z p_z \sum_{x,y} p_{(x,y)|z} \log_2 \frac{p_{(x,y)|z}}{p_{x|z} p_{y|z}}$$

$$IM(X;Y \mid Z) = \sum_z p_z \, D\left(p_{(x,y)|z} \,\middle\|\, p_{x|z} p_{y|z}\right)$$

Divergence de Kullback-Leibler

Divergence de Kullback-Leibler

- Rappel de définition :

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Mesure l'écart / l'**information de discrimination** entre les distributions de probabilités p et q.

Divergence de Kullback-Leibler

- Rappel de définition :

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Mesure l'écart / l'**information de discrimination** entre les distributions de probabilités p et q.

- Lien avec l'information mutuelle conditionnelle :

$$IM(X;Y \mid Z) = \sum_z p_z \, D\left(p_{(x,y)|z} \,\middle\|\, p_{x|z} p_{y|z}\right)$$

✓ Mesure l'écart entre la distribution jointe conditionnelle et le produit des marginales conditionnelles, moyenné sur les valeurs de Z.

✓ $IM(X;Y|Z) = H(Y|Z) + H(X|Z) - H(X,Y|Z)$

Information mutuelle et causalité