Rapport Projet
Proba-Stat

Torri Atte
Le Montagner Roman

April 2020

1 Première Partie

Question de compréhension

Question 1.

- $P(C|i)$ = la probabilité d'associer i au cluster C
- $P(i|C)$ = la probabilité de trouver i dans le cluster C

Question 2.

I(C, i) ressemble à la formule de l'information mutuelle, donc la quantité d'information qu'on peut avoir sur une des variables en observant l'autre variable.

Question 3.

On aurait aucune limite sur la taille des données.

Question 4.

En minimisant I(C, i) on arrive à avoir une compression sur la taille des données.

Dérivation de l'équation

Premièrement

$$\frac{\partial P(C)}{\partial P(C'|k)} = \frac{\partial \sum\limits_{i} P(C|i)P(i)}{\partial P(C'|k)}$$

Dans le cas où $C \neq C'$ on a $\frac{\partial P(C)}{\partial P(C'|k)} = 0$

Dans le cas où $C = C'$ on a $\frac{\partial P(C)}{\partial P(C'|k)} = \sum\limits_{i} \frac{\partial P(C|i)P(i)}{\partial P(C|k)} = P(k)$

$$\frac{\partial I(C,i)}{\partial P(C'|k)} = \frac{\partial \frac{1}{N} \sum\limits_{i=1}^{N} \sum\limits_{C=1}^{N_C} P(C|i) log(\frac{P(C|i)}{P(C)})}{\partial P(C'|k)} = \frac{1}{N} \sum_{i=1}^{N} \sum_{C=1}^{N_C} \frac{\partial (P(C|i) log(\frac{P(C|i)}{P(C)}))}{\partial P(C'|k)}$$

$$= \frac{1}{N}(\sum_{i=1}^{N}\sum_{C=1}^{N_C}\frac{\partial(P(C|i)}{\partial P(C'|k)}log(\frac{P(C|i)}{P(C)})) + \sum_{i=1}^{N}\sum_{C=1}^{N_C}P(C|i)\frac{\partial log(\frac{P(C|i)}{P(C)}))}{\partial P(C'|k)})$$

$$= \frac{1}{N}(log(\frac{P(C'|k)}{P(C')})) + \sum_{i=1}^{N}\sum_{C=1}^{N_C}P(C|i)\frac{\partial log(\frac{P(C|i)}{P(C)}))}{\partial P(C'|k)})$$

$$= \frac{1}{N}(log(\frac{P(C'|k)}{P(C')})) + \sum_{i=1}^{N}\sum_{C=1}^{N_C}P(C|i)\frac{P(C)}{P(C|i)}\frac{\partial\frac{P(C|i)}{P(C)}}{\partial P(C'|k)}$$

$$= \frac{1}{N}(log(\frac{P(C'|k)}{P(C')})) + \sum_{i=1}^{N}\sum_{C=1}^{N_C}P(C)\frac{P(C)\frac{\partial P(C|i)}{\partial P(C'|k)} - P(C|i)\frac{\partial P(C)}{\partial P(C'|k)}}{P(C)^2}$$

$$= \frac{1}{N}(log(\frac{P(C'|k)}{P(C')})) + \sum_{i=1}^{N}\sum_{C=1}^{N_C}\frac{\partial P(C|i)}{\partial P(C'|k)} - \sum_{i=1}^{N}\sum_{C=1}^{N_C}\frac{P(C|i)\frac{\partial P(C)}{\partial P(C'|k)}}{P(C)}$$

$$= \frac{1}{N}(log(\frac{P(C'|k)}{P(C')})) + 1 - \sum_{i=1}^{N}\frac{P(C'|i)P(k)}{P(C')}$$

Donc, $\frac{\partial I(C,i)}{\partial P(C'|k)} = \frac{1}{N}(log(\frac{P(C'|k)}{P(C')})) + 1 - \frac{\sum_{i=1}^{N}P(C'|i)P(k)}{\sum_{j=1}^{N}P(C'|j)P(j)}$

Deuxièmement

On note $P(i|C) = \frac{P(C|i)P(i)}{P(C)}$

$$\frac{\partial P(i|C)}{\partial P(C'|k)} = \frac{\partial\frac{P(C|i)P(i)}{P(C)}}{\partial P(C'|k)}$$

$$= \frac{P(C)\frac{P(C|i)P(i)}{\partial P(C'|k)} - P(C|i)P(i)\frac{\partial P(C)}{\partial P(C'|k)}}{P(C)^2}$$

Si $C \neq C'$ alors $\frac{\partial P(i|C)}{\partial P(C'|k)} = 0$
Si $C = C'$ alors

$$\frac{\partial P(i|C)}{\partial P(C'|k)} = \frac{P(C)\frac{P(C|i)P(i)}{\partial P(C|k)} - P(C|i)P(i)P(k)}{P(C)^2}$$

Si $i \neq k$ alors

$$\frac{\partial P(i|C)}{\partial P(C'|k)} = -\frac{P(C|i)P(i)P(k)}{P(C)^2}$$

Si $i = k$ alors

$$\frac{\partial P(i|C)}{\partial P(C'|k)} = \frac{P(C)P(k) - P(C|k)P(k)^2}{P(C)^2}$$

On veut dériver $\frac{\partial\langle s\rangle}{\partial P(C'|k)}$

$$\frac{\partial\langle s\rangle}{\partial P(C'|k)} = \frac{\partial(\sum_{C=1}^{N_c} P(C)s(C))}{\partial P(C'|k)}$$

$$= \sum_{C=1}^{N_c} \frac{\partial(P(C)s(C))}{\partial P(C'|k)}$$

$$= \sum_{C=1}^{N_c} \frac{\partial P(C)}{\partial P(C'|k)} s(C) + \sum_{C=1}^{N_c} P(C)\frac{\partial s(C)}{\partial P(C'|k)}$$

