ARMAND ADROHER SALVIA

UOC - TFG - EDM&LA

PATRONS DE CONNEXIÓ AL CV DE LA UOC

Aquesta és la memòria lliurada en motiu de la realització del *Treball de Fi de Grau* en el marc del programa *Educational Data Mining and Learning Analytics* durant el transcurs del semestre de tardor del curs 2014-2015.

Índex

Índex de figures

Índex de taules

Resum

La disciplina de la *Educational Data Ming and Learning Analytics* té per objecte emprar els mètodes propis de la *Descoberta de Coneixement en Bases de Dades* i l'*Aprenentatge Computacional* amb la finalitat de comprendre i millorar, si s'escau, els processos que tenen lloc en entorns d'aprenentatge. En aquest estudi es parteix d'un registre d'establiment i clausura de sessions dels usuaris al *Campus Virtual* de la *UOC* per a mirar d'obtenir resultats en aquesta direcció. Amb la finalitat d'ésser fidels a la informació que proporcionen les dades, en proposo una discretització que mesura la presència o absència, al *Campus Virtual*, de cada un dels usuaris, en funció de determinats intervals temporals. Això m'ha permès aplicar mètodes d'agregació que han donat lloc a un esbós de tipologia de comportament d'usuaris en el transcurs del semestre. En darrer lloc, el resultat central d'aquest estudi és l'obtenció d'un mètode simple per a l'estimació de models ocults de Markov que siguin capaços de predir, per a cada moment del desplegament temporal del semestre i amb un alt grau d'encerts, el trencament de la relació amb el *Campus Virtual* per part de cada usuari.

1
Determinació d'objectius

Encetem aquest estudi tot establint-ne el punt de partida i enunciant l'objectiu genèric d'anàlisi de dades que l'ha motivat.

1.1 Motivació de l'estudi

En motiu de la realització d'aquest treball de fi de grau (TFG), la UOC em va fer arribar un fitxer de text pla[1] que consistia en la representació d'una taula mitjançant el format *Comma Separated Values* (CSV). Presenta una estructura molt simple, en cada una de les línies de text que el componen hi trobem 4 valors. A partir de la informació que va acompanyat-ne el lliurament podem afirmar que:

[1] El fitxer en qüestió acompanya el lliurament d'aquesta memòria amb el nom `20131.con.txt` .

(1) Aquestes dades són extretes dels registres de sessions establertes al programari web que implementa el Campus Virtual de la UOC (CV).

(2) Els 4 camps a què corresponen cada un dels valors respectius són:

   (a) Una seqüència de dígits decimals, que és un identificador de l'usuari que ha establert aquesta sessió.

   (b) Les 3 marques de temps (*timestamp*) següents[2]:

      (i) La primera indica el moment en què s'ha iniciat la sessió al CV en nom d'aquest usuari.

      (ii) La segona indica el moment en què s'ha realitzat la darrera petició al CV de la UOC dins d'aquesta sessió.

      (iii) La darrera consisteix en el moment en què ha caducat la sessió.

[2] Si emprem la notació que admet la funció `strftime`, totes les marques de temps del fitxer segueixen el format "`%d/%m/%Y %H:%M:%S`", és a dir, tenen una precisió d'1 segon

A mode d'exemple, la figura 1.1 es mostra les primeres quatre línies d'aquest fitxer.

```
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
```

Figura 1.1: Primeres quatre línies del fitxer `20131.con.txt`

El programa en què s'emmarca aquest TFG proposava com a objectiu general aplicar, a aquest conjunt de dades les tècniques pròpies les disciplines d'*Educational Data Mining* i *Learning Analytics* (EDM&LA). En concordança llur naturalesa, aquest objectiu genèric es podria fer explícit, doncs, de la manera següent:

> Emprar les tècniques pròpies del *Descobriment de Coneixement en Bases de Dades* (KDD) i de l'*Aprenentatge Computacional* (ML) en aquest conjunt de dades amb la finalitat d'obtenir coneixement que millori els processos que tenen lloc en entorns d'aprenentatge (en aquest cas, el CV de la UOC).

En quins objectius concrets ha acabat reduint-se aquesta directiva general ha estat condicionat per part dels dos factors següents:

(1) Les característiques de les dades mateixes.

(2) El grau de coneixement que he pogut obtenir sobre el procés que les ha produïdes.

1.2 Naturalesa de les dades rebudes

Proporciono, a continuació, un esbós de quina és la primera abstracció que realitzarem a partir d'aquestes dades, així com també un resum de quina és la dimensió del conjunt que en resulta.

Abstracció elemental

Amb la finalitat de facilitar la notació en l'exposició, convenim que anomenarem `session` la classe dels objectes representats per part de cada una de les entrades del fitxer `20131.con.txt`, i `sessions` la relació que en resulta. Per altra banda, denotarem de la manera següent els atributs de l'esquema que la compon[3]:

[3] Noti's que s'hi ha afegit un atribut, `id`, amb un identificador únic per a assegurar-nos que es compleix el model relacional.

```
session(id, user_id, session_start, last_request, session_expiration)
```

Interpretarem que les seqüències de caràcters que integren les dades rebudes representen valors que pertanyen als tipus següents:

(1) $\texttt{id} \in \mathbb{N}$

(2) $\texttt{user\_id} \in ([\texttt{"0"},\texttt{"9"}] \cup \texttt{" "})^{13}$

(3) $\texttt{session\_start} \in \texttt{timestamp}$

(4) $\texttt{last\_request} \in \texttt{timestamp}$

(5) $\texttt{session\_expiration} \in \texttt{timestamp}$

On $([\texttt{"0"},\texttt{"9"}] \cup \texttt{" "})^{13}$ són les seqüenciès de longitud 13 construïbles mitjançant els dígits numèrics i l'espai, i `timestamp` és la marca de temps amb precisió de segons a què m'he referit anteriorment. Per altra banda, també definirem els següents atributs derivats:

(6) $\texttt{total\_duration} := [\texttt{session\_expiration}, \texttt{session\_start}) \in \texttt{interval}$

(7) $\texttt{activity\_duration} := [\texttt{last\_request}, \texttt{session\_start}) \in \texttt{interval}$

(8) $\texttt{inactivity\_duration} := [\texttt{session\_expiration}, \texttt{last\_request}) \in \texttt{interval}$

(9) $\texttt{central\_activity\_point} := (\texttt{session\_start} + \texttt{activity\_duration})/2 \in \texttt{timestamp}$

On `iterval` representa un interval tancat de valors consecutius del tipus `timestamp`.

En darrer lloc, també utilitzarem la funció l definida de la manera següent[4]:

[4] És a dir, la llargada, en segons, d'un valor del tipus `interval`

$$\begin{aligned} l : \texttt{interval} &\to (\mathbb{N} \times \mathrm{s}) \\ [t_0, t_1] &\mapsto t_1 - t_0 \end{aligned}$$

Podem dir que cada sessió es compon, doncs, de dos segments temporals consecutius. En la figura 1.2 se'n mostra una representació gràfica.

