Patrons de Connexió al CV de la UOC

Armand Adroher Salvia
UOC - TFG - GEI - EDM&LA

26 de gener del 2015

DADES REBUDES

- Camps

$$\langle \texttt{user\_id}, \texttt{session\_start}, \texttt{last\_request}, \texttt{session\_expiration} \rangle$$

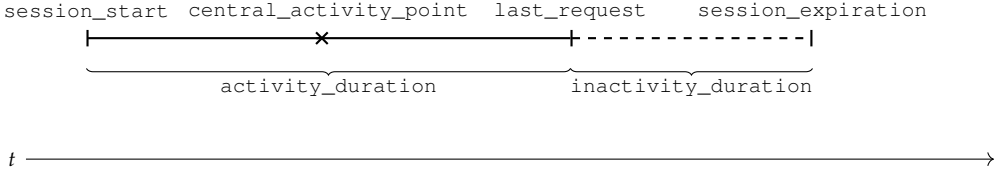
```
1 7149084242663;18/09/2013 00:00:03;18/09/2013 00:57:28;18/09/2013 02:04:32
2 6059394219413;18/09/2013 00:00:04;18/09/2013 00:00:15;18/09/2013 01:07:21
3 4139154106177;18/09/2013 00:00:07;18/09/2013 00:31:29;18/09/2013 01:39:07
4  858883854230;18/09/2013 00:00:07;18/09/2013 00:00:12;18/09/2013 01:07:21
```

- Dimensions

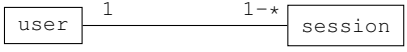
$$\#\langle \ldots \rangle \approx 8\,\text{M} \qquad \min(\texttt{session\_start}) \in \texttt{2013-09-18}$$
$$\#\texttt{user\_id} \approx 75\text{K} \qquad \max(\texttt{session\_expiration}) \in \texttt{2014-01-15}$$

MODEL DE DADES

- Sessions



- Usuaris

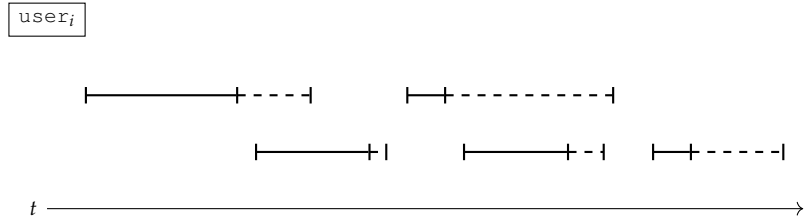


ANOMALIES

- Subsessions negatives

$$l(\texttt{activity\_duration}) < 0\,\text{s} \vee l(\texttt{inactivity\_duration}) < 0\,\text{s}$$

- Solapament de sessions



OBJECTIU GENERAL

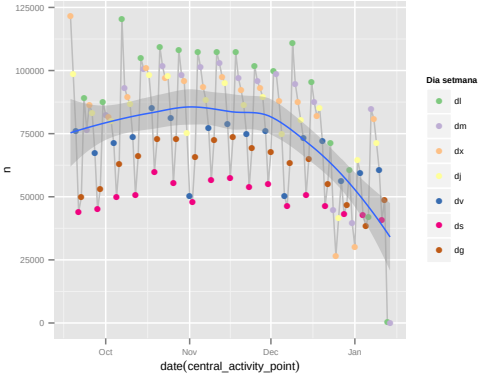
- Posar en pràctica la EDM&LA
  - KDD i ML per a millorar els processos d'aprenentatge
  - En el context del CV de la UOC

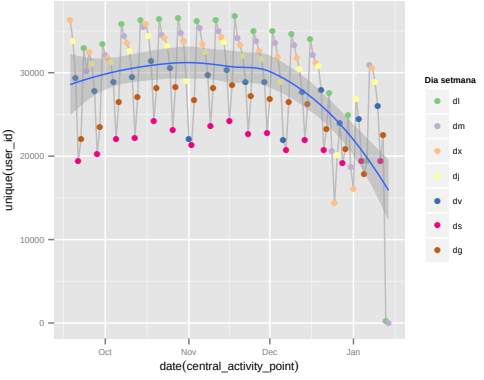
DIFICULTATS

- Procés de generació de les dades
  - Registre (*log*) de servidor?
  - Agregat en lots (*batch*) o en temps real?
- Tipus d'usuaris
  - Estudiants?
  - Consultors, PAS, *alumni*, etc.?
- Comportament de l'usuari
  - Accions en `session_start` i `last_request`
  - Què més?

