# Fouille de données et Medias Sociaux

# M2 DAC

TME 3. Diffusion

Ce TME a pour objectif de définir et expérimenter un des plus populaires modèles de diffusion: le modèle Independent Cascades. Il s'agit de mettre en place les mécanismes d'inférence et l'algorithme d'apprentissage des probabilités d'infection tel que défini dans [Saito et al., 2008] vu en cours.

## 1 Données

Deux fichiers de cascades ont été préparés pour expérimenter les modèles de diffusion: "/Vrac/lamprier/FDMS/cascades_train.txt" et "/Vrac/lamprier/FDMS/cascades_test.txt". Ces deux fichiers contiennent des episodes de diffusion issus d'une même distribution de probabilités et donc peuvent servir de jeux d'apprentissage et de test. Ces deux fichiers contiennent un episode de diffusion distinct par ligne, chaque ligne étant composée d'une liste de couples d'infection: <utilisateur>:<temps d'infection>

Habituellement, les modèles type IC se basent sur un graphe de relations explicites prédéfini. Cependant il a été montré que la diffusion ne suit pas forcément ces relations connues. Par ailleurs, elles ne sont pas toujours disponibles. Pour simplifier nous choisissons ici de nous baser sur le graphe complet de relations possibles dans le graphe. Pour alléger les calculs cependant, vous prendrez soin de ne considérer que les relations ayant au moins un exemple de diffusion possible dans l'ensemble d'apprentissage, car les autres relations obtiendraient de toutes façons une probabilité de 0 en fin de processus d'apprentissage.

## 2 IC: Inférence

En considérant un graphe de relations avec des probabilités de diffusion $\theta_{i,j}$ connues, il s'agit de mettre en place le mécanisme du modèle Independent Cascade, dans lequel chaque nouvel utilisateur infecté tente d'infecter chacun de ses successeurs dans le graphe à l'itération suivante. Pseudo-code de l'inférence à partir d'un ensemble de sources $S_0$:

1. $S_t = \emptyset$

2. Pour tous les utilisateurs $i \in S_{t-1}$
   - Pour tous les utilisateurs j successeur de i dans le graphe et tel que j par encore infecté avant $t$, avec une probabilité $\theta_{i,j}$ : $S_t \leftarrow S_t \cup j$
3. $t = t + 1$
4. Si $S_{t-1} \neq \emptyset$, retour en 1.

Comme l'algorithme d'inférence n'est pas déterministe, il s'agit de relancer cette simulation un certain nombre de fois et de considérer le nombre de fois où chauqe utilisateur est infecté pour extraire leur probabilité d'infection sachant les sources.

## 3 IC: Apprentissage

L'apprentissage des probabilité d'IC se fait par un algorithme EM où l'on cherche à maximiser la vraissemblance des paramètres $\theta$ connaissant les probabilités à l'étape précédente $\hat{\theta}$:

1. $\hat{\theta} = Random$
2. Tant qu'on n'a pas atteint des paramètres stables (ou que la vraissemblance augmente) :
   (a) Pour tout $D \in \mathcal{D}$: calculer la probabilité de tout $u \in D$ d'être infecté au temps $t_u^D$ selon les paramètres courants $\hat{\theta}$:
   $$\hat{P}_{t_u^D}(u) = 1 - \prod_{v \in Preds(u) \wedge t_v^D = t_u^D - 1} 1 - \hat{\theta}_{v,u}$$
   (b) On pose l'espérance de vraissemblance :
   $$Q\left(\theta; \hat{\theta}\right) = \sum_{D \in \mathcal{D}} \Phi^D\left(\theta; \hat{\theta}\right) + \sum_{\substack{(u,v), u \in D \wedge v \in Succs(u) \wedge \\ ((v \notin D) \vee (v \in D \wedge t_v^D > t_u^D + 1))}} \log\left(1 - \theta_{u,v}\right)$$
   Avec $\Phi^D\left(\theta; \hat{\theta}\right) =$
   $$\sum_{\substack{(u,v) \in D^2, v \in Succs(u) \\ \wedge t_v^D = t_u^D + 1}} \frac{\hat{\theta}_{u,v}}{\hat{P}_{t_v^D}(v)} \log(\theta_{u,v}) \; + \; (1 - \frac{\hat{\theta}_{u,v}}{\hat{P}_{t_v^D}(v)}) \log(1 - \theta_{u,v})$$
   (c) On maximise :
   $$\theta^* = \arg\max_{\theta} Q\left(\theta; \hat{\theta}\right)$$
   (d) $\hat{\theta} = \theta^*$

A chaque étape de l'algorithme, la maximisation de la vraissemblance se fait en annulant dérivée et réalisant la mise à jour suivante pour tous les couples $i, j$ avec au moins un exemple de possible diffusion dans l'ensemble d'apprentissage:

$$\theta^*_{u,v} = \frac{\sum_{D \in \mathcal{D}^+_{u,v}} \frac{\hat{\theta}_{u,v}}{\hat{P}_{t^D_v}(v)}}{|\mathcal{D}^+_{u,v}| \ + \ |\mathcal{D}^-_{u,v}|} \text{ Avec :}$$

$$\mathcal{D}^+_{u,v} = \{D \in \mathcal{D} | (u,v) \in D^2 \wedge t^D_v = t^D_u + 1\}$$
$$\mathcal{D}^-_{u,v} = \{D \in \mathcal{D} | u \in D \wedge ((v \notin D) \vee (v \in D \wedge t^D_v > t^D_u + 1))\}$$

Il est à noter qu'IC ne considère que des infections contigües: chaque infection ne peut être expliquée que par les infectés à l'étape précédente, ce qui réduit grandement les exemples d'apprentissage et paraît peu réaliste si l'on considère des données issues de réseaux sociaux. Nous proposons alors de considérer plutôt la proabilité d'infection et les ensembles suivants pour la mise à jour des paramètres:

$$\hat{P}_{t^D_u}(u) = 1 - \prod_{v \in Preds(u) \wedge t^D_v < t^D_u} 1 - \hat{\theta}_{v,u}$$

$$\mathcal{D}^+_{u,v} = \{D \in \mathcal{D} | (u,v) \in D^2 \wedge t^D_v > t^D_u\}$$
$$\mathcal{D}^-_{u,v} = \{D \in \mathcal{D} | u \in D \wedge v \notin D\}$$

Bien que de tels ensembles ne permettent plus de prédire les temps d'infection en inférence (délais d'infection uniformes), ils permettent généralement d'obtenir de meilleurs résultats pour la prédiction de l'ensemble de utilisateurs infectés finaux.

## 4 Evaluation

Plusieurs mesures d'évaluation peuvent être considérées. On propose ici d'envisager une mesure de précision moyenne MAP, classique en recherche d'information pour évaluer les listes de documents retournés par les moteurs de recherche. En considérant pour chaque cascade de test les utilisateurs infectés à la première itération de la diffusion comme les sources, il s'agit de prédire une probabilité d'infection pour tous les autres infectés de la diffusion supérieure à celle des utilisateurs non infectés. Soit pour chaque épisode $D$, la liste de tous les utilisateurs $U^D$ ordonnée en ordre décroissant selon les probabilités d'infection définies par notre algorithme d'inférence, on considèrera alors la mesure MAP définie de la manière suivante:

$$MAP = \frac{1}{\mathcal{D}} \sum_{D \in |\mathcal{D}|} \frac{1}{|D|} \sum_{i=1}^{|U^D|} \frac{|\{U^D_1 .. U^D_i\} \cap D|}{i}$$