Clustering et étiquetage de clusters

Nicolas Dugué - Master ATAL

14 octobre 2020

K-moyennes

- Les données : $\{x_1, x_2, \cdots, x_n\}$ avec $\forall i, x_i \in \mathbb{R}^p$, les x_i les individus (documents) décrits par p attributs (termes), soit la matrice X à n lignes et p colonnes ;
- Ce qu'on veut minimiser : $||X - A \cdot \mu||^2$ avec
  - A une matrice d'affectation à n lignes et k colonnes t.q. $A_{ij} = 1$ si x_i appartient au cluster j, et 0 sinon ;
  - μ est une matrice à k lignes et p colonnes représentant les k centroïdes.

K-moyennes

- Les données : $\{x_1, x_2, \cdots, x_n\}$ avec $\forall i, x_i \in \mathbb{R}^p$, les x_i les individus (documents) décrits par p attributs (termes), soit la matrice X à n lignes et p colonnes ;
- Ce qu'on veut minimiser : $||X - A \cdot \mu||^2$

$$\mathcal{L}(A, \mu) = \sum_{i=1}^{n} \sum_{j=1}^{k} A_{ij} \sum_{d=1}^{p} (X_{id} - \mu_{kd})^2$$

K-moyennes

$$\underset{A,\mu}{\mathit{argmin}}\ \mathcal{L}(A,\mu) = \sum_{i=1}^{n}\sum_{j=1}^{k} A_{ij} \sum_{d=1}^{p} (X_{id} - \mu_{kd})^2$$

Méthode type *EM* :

- Initialisation aléatoire de μ puis, jusqu'à convergence :
  - Mettre à jour A en considérant μ fixée
  - Mettre à jour μ en considérant A fixée

K-moyennes

$$\underset{A,\mu}{argmin}\ \mathcal{L}(A,\mu) = \sum_{i=1}^{n}\sum_{j=1}^{k} A_{ij}\sum_{d=1}^{p}(X_{id}-\mu_{kd})^2$$

Méthode type *EM* :

- Initialisation aléatoire de μ puis, jusqu'à convergence :
  - Mettre à jour A en considérant μ fixée : $A_{ij} = \underset{j}{argmin}||X_i - \mu_j||_2$
  - Mettre à jour μ en considérant A fixée : Comment calculer la matrice μ optimale ?

K-moyennes : calculer μ optimal

$$\frac{\delta}{\delta \mu_{ml}} \mathcal{L} = \frac{\delta}{\delta \mu_{ml}} \sum_{i=1}^{n} \sum_{j=1}^{k} A_{ij} \sum_{d=1}^{p} (X_{id} - \mu_{kd})^2$$

$$= \frac{\delta}{\delta \mu_{ml}} \sum_{i=1}^{n} A_{im} \sum_{d=1}^{p} (X_{id} - \mu_{md})^2$$

$$= \frac{\delta}{\delta \mu_{ml}} \sum_{i=1}^{n} A_{im} (X_{il} - \mu_{ml})^2$$

$$= \sum_{i=1}^{n} A_{im} \frac{\delta}{\delta \mu_{ml}} (X_{il} - \mu_{ml})^2$$

$$= \sum_{i=1}^{n} A_{im} \frac{\delta}{\delta \mu_{ml}} (X_{il}^2 - 2 X_{il} \mu_{ml} + \mu_{ml}^2)$$

$$= \sum_{i=1}^{n} A_{im} (-2 X_{il} + 2 \mu_{ml})$$

K-moyennes : calculer μ optimal

$$\frac{\delta}{\delta \mu_{ml}} \mathcal{L} = \sum_{i=1}^{n} A_{im}(-2X_{il} + 2\mu_{ml}) = 0$$

$$2\sum_{i=1}^{n} A_{im}\mu_{ml} - 2\sum_{i=1}^{n} A_{im}X_{il} = 0$$

$$\sum_{i=1}^{n} A_{im}\mu_{ml} = \sum_{i=1}^{n} A_{im}X_{il}$$

$$\mu_{ml} = \frac{\sum_{i=1}^{n} A_{im}X_{il}}{\sum_{i=1}^{n} A_{im}}$$

Plan

1. Clustering
   - K-moyennes
   - NMF
2. Étiquetage des clusters
   - Avec les titres des documents
   - χ^2
   - Mutual Information
   - Feature F-mesure

NMF : Non Negative Matrix Factorization

$$\underset{W,H}{argmin} \, ||X - W \cdot H||$$

avec W et H qui ne contiennent que des valeurs positives ;

- Factorisation de matrice : lien avec approche Glove ;
- Lien avec le clustering ;
- Interprétabilité : W à n lignes et k colonnes la matrice document-thématique, et H à k lignes et n colonnes la matrice thématique-terme.