DISTRIBUCIÓ DE VALORS

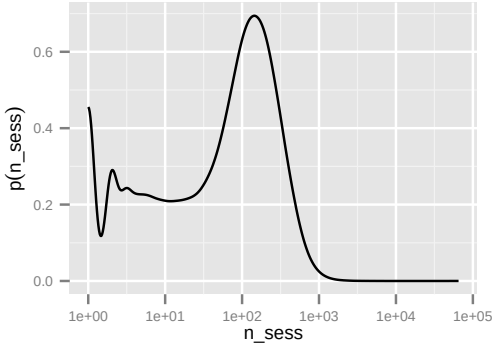
- Patró general de comportament
- Consonància amb el sentit comú

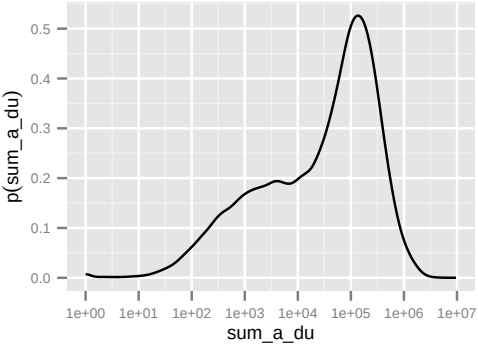




DISPERSIÓ DE VALORS

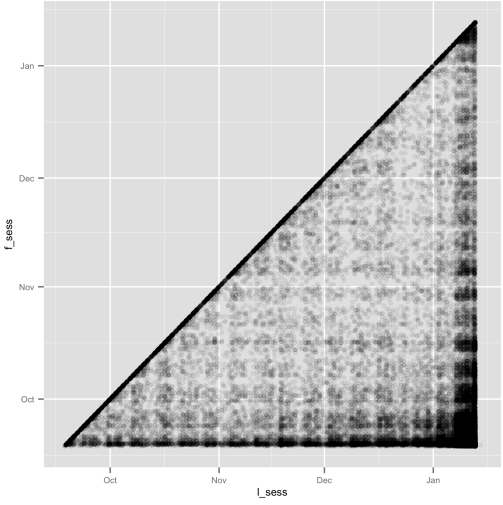
- Molts valors petits
- Molt pocs valors molt grans





SUPERACIÓ DE LES DIFICULTATS

- Limitar-se a la informació rebuda
- Inici $\rightarrow$ final de la interacció
- Presència per unitat de temps



NOUS ATRIBUTS

- Discretització temporal

$$\texttt{semester} = I_0, \ldots, I_m \qquad t_0 \in I_0 \qquad I_j = [t_0 + dj, t_0 + d(j+1)]$$

- Atributs booleans

$$\texttt{user}_i = \langle \texttt{id}_i, a_{i,0}, \ldots, a_{i,m} \rangle \quad (a_{i,j} \in \{0,1\})$$

$$a_{i,j} = \begin{cases} 1 \text{ si } \{v \in \texttt{sessions}_i : \phi(v,j)\} \neq \emptyset \\ 0 \text{ altrament} \end{cases}$$

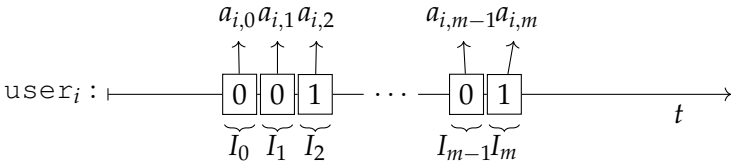
$$\phi(v,j) = (v.\texttt{session\_start} \leq t_0 + d(j+1)) \wedge (t_0 + dj \leq v.\texttt{last\_request})$$

- Dos valors de d

$$d \leftarrow 1 \text{ dia} \qquad\qquad d \leftarrow 1 \text{ setmana}$$

NOUS OBJECTES

- Usuari i-èssim $\mapsto$ seqüència de valors booleans



- $a_{i,j} = 1$ si i només si l'usuari i-èssim *ha estat present al CV* durant I_j

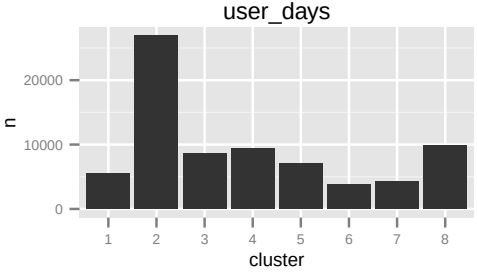
K-MEANS

- k-means (*Lloyd*) sobre

$$a_{i,0}, \ldots, a_{i,m} \in \{0,1\}^{m+1}$$

- Tipus de comportaments

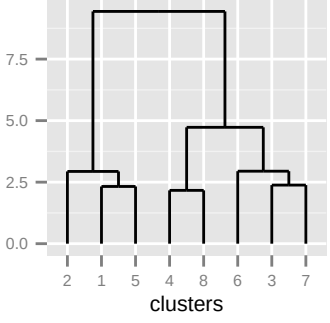
$$\mu_1, \ldots, \mu_k \Rightarrow \text{ caracterització}$$