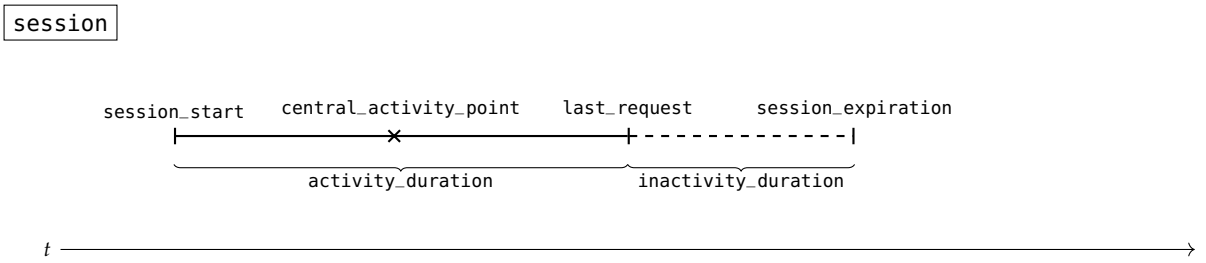


Figura 1.2: Diagrama del significat de cada un dels atributs pel que fa l'objecte `session`.

Amb tot, noti's que a cada un dels usuaris que han establert les sessions de què tenim notícia, els quals són designats unívocament per mitjà de `user_id`, li correspon un conjunt de sessions. Com és evident, doncs, podem prendre `user_id` com a clau forània. En la figura 1.3 es mostra el senzill diagrama de classes que representa aquesta relació.

Figura 1.3: Diagrama de classes de la relació entre usuaris i sessions.

Candidats de tipus objectes a modelar

Després d'aquest primer cop d'ull, aturem-nos un moment a fer una llista de les classes d'objectes que podem abstraure a partir d'aquestes dades. La finalitat d'aquest exercici és donar un primer pas per a la concreció dels objectius del nostre estudi.

(1) **Sessió**: Evidentment, en interpretar el contingut de `20131.con.txt` com una relació, ja estem considerant que les mateixes sessions són un possible objecte a modelar. De fet, podem considerar que són la classe primària que se'ns hi presenta, essent-ne les altres derivades.

(2) **Cicles temporals**: Ja sigui per causes biològiques o socioculturals, el comportament dels humans presenta ritmes que segueixen cicles temporals. En aquest cas, estem considerant els humans que prenen part en el procés d'aprenentatge que té lloc en l'entorn educatiu del CV de la UOC. Per tant, aquestes unitats temporals poden ser un bon marc per a l'anàlisi de llur comportament. Podem identificar-ne, com a mínim, els següents:

   (a) **Dia**, amb parts com les hores, matí, tarda, nit, etc.

   (b) **Setmana**, amb parts com els dies, laboral/festiu, etc.

   (c) **Semestre**, és a dir, la totalitat del rang temporal en què es situen les dades rebudes.

   La caracterització d'aquests cicles vindrà donada, doncs, tant per la presència o absència de com per característiques de les sessions que corresponen a cada una de les seves parts.

(3) **Usuari**: Si, com hem dit, els usuaris són identificats per mitjà de `user_id`, podem caracteritzar el comportament de cada un d'ells en funció dels dos tipus d'objectes anteriors, això és, de les sessions que ha establert i dels cicles temporals als quals aquestes pertanyen.

Dimensions bàsiques

Donem ara una idea general de les dimensions de les dades a tractar.

El fitxer `20131.con.txt` en format text conté 661.17 MB d'informació[5] i consta de 8837062 línies, que corresponen als casos que constitueixen la població del nostre estudi.

[5] En codificació de caràcters ASCII.

Per altra banda, en la taula 1.1 s'indica quines són les cardinalitats dels respectius conjunts de valors diferents corresponents a cada un dels atributs de `session`.

atribut a	$\lvert\{x : x = \mathrm{val}(a)\}\rvert$
`id`	8837062
`user_id`	75609
`session_start`	5059939
`last_request`	5072362
`session_expiration`	2135158

Taula 1.1: Cardinalitat de valors distints que prenen cada un dels atributs dels objectes de la classe `session`.

Adverteixi's que les sessions han estat generades per part de 75608 usuaris diferents.

En darrer lloc l'abast temporal de les dades és comprès entre els dos valors següents[6]:

$$\min(\texttt{session\_start}) = \texttt{2013-09-18 00:00:03}$$
$$\max(\texttt{session\_expiration}) = \texttt{2014-01-15 21:57:13}$$

[6] Observi's que el format de les marques de temps que he utilitzat en aquest cas no és pas el de les cadenes de text rebudes en el fitxer de dades, sinó l'estàndard ISO 8601. A fi de seguir la notació més comuna, a partir d'ara ho faré així.

Com es pot veure, el rang temporal de les dades va des de la tercera setmana de setembre del 2013 fins a mitjans de gener del 2014. Aquest període de poc menys de 4 mesos correspon de manera aproximada al desplegament temporal d'un semestre de tardor dels cursos ordinaris de la UOC.

1.3 Grau de coneixement del procés de generació de les dades

Allò que ens interessa en un estudi com aquest no són les dades mateixes, sinó més aviat els esdeveniments que, en tenir lloc, n'han provocat l'aparició, a mode de rastres o indicis. És a causa d'aquest fet que es fa necessari interpretar-les. A l'hora de poder saber satisfactòriament quin n'és el significat, és a dir, quina informació proporcionen i en quina mesura ho fan, és especialment valuosa la recol·lecció del màxim de coneixement possible sobre el procés que les ha generades.

Desafortunadament en realitat no he disposat de gaire més informació pel que fa aquesta qüestió que l'escueta descripció que ja s'ha proporcionat en la subsecció anterior. Sabem que cada una de les entrades del fitxer correspon unívocament a la sessió d'un usuari en el servidor del CV. D'aquest fet podem deduir, com a mínim, les següents conclusions:

(1) Atès que hom pot interactuar amb el CV de la UOC per mitjà d'un navegador web, algun tram de les connexions corresponents a part de les sessions establertes en la comunicació entre el client i el servidor ha tingut lloc per mitjà del *Hypertext Transfer Protocol* (HTTP).

(2) Aquest és un protocol sense estat (*stateless protocol*)[**?** , cap. 2], és a dir, es tracta cada una de les transaccions[7] com quelcom independent d'aquelles que el mateix client ha iniciat anteriorment. Aquesta dificultat és sovint eludida per mitjà de galetes (*cookies*) residents en el client, que l'identifiquen en cada transacció, o per mitjà de camps ocults en formularis enviats per mitjà del mètode `POST`.

(3) Dels dos punts anteriors se'n dedueix, doncs, que `last_request` tan sols indica quin és el segon en què el servidor del CV ha rebut la darrera petició per part del client que `user_id` identifica. Noti's que això no implica que l'usuari que li correspon *no hagi fet res*, en relació al procés d'aprenentatge, durant l'interval `inactivity_duration`[8]. En altres paraules, solament tenim notícia d'aquests dos moments d'interació entre client i servidor, però poca cosa més.

[7] És a dir, cada un dels parells de petició-resposta presos com una unitat atòmica.