NMF : Non Negative Matrix Factorization

$$\underset{W,H}{argmin}\mathcal{L}(W,H) = \underset{W,H}{argmin}\sum_{i=1}^{n}\sum_{d=1}^{p}(X_{id} - \sum_{j=1}^{k} W_{ij}H_{jd})^2$$

Comment apprendre W et H ?
Initialiser aléatoirement W et H avec des valeurs positives et alterner :

- Mise-à-jour de W ;
- Mise-à-jour de H.

Mettre à jour : Gradient NMF

$$\frac{\delta}{\delta W_{ml}}\mathcal{L}(W,H) = \frac{\delta}{\delta W_{ml}}\sum_{i=1}^{n}\sum_{d=1}^{p}(X_{id} - \sum_{j=1}^{k} W_{ij}H_{jd})^2$$

$$= \frac{\delta}{\delta W_{ml}}\sum_{i=1}^{n}\sum_{d=1}^{p}(X_{id}^2 - 2X_{id}\sum_{j=1}^{k} W_{ij}H_{jd} + (\sum_{j=1}^{k} W_{ij}H_{jd})^2)$$

$$= \frac{\delta}{\delta W_{ml}}\sum_{d=1}^{p}(X_{md}^2 - 2X_{md}\sum_{j=1}^{k} W_{mj}H_{jd} + (\sum_{j=1}^{k} W_{mj}H_{jd})^2)$$

$$= \frac{\delta}{\delta W_{ml}}(X_{ml}^2 - 2X_{ml}\sum_{j=1}^{k} W_{mj}H_{jl} + (\sum_{j=1}^{k} W_{mj}H_{jl})^2)$$

$$= \frac{\delta}{\delta W_{ml}}(X_{ml}^2 - 2X_{ml}W_{ml}H_{jl} + (W_{ml}H_{jl})^2)$$

$$= -2X_{ml}H_{jl} + 2W_{ml}H_{jl}^2$$

Gradient NMF

$\frac{\delta}{\delta W_{ml}} \mathcal{L}(W, H) = -2X_{ml}H_{jl} + 2W_{ml}H_{jl}^2$
On a donc les gradients :

- $\nabla_W \mathcal{L}(W, H) = -2XH^t + 2WHH^t$
- $\nabla_H \mathcal{L}(W, H) = -2W^tX + 2W^tWH$

et des règles de mise-à-jour :

- $W \leftarrow W - \eta_W \cdot \nabla_W$
- $H \leftarrow H - \eta_H \cdot \nabla_H$

Mais...

Mais risque d'obtenir des valeurs négatives dans H et W

Gradient NMF

- $\nabla_W \mathcal{L}(W, H) = -2XH^t + 2WHH^t$
- $\nabla_H \mathcal{L}(W, H) = -2W^tX + 2W^tWH$

et des règles de mise-à-jour :

- $W \leftarrow W - \eta_W \cdot \nabla_W$
- $H \leftarrow H - \eta_H \cdot \nabla_H$

On fixe $\eta_W = \frac{W}{WHH^t}$ et obtient ainsi la règle :

$$W \leftarrow W - \eta_W \cdot \nabla_W = W + \frac{W}{WHH^t} \cdot (XH^t - WHH^t) = W \frac{XH^t}{WHH^t}$$

NMF

- Le gradient résoud tout !
- Jouer sur le pas d'apprentissage nous permet d'obtenir des règles qui garantissent la positivité ;
- Les problèmes d'apprentissage sont connectés : Glove et NMF... Mais aussi K-moyennes qui est un cas particulier de la NMF.

Plan

1. Clustering
   - K-moyennes
   - NMF
2. Étiquetage des clusters
   - Avec les titres des documents
   - χ^2
   - Mutual Information
   - Feature F-mesure

Étiqueter les clusters

Trouver des termes :

- qui décrivent les clusters ;
- qui sont typiques des clusters.