- Agregació jeràrquica sobre

$$\mu_1, \ldots, \mu_k$$

- Enllaç complet

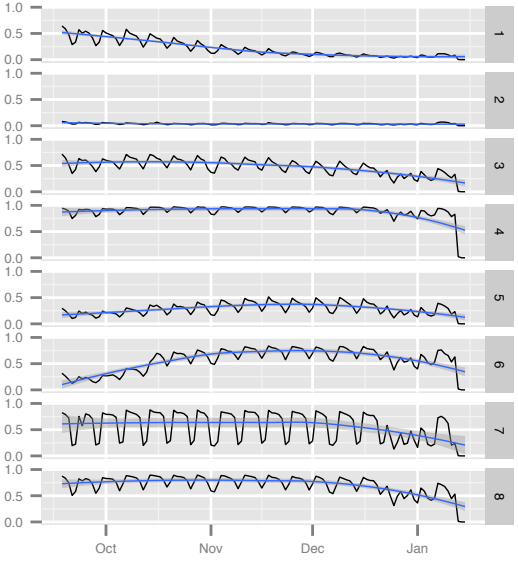


PATRONS

- Tipologia de comportaments

A. **Absents** $\{1, 2, 5\}$ (52.2%)
   a. **Esporàdics** $\{2\}$
   b. **Capbussadors** $\{1\}$
   c. **Tímids** $\{5\}$

B. **Presents** $\{3, 4, 6, 7, 8\}$ (47.8%)
   a. **Endarrerits** $\{6\}$
   b. **Fatigats** $\{3\}$
   c. **Setmanaris** $\{7\}$
   d. **Persistents** $\{3, 7\}$
      i. **Dedicats** $\{8\}$
      ii. **Obsessionats** $\{4\}$



LIMITACIONS

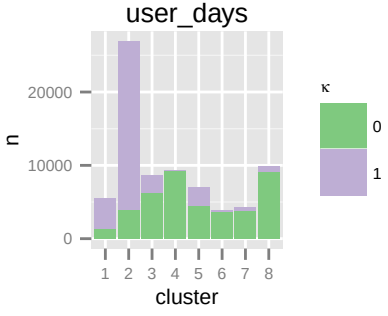
- A partir de dades pobres

$$P(a_{i,j} = 1 | \texttt{user}_i \in \texttt{students}) \quad ?$$

- *A posteriori*
  Després del final del semestre
  $\Rightarrow$ No es pot actuar en temps real

- Prediccions poc significatives

$$\kappa_i = \begin{cases} 1 \text{ si } \texttt{l\_sess}_i < Q_2(\texttt{l\_sess}) \\ 0 \text{ altrament} \end{cases}$$



CADENES DE MARKOV

- $a_0, \ldots, a_m$ seqüència d'estats $\{0, 1\}$

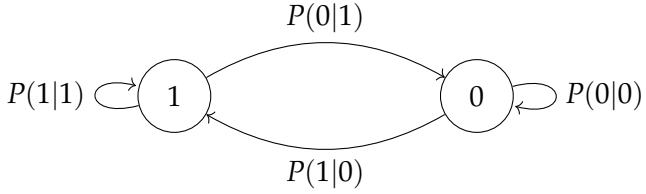
$$\begin{aligned} \texttt{user}_1 &\mapsto \quad 00001011101001\ldots \\ &\vdots \\ \texttt{user}_n &\mapsto \quad \underbrace{11001010010101\ldots}_{m+1} \end{aligned}$$

- Representació de $\texttt{user}_i$

$$\begin{pmatrix} P(0|0) & P(1|0) \\ P(0|1) & P(1|1) \end{pmatrix} = \begin{pmatrix} 1-\alpha & \alpha \\ 1-\beta & \beta \end{pmatrix}$$

Paràmetres a estimar: α, β

- DTMC homogènia de grau 1

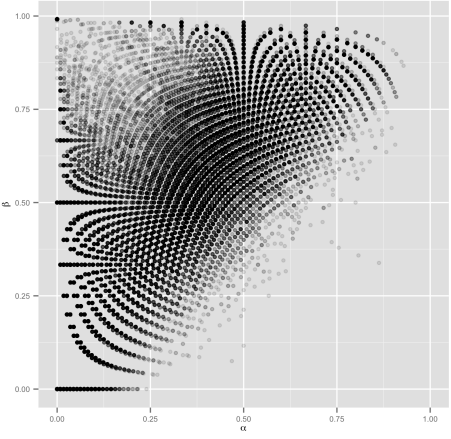


Estimació per a DTMC

- Estimador $\widehat{p}_{x,y}$

$$x, y \in \{0, 1\}$$
$$n_{x,y} = \text{transicions } x \to y$$
$$\widehat{p}_{x,y} = \frac{n_{x,y}}{\sum_{z \in \{0,1\}} n_{x,z}}$$
$$\alpha = \frac{n_{0,1}}{\sum_{z \in \{0,1\}} n_{0,z}} \qquad \beta = \frac{n_{1,1}}{\sum_{z \in \{0,1\}} n_{1,z}}$$