[8] Podria per, exemple, haver passat una bona estona llegint un document llarg

En aquest sentit, és especialment rellevant tenir ben present *quina informació no proporcionen* aquestes dades. Altrament, correríem el risc d'obtenir conclusions errònies tot construint arguments correctes a partir de suposicions infundades. Enumerem doncs quina és la

informació que podríem caure en l'error de pensar que proporcionen però que en realitat no ho fan.

(1) En primer lloc, notem que no sabem res sobre el procés intern per mitjà del qual es recullen aquestes dades. És a dir, no sabem si consisteix en un *log* escrit en temps real per part d'un servidor únic (o del node director d'un *cluster*) o bé si són el resultat d'una agregació processada per lots (*batch*) dels registres de nodes distints que executen instàncies diferents del programari del servidor del CV.

    Tenim raons per a pensar que ens trobem en aquesta segon situació. Per exemple, en el cas que exposen Masip et al.[**?**], les dades d'accés al CV de la UOC són obtingudes a partir d'una combinació dels registres dels diferents servidors *front-end* en què s'executa el programari que l'implementa. Aquesta combinació s'efectua per mitjà d'un procés per lots diàriament durant la matinada. Així, doncs, hem de comptar amb les particularitats d'un procés com aquest, en concret, amb la possibilitat de l'existència d'inconsistències en les dades.

(2) D'altra banda, tampoc no sabem a quina mena d'usuari correspon cada un dels identificadors. És a dir, si bé és correcte considerar que una part considerable dels identificadors d'usuari correspon a estudiants que es troben seguint un curs ordinari durant el semestre, no ho és pas suposar que ho fan tots. De fet, no podem suposar ni que són majoria. I és que cal tenir present que, a part d'estudiants, les sessions representades en les dades rebudes poden haver estat generades per part d'altres tipus d'usuaris. És raonable pensar, com a mínim, en els següents:

    (a) Consultors, professors i altre personal docent.

    (b) Personal purament acadèmic o de recerca.

    (c) Membres de l'administració i serveis.

    (d) Alumnes que no s'han matriculat en aquest semestre però que estan seguint alguns estudis a la UOC.

    (e) Exalumnes de la UOC que, com a membres de *UOC Alumni*, tinguin accés al CV.

    (f) *Bots* que estableixin sessions amb el servidor com a resultat de l'execució de tasques automàtiques.

    No hem rebut cap informació de part de l'organització del programa d'aquest treball de fi de grau (TFG) pel que fa aquest aspecte. L'única cosa que podem afirmar amb seguretat és que aquests usuaris són els agents d'aquestes sessions en el CV.

(3) En tercer lloc, cal notar que tampoc tenim cap informació sobre quins són els fusos horaris a què corresponen els següents elements en joc:

(a) No coneixem el fus horari a definit en la configuració de cada un dels sistemes en què s'executa el programari que ha generat aquestes dades.

Suposaré, no obstant, que es tracta del mateix per a totes les entrades, a saber:

$$\texttt{CET} = \texttt{UTC+01:00}$$

(b) No sabem tampoc quin és el fus horari corresponent la localització geogràfica des d'on ha tingut lloc la connexió per part de cada un dels usuaris que corresponen a clients controlats per humans, que han establert les sessions en cada cas. Ni tan sols disposem d'informació més general sobre aquest aspecte. En concret, no sabem quina és la proporció de sessions que s'estableixen des del fus `CET`.

Aquest fet provoca que tots els estudis d'anàlisi de dades referents al comportament de les sessions en funció del *moment del dia* corrin el risc d'ésser considerablement esbiaixiats. Com es veurà tot seguit, he renunciat a emetre cap judici ferm en aquest sentit.

2
Neteja i transformació de dades

En el marc del cicle de vida del KDD, el procés de tranformació i neteja de les dades en brut ha d'estar exclusivament al servei de l'objectiu que ens haguem proposat. Començarem, per tant, tot fent una descripició inicial de les dades rebudes. Aquesta consistirà, en primer lloc, en detectar-ne les possibles anomalies i considerar l'eliminació o normalització dels casos defectuosos. Seguidament reconeixerem quin dels seus trets poden suposar un obtacle per a assolir els nostres objectius i indicarem, si s'escau, quines són les transformacions adients per tal de superar-los.

Marques de temps incompletes

D'entre les entrades del fitxer `20131.con.txt`, 370 (0.004%) contenen, com a valor d'algun dels seus camps, una marca de temps que fretura de la indicació horària[1]. D'altra banda, no hi ha cap de les altres instàncies que presenti valors del tipus `timestamp` que caiguin exactament en el segon `00:00:00`. Per tant, suposaré que en les primeres s'hi representa precisament aquest instant.

[1] És a dir, que segueix el patró "`%Y-%m-%d`"

Subsessions de durada negativa

Si definim el conjunt:

$$\texttt{neg} = \{v \in \texttt{sessions} : l(v.\texttt{activity\_duration}) < 0\,\text{s} \vee l(v.\texttt{inactivity\_duration}) < 0\,\text{s}\}$$

aleshores $|\texttt{neg}| = 467$ (0.005%). El fet aparentment absurd que hi hagi parts de les sessions que constin com a tenint una durada negativa s'explica fonamentalment per dues causes. Exposem-les:

(1) En 47 dels elements de `neg` és el cas que $l(\texttt{activity\_duration}) < 0$. En totes elles tant `session_start` com `last_request` cauen entre les 2:00 i les 3:00 del matí del dia 27/10/2013, que és quan va tenir lloc, aquell any, el pas de l'horari d'estiu al d'hivern. Això no obstant, si bé resulta que aquesta duració negativa es deu a aquest canvi, *això no significa que totes les distorsions introduïdes per part d'aquest canvi hagin provocat duracions negatives*. Poden donar-se, per exemple, situacions en què:

   (a) Dues sessions constin com a simultànies i no ho siguin (sessions v_0 i v_1 de la figura 2.1).

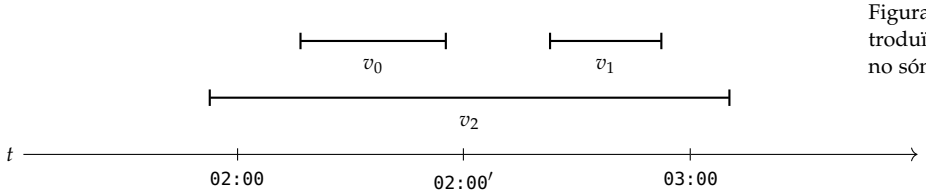


Figura 2.1: Exemples de distorsions introduïdes per part del canvi horari que no són detectables.

(b) Una sessió hagi començat abans de les 2:00 o hagi acabat després de les 3:00 y aquest fet la faci constar com havent durat una hora menys del que en realitat durà (sessió v_2 de la figura 2.1).

Les sessions que compleixen la condició següent (és a dir, aquelles que són problemàtiques a causa d'aquest fet), són 1003 (0.01%):

$$\begin{aligned}(&\texttt{session\_start} \in [\texttt{2013-10-27 02:00:00}, \texttt{2013-10-27 03:00:00}))\\ &\vee(\texttt{last\_request} \in [\texttt{2013-10-27 02:00:00}, \texttt{2013-10-27 03:00:00}))\\ &\vee(\texttt{session\_start} < \texttt{2013-10-27 02:00:00} \wedge \texttt{last\_request} < \texttt{2013-10-27 03:00:00})\end{aligned}$$

Davant la impossibilitat de destriar les distorsions provocades per aquest fenomen i de les dades que en realitat són fiables, així com de l'escàs pes que tenen en la totalitat del conjunt de dades, he pres la decisió d'eliminar solament aquelles en què $l(\texttt{activity\_duration})$ és negatiu, amb la finalitat de simplificar les operacions que es durant a terme més endavant.