Représentatif VS Discriminant

En choisissant les titres des documents les plus proches des centroïdes

Plan

Étiqueter les clusters avec le χ^2

	Cluster 0	$\overline{Cluster0}$	
Chaussette	49	27.652	
$\overline{Chaussette}$	141	774.106	

La classe 0 et la variable *Chaussette* sont-elles indépendantes ?
Calculons les valeurs attendues dans le cas de l'indépendance.

Étiqueter les clusters avec le χ^2

	Cluster 0	$\overline{Cluster0}$	
Chaussette	49	27.652	27.701
$\overline{Chaussette}$	141	774.106	
	190		801.948

La classe 0 et la variable *Chaussette* sont-elles indépendantes ?
Calculons les valeurs attendues dans le cas de l'indépendance.

Étiqueter les clusters avec le χ^2

	Cluster 0		$\overline{Cluster0}$	
Chaussette	49	6,6	27.652	27.701
$\overline{Chaussette}$	141		774.106	
	190			801.948

La classe 0 et la variable *Chaussette* sont-elles indépendantes ?
Calculons les valeurs attendues dans le cas de l'indépendance.

Étiqueter les clusters avec le χ^2

	Cluster 0		$\overline{Cluster0}$	
Chaussette	49	6,6	27.652	27.701
$\overline{Chaussette}$	141		774.106	
			801.758	801.948

La classe 0 et la variable *Chaussette* sont-elles indépendantes ?
Calculons les valeurs attendues dans le cas de l'indépendance.

Étiqueter les clusters avec le χ^2

	Cluster 0		$\overline{Cluster0}$		
Chaussette	49	6,6	27.652	27.694,4	27.701
$\overline{Chaussette}$	141		774.106		
			801.758		801.948

La classe 0 et la variable *Chaussette* sont-elles indépendantes ?
Calculons les valeurs attendues dans le cas de l'indépendance.

Étiqueter les clusters avec le χ^2

	Cluster 0		$\overline{Cluster0}$		
Chaussette	49	6,6	27.652	27.694,4	
$\overline{Chaussette}$	141		774.106		774.247
	190				801.948

La classe 0 et la variable *Chaussette* sont-elles indépendantes ?
Calculons les valeurs attendues dans le cas de l'indépendance.

Étiqueter les clusters avec le χ^2

	Cluster 0		$\overline{Cluster0}$		
Chaussette	49	6,6	27.652	27.694,4	
$\overline{Chaussette}$	141	183,4	774.106		774.247
	190				801.948

La classe 0 et la variable *Chaussette* sont-elles indépendantes ?
Calculons les valeurs attendues dans le cas de l'indépendance.

Étiqueter les clusters avec le χ^2

	Cluster 0		$\overline{Cluster0}$		
Chaussette	49	6,6	27.652	27.694,4	
$\overline{Chaussette}$	141	183,4	774.106		774.247
			801.758		801.948

La classe 0 et la variable *Chaussette* sont-elles indépendantes ?
Calculons les valeurs attendues dans le cas de l'indépendance.

Étiqueter les clusters avec le χ^2

	Cluster 0		$\overline{Cluster0}$		
Chaussette	49	6,6	27.652	27.694,4	
$\overline{Chaussette}$	141	183,4	774.106	774.063,6	774.247
			801.758		801.948

La classe 0 et la variable *Chaussette* sont-elles indépendantes ?
Calculons les valeurs attendues dans le cas de l'indépendance.

Étiqueter les clusters avec le χ^2

	Cluster 0		$\overline{Cluster0}$	
Chaussette	49	6,6	27.652	27.694,4
$\overline{Chaussette}$	141	183,4	774.106	774.063,6

$$\chi^2 = \sum_d \frac{(O_d - E_d)^2}{E_d} = 284$$

Étiqueter les clusters avec le χ^2

	Cluster 0		$\overline{Cluster0}$	
Chaussette	49	6,6	27.652	27.694,4
$\overline{Chaussette}$	141	183,4	774.106	774.063,6

$$\chi^2 = \sum_d \frac{(O_d - E_d)^2}{E_d} = 284$$

Plus le χ^2 est élevé, moins l'hypothèse d'indépendance entre la classe et la présence de l'attribut est probable.