- Resultats dispersos



Model ocult de Markov

- Estats ocults $S = \{A, Q\}$

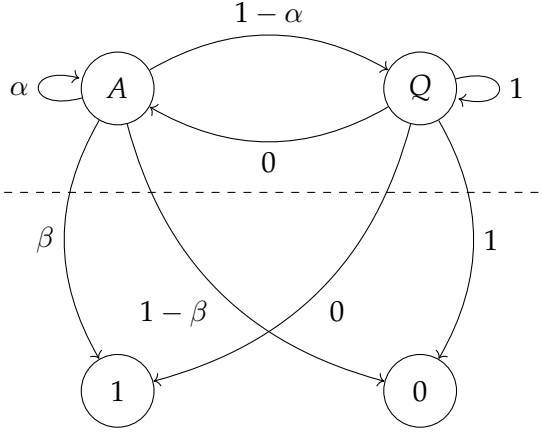
$x_{i,j} = Q \Leftrightarrow \texttt{user}_i$ ha abandonat el CV

- Observacions $a_{i,j} \in \{0, 1\}$
- Transicions

$$\begin{pmatrix} P(Q|Q) & P(A|Q) \\ P(Q|A) & P(A|A) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1-\alpha & \alpha \end{pmatrix}$$

- Emissions

$$\begin{pmatrix} P(0|Q) & P(1|Q) \\ P(0|A) & P(1|A) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1-\beta & \beta \end{pmatrix}$$



ESTIMACIÓ PER A HMM

- Per a α, prenem λ tal que $Q_2(\texttt{l\_sess}) \in I_\lambda$

$$\begin{aligned}
\begin{pmatrix} P(Q) & P(A) \end{pmatrix} \begin{pmatrix} P(Q|Q) & P(A|Q) \\ P(Q|A) & P(A|A) \end{pmatrix}^\lambda &= \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1-\alpha & \alpha \end{pmatrix}^\lambda \\
&= \begin{pmatrix} 1-\alpha^\lambda & \alpha^\lambda \end{pmatrix} \\
\alpha^\lambda = \frac{1}{2} \quad \Rightarrow \quad \alpha &= \left(\frac{1}{2}\right)^{\frac{1}{\lambda}}
\end{aligned}$$

- Per a β, prenem u tal que $\texttt{user}_i\texttt{.l\_sess} \in I_u$

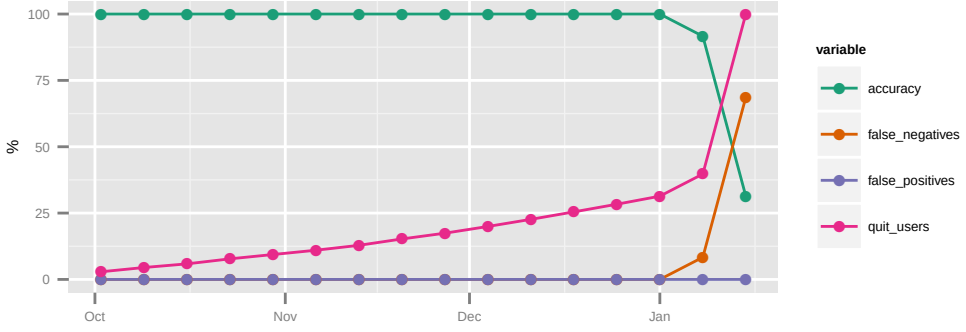
$$\texttt{user}_i\texttt{.a\_rate} = \frac{1}{u+1} \sum_{j=0}^{u} a_{i,j}$$

$$\beta = \frac{1}{|\texttt{users}|} \sum_{v \in \texttt{users}} v\texttt{.a\_rate}$$

PREDICCIÓ DE L'ABANDONAMENT

- En temps real
  Per a cada interval I_j i cada usuari i-èssim

$$\texttt{viterbi}(HMM(\alpha,\beta),\langle a_{i,0},\ldots,a_{i,j}\rangle) \quad = \quad \langle x_{i,0},\ldots,x_{i,j}\rangle$$
$$\texttt{user}_i \text{ ha abandonat} \quad \Leftrightarrow \quad x_{i,j} = Q$$



CONCLUSIONS

- La discretització en presència/absència per interval temporal és útil per a dotar de significat dades com aquestes.
- L'agregació per k-means permet fer un esbós *a posteriori* de la tipologia d'usuaris segons el patró de connexió.
- La predicció *a priori* de l'abandonament del CV per mitjà de models ocults de Markov és eficaç en el cas de la discretització setmanal.