Si $C' \neq C$ alors $\frac{\partial\langle s\rangle}{\partial P(C'|k)} = 0$ Si $C' = C$ alors

$$P(k)s(C') + P(C')\frac{\partial s(C')}{\partial P(C'|k)}$$

$$= ...$$

$$= P(k)(1-r)s(C') + P(k)\sum_{s=1}^{r} s(C'; i^{(s)} = k)$$

Finalement

On considère $\mathcal{F} = \langle s\rangle - TI(C,i)$. Pour maximiser cette fonction on veut trouver le point où la dérivée s'annule, donc

$$\frac{\partial\mathcal{F}}{\partial P(C'|k)} = \frac{\partial\langle s\rangle}{\partial P(C'|k)} - T\frac{\partial I(C,i)}{\partial P(C'|k)} = 0$$

$$\iff P(k)(1-r)s(C) + P(k)\sum_{r'=1}^{r} s(C; i^{(r')=k}) = T\frac{1}{N}\left(log(\frac{P(C|k)}{P(C)}) + 1 - \frac{\sum_i P(C|i)P(k)}{\sum_j P(C|j)P(j)}\right)$$

$$\iff \frac{1}{N}(1-r)s(C) + \frac{1}{N}\sum_{r'=1}^{r} s(C; i^{(r')=k}) = T\frac{1}{N}\left(log(\frac{P(C|k)}{P(C)}) + 1 - \frac{\sum_i P(C|i)\frac{1}{N}}{\sum_j P(C|j)\frac{1}{N}}\right)$$

$$\iff (1-r)s(C) + \sum_{r'=1}^{r} s(C; i^{(r')=k}) = T\left(log(\frac{P(C|k)}{P(C)})\right)$$

$$\iff P(C|k) = P(C)exp\left(\frac{1}{T}\left((1-r)s(C) + \sum_{r'=1}^{r} s(C; i^{(r')=k})\right)\right)$$

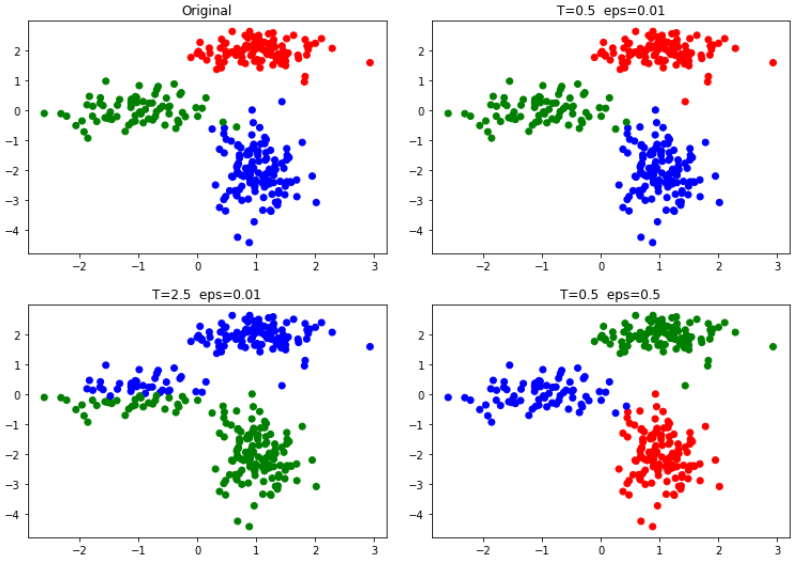
Similarité pour r=2

$$s(C) = \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} P(i_1|C)P(i_2|C)s(i_1, i_2)$$

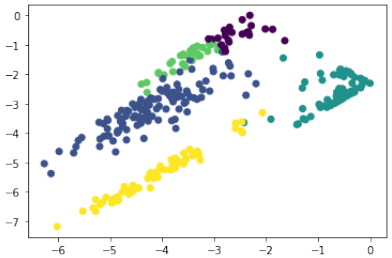
$$P(C|i) = \frac{P(C)}{Z(i;T)} exp\left(\frac{1}{T}\left(s(C; i^{(1)}) + s(C; i^{(2)}) - s(C)\right)\right)$$

Jeux de données

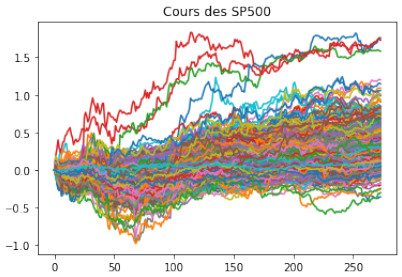
On remarque que l'algorithme marche très bien sur les données artificielles avec $N = 300$, $T = 0.5$ et $\epsilon = 0.01$. Par tests on a pu remarquer qu'il marches plutôt bien pour tout $T \in]0, 2[$, par contre pour $T \geq 2$ les données ne se clusterisent plus très bien. Finalement on remarque aussi que la valeur de ϵ a peu d'effet sur le résultat des clusters comparé à T, par exemple ci-dessous $\epsilon = 0.5$ avec $T = 0.5$.



Sur les données sans scatterplot, on a les meilleurs résultats de clustering avec 5 clusters.



On a ci-dessous un graphique représentant l'évolution des cours en bourse des entreprises de l'index SP500.



Nous avons testé plusieurs nombres de cluster différents, mais n'avons pas eu de résultat conclusif sur le nombre optimal de clusters. Voici ci-dessous un exemple avec $C = 15$ et $T = 15$