(2) En 431 casos (els conjunts definits per part de les dues condicions anteriors no són disjunts), el que s'ha esdevingut és que

$$l(\texttt{inactivity\_duration}) < 0\,\mathrm{s}$$

Això no obstant, en totes elles resulta que

$$|l(\texttt{inactivity\_duration})| < 5\,\mathrm{s}$$

i, de fet, en la majoria dels casos la diferència és solament d'un segon. Suposarem que això és causat per decalatges en l'arribada de peticions al servidor o, en general, a altres errors de funcionament intern. En aquests casos assumirem que

$$\texttt{last\_request} = \texttt{session\_expiration}$$

(3) En darrer lloc, només hi ha una sessió amb subsessions de durada negativa que no s'expliqui per cap d'aquestes dues causes. Serà igualment eliminada.

3
Estadística descriptiva

Atenguem-nos ara a la distribució dels valors dels diferents atributs dels objectes candidats a modelar que hem apuntat en la secció 1.2. Tot seguint el mateix ordre expositiu, començarem veient, de primer, les característiques de les sessions en general, de certs cicles temporals en segon lloc i, finalment, dels usuaris a què pertanyen.

3.1 Característiques dels atributs de les sessions

Comencem donant un cop d'ull a la distribució dels diferents valors dels atributs de l'objecte `session`. En la taula 3.1 es mostren els alguns dels percentils (100-quantils) dels valors dels atributs `activity_duration` i `inactivity_duration`. Observem-hi el següent:

100-quantil k-èssim	$\lfloor k/100 \cdot \lvert\texttt{sessions}\rvert \rfloor$	100-$q_{k\%}(l(\texttt{a\_du}))$	100-$q_{k\%}(l(\texttt{ina\_du}))$
1	88365	0	0
5	441829	3	0
10	883659	9	0
25	2209148	41	0
50	4418297	203	4023
75	6627446	1039	4049
90	7952935	3401	4064
95	8394765	5241	4074
99	8748229	10869	4083

Taula 3.1: Percentils rellevants dels valors dels atributs de `session`. Els valors de les dues darreres columnes es mostren en segons. En la notació de les capçaleres `a_du` = `activity_duration` i `ina_du` = `inactivity_duration`.

(1) Pel que fa `activity_duration` la distribució dels valors és especialment desigual. La figura 3.1 mostra una estimació de la funció de densitat de probabilitat (PDF) [**?**][1]. Observem dues característiques especialment rellevants en la distribució dels valors d'aquest atribut.

   (a) En primer lloc la proporció de sessions amb una durada desmesuradament curta és notablement gran. A la taula 1 es mostra el nombre de instàncies (n_k) i la freqüència (f_k) acumulades per a 6 k-valors de `activity_duration`. Tornarem a aquesta qüestió més endavant, però de moment noti's que

[1] Noti's addicionalment que aquesta estimació s'ha calculat en tots els casos mitjançant la funció `density` del paquet base del programari R [**?**], amb el nucli gaussià que empra per defecte.

un 30% de les sessions presenten una duració d'activitat d'un minut o menys.

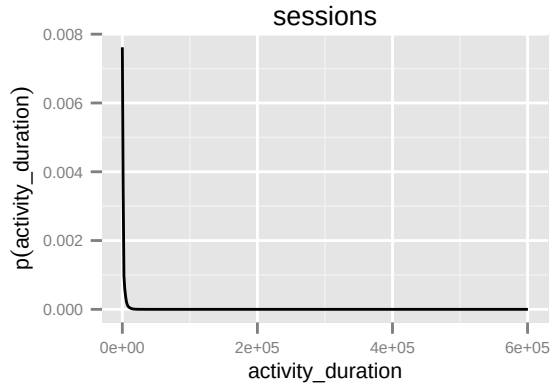


Figura 3.1: Estimació de la PDF de `activity_duration`

(b) En segon lloc, val a dir que també presenta valors extrems per la dreta. Si bé el percentil 99 és a 10 879 s (3.02 h), noti's que a partir d'aquest punt els seus valors prenen valors desmesuradament alts. La taula 3.3 mostra els 5 valors més alts que pren aquest atribut.

Podem mirar de prescindir d'aquests valors extrems i observar la distribució resultant si definim la condició c

$$c := l(\texttt{activity\_duration}) > 60\,\text{s} \\ \wedge\, l(\texttt{activity\_duration}) < 100\text{-}q_{95\%}(l(\texttt{activity\_duration}))$$

i obtenim la relació $\texttt{trim\_sessions} \subseteq \texttt{sessions}$, de les tuples que la compleixen

$$\texttt{trim\_sessions} = \sigma_c(\texttt{sessions})$$

Noti's que aleshores $|\texttt{trim\_sessions}| = 5713445$ (el 64.5% de la cardinalitat de `sessions`).

D'altra banda, en la figura 3.2 es mostra l'estimació de la PDF de `trim_sessions.activity_duration`.

Observi's que, tot i que la taxa de decreixement de la probabilitat és menor, aquesta gràfica presenta un aspecte similar al de la figura 3.1. La moda de `trim_sessions.activity_duration` en segueix essent el valor mínim i, per tant, el pic de la distribució es situa a 61 s.

(2) La distribució del valors de `inactivity_duration` és tota una altra. Com es pot veure en la taula 3.1 el seu valor és 0 s fins a haver superat amb escreix el 25è percentil (recordi's que hem igualat a 0 les durades negatives en aquest cas). En la figura 3.3 es mostra una estimació de la PDF dels valors d'aquest atribut.

És fàcil reconèixer dos pics, que, de fet, divideixen les sessions en dues classes:

$l(\texttt{a\_du}) \leq k$	n_k	f_k
0	219359	2.5%
1	337924	3.8%
5	618676	7 %
10	979324	11.1 %
30	1872169	21.2 %
60	2681220	30.34 %

Taula 3.2: Distribució dels valors més petits de $l(\texttt{activity\_duration})$. Les abreviacions dels noms dels atributs són les mateixes que en la taula 3.1. Com es pot observar, La freqüència (f_k) s'expressa en percentatges.

i	x_i
⋮	⋮
$n-4$	429742
$n-3$	512625
$n-2$	520141
$n-1$	562338
n	600528

Taula 3.3: Els 5 valors més grans de `activity_durtation`.