Plan

Étiqueter les clusters avec la Mutual Information

$$PMI = log\Big(\frac{P(x,y)}{P(x)P(y)}\Big)$$

À quel point x et y co-occurrent plus ou moins que par chance ?

MI (Mutual Information) : à quel point x apparait plus ou moins dans C qu'ailleurs

$$PMI = \sum_{x \in (Chaussette, \overline{Chaussette})} \sum_{c \in (Cluster0, \overline{Cluster0})} P(x,c) log\Big(\frac{P(x,c)}{P(x)P(c)}\Big)$$

	Cluster 0	$\overline{Cluster0}$
Chaussette	49	27.652
$\overline{Chaussette}$	141	774.106

Plan

Un exemple supervisé

Le corpus des conversations de ma grand-mère avec la voisine

- 2 patois : Le berrichon (Be), le bourbonnais (Bo) ;
- 6 conversations dans l'un de ces deux patois ;
- 3 mots et leur fréquence.

Tazon	Arcandier	Nigeasson	Classe
9	5	5	Be
9	10	5	Be
9	20	6	Be
5	15	5	Bo
6	25	6	Bo
5	25	5	Bo

Représentativité Versus Typicité

- $\mathbf{FP}_c(f) = \frac{W_c^f}{W_c} \longrightarrow$ **représentativité, dominance**

Représentativité Versus Typicité

- $\mathbf{FP}_c(f) = \frac{W_c^f}{W_c} \longrightarrow$ **représentativité, dominance**
- $\mathbf{FR}_c(f) = \frac{W_c^f}{W^f} \longrightarrow$ **typicité, saillance**

Représentativité Versus Typicité

- $\mathbf{FP}_c(f) = \frac{W_c^f}{W_c} \longrightarrow$ **représentativité, dominance**
- $\mathbf{FR}_c(f) = \frac{W_c^f}{W^f} \longrightarrow$ **typicité, saillance**
- **FF** la moyenne harmonique.

Un exemple supervisé

Tazon	Arcandier	Nigeasson	Classe
9	5	5	Be
9	10	5	Be
9	20	6	Be
5	15	5	Bo
6	25	6	Bo
5	25	5	Bo

$$W_{Be}^{Tazon} = 27$$

Un exemple supervisé

Tazon	Arcandier	Nigeasson	Classe
9	5	5	Be
9	10	5	Be
9	20	6	Be
5	15	5	Bo
6	25	6	Bo
5	25	5	Bo

$$W_{Be}^{Tazon} = 27$$

$$W^{Tazon} = 43$$

Un exemple supervisé

Tazon	Arcandier	Nigeasson	Classe
9	**5**	**5**	Be
9	**10**	**5**	Be
9	**20**	**6**	Be
5	15	5	Bo
6	25	6	Bo
5	25	5	Bo

$W_{Be}^{Tazon} = 27$

$W^{Tazon} = 43$

$W_{Be} = 78$

$$FR_{Be}(Tazon) = \frac{27}{43}$$

$$FP_{Be}(Tazon) = \frac{27}{78}$$

Sélection de variables

$$S_c = \left\{ f \in F | FF_c(f) > \overline{FF(f)}, FF_c(f) > \overline{FF} \right\}$$

avec

- $\overline{FF(f)}$ F-Mesure moyenne de f
- $\overline{FF}$ la F-Mesure moyenne

Sélection de variables

Tazon	Arcandier	Nigeasson	
0.46	0.39	**0.3**	$FF_{Be}(f)$
0.22	0.66	**0.24**	$FF_{Bo}(f)$
0.34	0.53	0.27	$\overline{FF(f)}$
0.38			$\overline{FF}$

$\rightarrow$ *Nigeasson* pas sélectionnée

Références

[ASL05] Shadi Al Shehabi and Jean-Charles Lamirel.

Multi-topographic neural network communication and generalization for multi-viewpoint analysis.

In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, volume 3, pages 1564–1569. IEEE, 2005.

[Lam12] Jean-Charles Lamirel.

A new approach for automatizing the analysis of research topics dynamics : application to optoelectronics research.

Scientometrics, 93(1) :151–166, 2012.