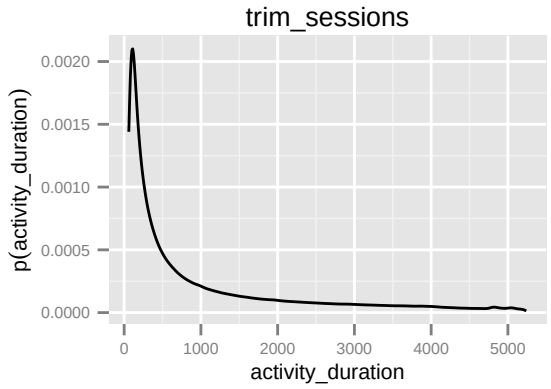


Figura 3.2: Estimació de la densitat de probabilitat dels valors `trim_sessions.activity_duration`

(a) La d'aquelles sessions que s'han tancat per mitjà de la sortida manual per part de l'usuari.

(b) La d'aquelles que han caducat automàticament després d'un temps estipulat en el programari del CV. Sembla que aquest temps ronda els 4008.2 s, que és la mitjana de la durada d'aquest atribut si prescindim dels valors nuls.

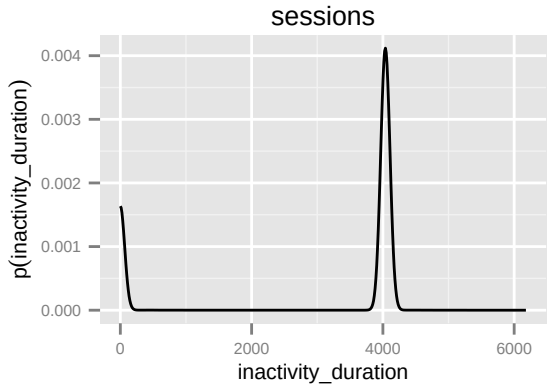


Figura 3.3: Estimació de la PDF de $l($`inactivity_duration`$)$.

3.2 *Descripció dels cicles temporals*

Analitzem tot seguit les distribucions de valors dels atributs dels objectes que corresponen als diferents cicles temporals que hem abstret en la secció 1.2.

El dia

El primer d'aquests cicles temporals és del dia en què tenen lloc les diferents sessions[2].

La gràfica present en la figura 3.4 inclou les corbes corresponents a les dues variables següents:

[2] Les dades resultants dels resums d'aquesta subsecció es troben al fitxer `session_days.csv`

(1) Per una banda, `n_sessions` representa el nombre de sessions que tenen `central_activity_point` en l'hora del dia indicada per part de la variable independent com a proporció del màxim del seu rang, és a dir, com a percentatge de

$$\max(|\{v \in \texttt{sessions} : \texttt{hour}(v.\texttt{central\_activity\_point}) = x\}|)$$

on la funció `hour` retorna l'hora del dia a què pertany un valor del tipus `timestamp`.

(2) Similarment, `n_users` expressa el nombre d'usuaris diferents que han establert sessions amb `central_activity_point` igual al valor de l'eix de les abscisses també com a proporció del màxim del seu rang.

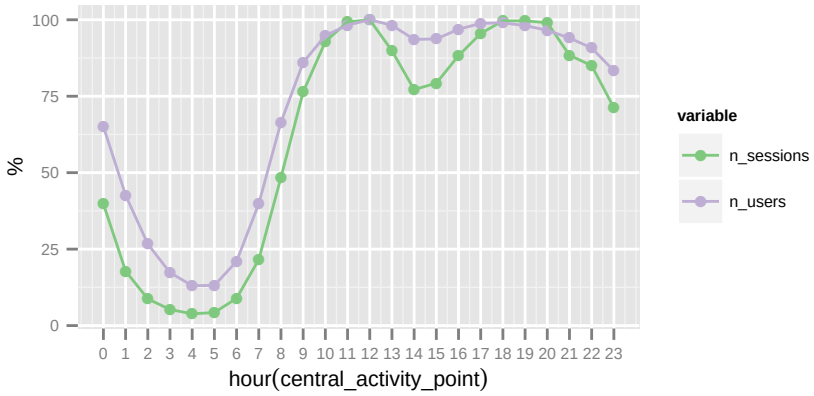


Figura 3.4: Proporció de sessions i usuaris per hora del dia en què s'esdevé `central_activity_point`.

Les dues presenten uns trets generals similars:

(1) Hi veiem dos pics, un al voltant de les 12 h i l'altre a l'entorn de l'intèrval $[18\,\text{h}, 20\,\text{h}]$.

(2) Per altra banda, dues valls, un mínim local a les 14 h i un d'absolut a les 4 h.

L'evolució que presenten els valors d'aquestes dues variables és completament compatible amb la idea que ens podríem haver format intuïtivament dels hàbits diaris de connexió dels estudiants de la UOC. Notem-ne el següent:

(1) En primer lloc, sembla que els dos punts de màxima activitat es troben a mig matí i a darrera hora de la tarda. Aquest esquema encaixa tant amb els estudiants que es dediquen a l'estudi a jornada completa i paren per dinar, com amb aquells que només hi dediquen o bé el matí o bé la tarda.

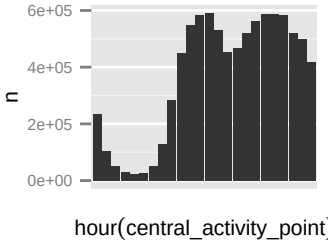
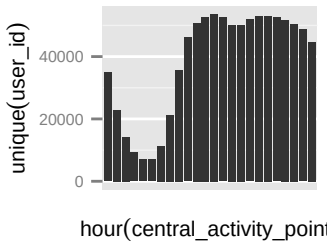


Figura 3.5: Nombre de sessions per hora del dia en què es dóna `central_activity_point`.



(2) Per altra banda, el mínim absolut situat a les hores mitjes de la matinada encaixa amb la suposició segons la qual a aquestes hores la majoria d'estudiants estan dormint.

(3) El desglossament d'aquestes variables en funció dels dies laborables i dels que pertanyen al cap de setmana no manifesta, no obstant, cap diferència notable respecte la totalitat dels casos. En els histogrames presents en les figures 3.7 i 3.8 se'n pot observar les distribucions respectives.

    Com es pot veure, l'evolució general és la mateixa. Ara bé, el cap de setmana, el màxim absolut pel que fa al nombre de sessions té lloc a la tarda, i el mínim relatiu de l'hora de dinar és més pronunciat.

(4) En darrer lloc hi ha un altre fet rellevant a tenir en compte que resulta d'aquesta distribució per hora del dia. En el capítol 2 em referia a la manca d'informació sobre quin és el fus horari corresponent a la localització geogràfica del client que estableix la sessió en cada cas. La distribució dels valors de `central_activity_point` en funció de l'hora del dia encaixen amb el patró dels ritmes diaris que hom pot observar, com a mínim, en els llocs geogràfics següents:

    (a) Pel que fa el fus `CET=UTC+01:00`, és coherent aquella que podem observar a l'Estat Espanyol. Pel que sé, als altres territoris que hi pertanyen tant els àpats com el moment de descans nocturn acostuma a tenir lloc una o dues hores abans.

    (b) També encaixa amb els ritmes propis dels territoris europeus que es troben situats en el fus `WET=UTC+00:00`, ja que el ritme diari també és desfasat una o dues hores abans en relació a l'espanyol.

    Certament, desconeixem quina és la proporció d'usuaris del CV que resideixen habitualment al Regne Unit, Irlanda o Portugal, d'entre aquells que ho fan a l'estranger. Si no en són una part extraordinàriament gran, aleshores es pot concloure que en realitat la gran majoria s'hi connecten des de territori espanyol.

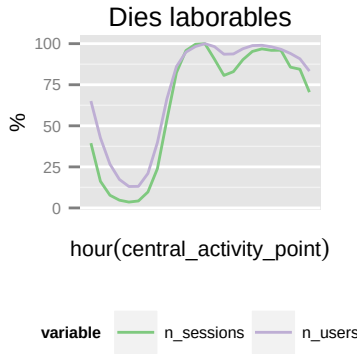


Figura 3.7: Proporció de sessions per hora del dia en què es dóna `central_activity_point` de dilluns a divendres.

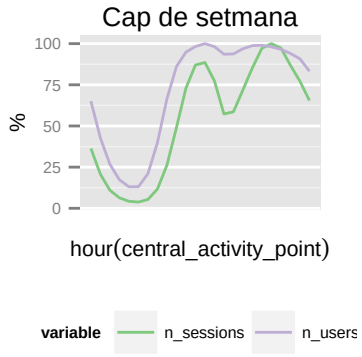


Figura 3.8: Proporció de sessions per hora del dia en què es dóna `central_activity_point` els caps de setmana.

La setmana

Observem tot seguit el comportament de les mateixes variables en funció dels dies de la setmana (figura 3.9)[3].

Altra vegada, les dues corbes presenten una estructura similar tot i que en aquest cas el nombre de per dia de la setmana presenti trets més accentuats. Com també podríem haver suposat *a priori*, tant aquest valor com el d'usuaris únics davalla durant el cap de setmana, essent el dissabte el dia amb menys activitat.

Per a una distribució en nombres absoluts de les variables consulti's els histogrames respectius de les figures 3.10 i 3.11.

[3] Les dades resultants dels resums d'aquesta subsecció es troben al fitxer `session_weeks.csv`

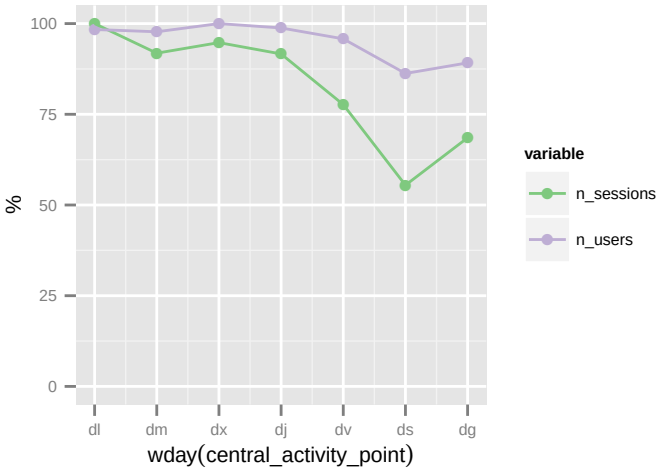


Figura 3.9: Nombre de sessions per dia de la setmana què es dóna `central_activity_point`.

La interpretació de la diferència entre els comportaments de les dues variables té a veure amb el fenomen que ja hem observat en la subsecció anterior sobre els comportaments horaris que presenten en dies laborables (figura 3.7) i el cap de setmana (figura 3.8), respectivament. És raonable creure que hi ha una proporció considerable d'usuaris que tot i que segeuixen visitant el CV durant el cap de setmana (tal com havien fet durant els dies laborables), en aquest cas ho fan amb menys intensitat.

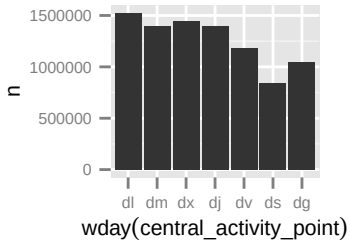


Figura 3.10: Nombre de sessions per dia de la setmana què es dóna `central_activity_point`.

El semestre

Completem ara la nostra inspecció inicial dels cicles temporals tot atenent-nos al semestre en la seva totalitat[4]. Això consisteix en posar la nostra atenció en les característiques de les seves diferents parts en funció de la presència o absència de sessions i/o usuaris, i dels trets que aquestes presenten.

Per a fer-ho, observem els resultats que s'exposen en les gràfiques de les figures 3.12 i 3.13, respectivament. Hi trobem un diagrama de punts i línies que relaciona el dia del semestre amb cada una de les dues variables que hem vingut observant fins ara. En el primer cas, les ordenades indiquen el nombre de sessions l'atribut `central_activity_point` de les quals cau en la data que correspon al valor de la variable independent. En el segon allò que indica la variable dependent és el nombre d'usuaris únics que són els agents d'aquestes sessions.

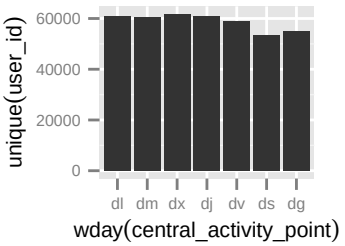


Figura 3.11: Nombre d'usuaris únics per dia de la setmana.

En tots dos casos, els punts mostren els valors obtinguts i la línia grisa fa palesa llur precedència cronològica. A més, com es pot veure, el color dels punts indica el dia de a setmana a què corresponen. En darrer lloc, noti's que en tots tres casos també s'hi ha sobreimprès, en color blau, la corba resultant de l'aplicació de la regressió local (LOESS) [**?**][5], envoltada d'una zona ombrejada que representa l'evolució de l'amplada de l'interval de confiança.

[4] Les dades resultants dels resums d'aquesta subsecció es troben al fitxer `session_semester.csv`

[5] El càlcul concret d'aquesta regressió no paramètrica s'ha dut a terme per mitjà de la funció `loess` del paquet `stats` de programari R [**?**]

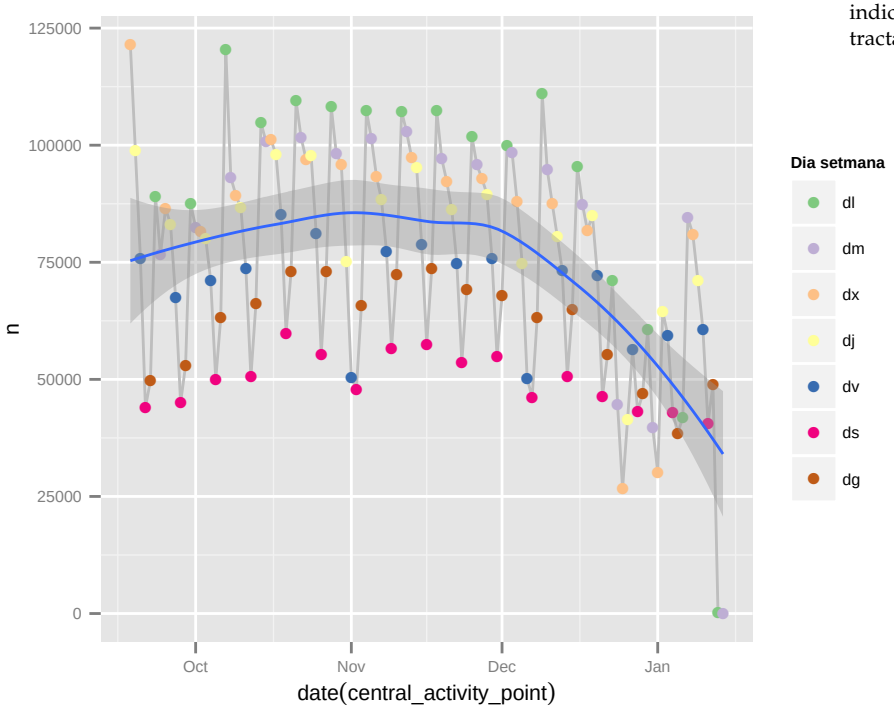


Figura 3.12: Nombre de sessions per dia del semestre. El color dels punts indica el dia de la setmana de què es tracta.

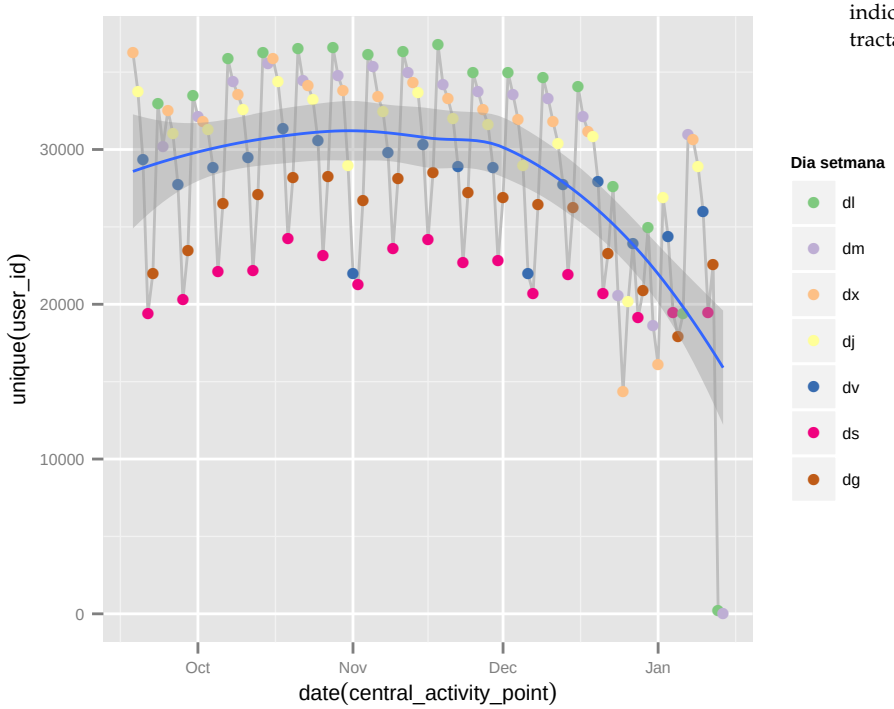


Figura 3.13: Nombre d'usuaris únics per dia del semestre. El color dels punts indica el dia de la setmana de què es tracta.

Dels dos resultats, cal notar-ne el següent:

(1) La dispersió dels valors s'explica per la variació del dia de la setmana a què pertanyen, la qual esdevé una raó suficient per a haver-nos plantejat la setmana com una unitat temporal digna d'interès, tal com hem fet en la secció 1.2. Si parem atenció al cicle [`dl,dg`], veurem que, en quasi tots els casos, las distribució dels valors és coherent amb els resultats setmanals que es mostren en la gràfica de la figura 3.9. La regularitat d'aquest patró només es trenca a partir de la darrera setmana de desembre, que és el moment en què comencen les festes de Nadal, durant les quals és raonable pensar que s'alteren els hàbits diaris dels usuaris.

(2) En tots dos casos la corba de la LOESS mostra una tendència similar. Podem distingir-hi els tres trams següents:

   (a) Un primer tram d'evolució tímidament creixent des de l'inici de curs fins a primers de novembre.

   (b) Un de segon que ocupa tot el novembre. Tot i que en general el pendent és negatiu és tant a prop del zero que el podem veure com un altiplà amb un lleu mínim local.

   (c) Un darrer tram que s'inicia amb l'entrada del desembre i que presenta un decreixement accelerat fins al final del semestre.

   Tot i que en tot moment hem de tenir present que no sabem del cert quina és la proporcio d'usuaris que són estudiants, llur comportament general també és coherent amb la imatge preconcebuda que ens proporciona el sentit comú sobre quina és l'evolució de la relació que mantenen amb l'activitat acadèmica que segueixen i, per tant, amb el CV. Pel que fa l'esquema general preconcebut sobre l'evolució del curs acadèmic, podem interpretar així els tres trams que acabem d'identificar:

   (a) El primer tram consisteix en un *període d'incorporació* dels estudiants. Ja sigui perquè n'hi ha que no inverteixen una gran dedicació a l'estudi just a principi de curs, perquè sempre es deixa un temps prudencial fins el lliurament de la primera *Prova d'Avaluació Continuada* (PAC) o perquè n'hi ha que, fruit d'una ampliació de matrícula, amplien la presència al CV, durant aquest tram l'activitat augmenta progressivament.

   (b) En el segon hi trobem un *màxim d'activitat*. Així doncs, consisteix en el nucli del curs, on la suma de la dedicació de tots els estudiants és la més alta.

   (c) En el tercer hi podem identificar el que podríem anomenar com un *període de garbellat i clausura*. En aquest cas, tots aquells estudiants que, per les raons que sigui, ja tenen clar que no superaran una part o la totalitat de les assignatures a què estan matriculats, redueixen la seva activitat al CV o

l'abandonen completament. Cal tenir present, no obstant, que durant el període comprès entre la segona setmana de gener i el final de semestre l'activitat acadèmica no exigeix una atenció constant al CV fins i tot per a aquells estudiants que han seguit el curs satisfactòriament.

3.3 *Característiques dels atributs dels usuaris*

Per a tancar aquesta secció mirem de fer un estudi preeliminar sobre el darrer dels objectes a modelar que hem apuntat en la secció 1.2, a saber, els usuaris, que representarem per mitjà de l'objecte `user`[6]. Com fa palès, aquest és un objecte derivat. Denotarem l'esquema de la classe que els defineix de la manera següent:

[6] Les dades resultants dels resums d'aquesta subsecció es troben al fitxer `users.csv`

$$\texttt{user(user\_id, n\_sess, sum\_a\_du, m\_a\_du, m\_wday, m\_hour, f\_sess, l\_sess)}$$

Per a cada valor distint i-èssim de l'atribut `user_id` de `session`, construirem doncs un objecte $\texttt{user}_i$ els valors dels atributs del qual calcularem com s'indica a continuació. Sigui

$$\texttt{sessions}_i := \sigma_{(\texttt{user\_id}=\texttt{user\_id}_i)}(sessions)$$

Aleshores hi definim els atributs següents:

(1) $\texttt{id}_i := \texttt{user\_id}_i \in ([\texttt{"0"},\texttt{"9"}] \cup \{\text{“”}\})13$

(2) $\texttt{n\_sess}_i := |\texttt{sessions}_i| \in \mathbb{N}$

(3) $\texttt{sum\_a\_du}_i := \sum_{v \in \texttt{sessions}_i} l(v.\texttt{activity\_duration}) \in \mathbb{N} \times \mathrm{s}$

(4) $\texttt{m\_a\_du}_i := \dfrac{\texttt{sum\_a\_du}_i}{\texttt{n\_sess}_i} \in \mathbb{R} \times \mathrm{s}$

(5) $\texttt{m\_wday}_i := \dfrac{1}{\texttt{n\_sess}_i} \sum_{v \in \texttt{sessions}_i} \texttt{wday}(v.\texttt{central\_activity\_point}) \in [0,6] \subseteq \mathbb{R}$

(6) $\texttt{m\_hour}_i := \dfrac{1}{\texttt{n\_sess}_i} \sum_{v \in \texttt{sessions}_i} \texttt{hour}(v.\texttt{central\_activity\_point}) \in [0,23] \subseteq \mathbb{N}$

(7) $\texttt{f\_sess}_i := \min(\{v.\texttt{central\_activity\_point} : v \in \texttt{sessions}_i\}) \in \texttt{timestamp}$

(8) $\texttt{l\_sess}_i := \max(\{v.\texttt{central\_activity\_point} : v \in \texttt{sessions}_i\}) \in \texttt{timestamp}$

Com es pot observar, de moment ens hem centrat fonamentalment en dos aspectes de la relació entre els usuaris i les sessions que han establert.

(1) En les mitjanes mostrals dels valors dels atributs de les sessions que corresponen a cada usuari.

(2) En la durada de l'activitat de cada usuari, delimitada pels valors `central_activity_point` corresponents a les seves primera i darrera sessions.

Anomenarem `users` la relació que conté tots els casos de `user` obtinguts, la qual, recordem, presenta la següent cardinalitat:

$$|\texttt{users}| = 75609$$

Distribució dels valors de les mitjanes mostrals

Una simple inspecció de certs percentils dels valors que prenen les mitjanes mostrals que dónen valor a 5 atributs d'aquest nou objecte ens permet copsar ràpidament que en aquest cas també presenten una disparitat molt acusada. En tenim una mostra en la taula 3.4.

100-quantil	$\lfloor k/100 \cdot \lvert\texttt{users}\rvert \rfloor$	n_sessions	sum_a_du	m_a_du	m_wday	m_hour
1	756	1	28	18.5	0.00	6.0
5	3780	1	185	90.6	1.00	10.5
10	7560	2	488	167.4	1.75	11.9
25	18902	7	3674	383.6	2.24	13.5
50	37804	68	47500	749.0	2.64	14.8
75	56706	167	161972	1248.9	3.00	16.1
90	68048	296	341317	1814.9	3.50	17.6
95	71828	399	515577	2252.9	4.00	19.0
99	74852	670	1087294	3785.7	6.00	22.0

Taula 3.4: Percentils rellevants dels valors numèrics de user.

De fet, en n_sessions, sum_a_du i m_a_du, hi ha tants valors petits i tenen valors molt alts un nombre tan petit de casos que moltes de les representacions gràfiques que en poguem fer són de poca utilitat. Llurs distribucions presenten un rang interquartílic tan petit en relació als valors més elevats que en un diagrama de caixes aquestes se'n apareixerien com un segment horitzonal. Per altra banda, una gràfica de l'estimació de la PDF d'alguns d'ells tindria un aspecte encara més similar a una "L" del que es pot apreciar en la corba de la gràfica de la figura3.1. És per aquesta raó que representarem aquesta distribució mitjancant corbes d'estimació de la densitat de probabilitat en gràfiques l'escala de les abscisses de les quals sigui logarítmica en base 10.

En primer lloc, en la figura 3.14 hi podem trobar una representació d'aquesta mena pel que fa els valors de n_sess. Podem observar-hi com el gruix dels usuaris del CV han establert un nombre de sessions que ronda les 100.

n_sess $\leq k$	n_k	f_k
1	6880	9 %
2	10978	13.5 %
3	13704	18.1 %
4	15542	20.5 %
5	17030	22.5 %

Taula 3.5: Distribució dels valors més petits de n_sess. Com es pot observar, la freqüència acumulada (f_k) s'expressa en percentatges.

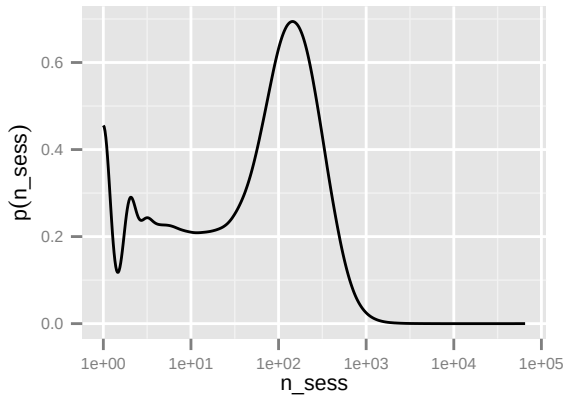


Figura 3.14: Estimació de la PDF dels valors de l'atribut `n_sess`. L'escala de les abscisses és logarítmica en base 10.

El fenomen que sí que mereix especial atenció és, no obstant, l'acumulació de valors molt a prop de l'1. La taula 3.5 mostra el nombre d'observacions i la freqüència corresponents als cinc valors més petits de `n_sess`. És rellevant constatar com el 9% de tots els usuaris només han establert una sessió amb el CV, així com quasi una quarta part només n'han establert 5 o menys. Per altra banda, si bé el percentil 99è d'aquest atribut es situa a 670 sessions, a partir d'aquest punt els valors es disparen molt ràpidament. En la taula 3.6 es mostren els 5 valors més grans que pren.

i	x_i
$\vdots$	$\vdots$
$n-4$	3346
$n-3$	3698
$n-2$	3730
$n-1$	5408
n	65160

Taula 3.6: Els 5 valors més grans de `n_sessions`.

La distribució dels valors també presenta una gran disparitat si ens atenem a l'atribut `sum_a_dur`, això és, a la suma total dels segons en què l'usuari ha estat actiu. La gràfica de la figura 3.14 mostra la corba de l'estimació de la PDF per als valors d'aquest atribut, també amb una escala logarítmica en l'eix de les abscisses.

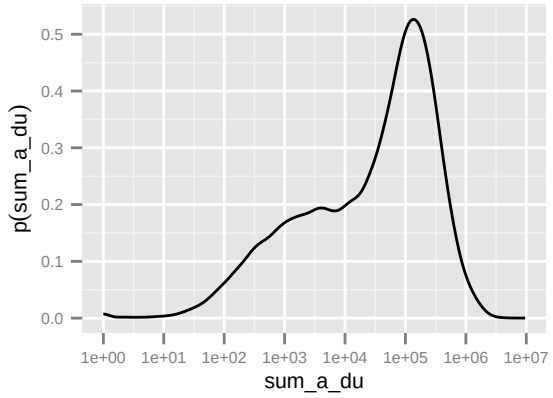


Figura 3.15: Estimació de la PDF dels valors de l'atribut `sum_a_du`. L'escala de les abscisses és logarítmica en base 10.

A primer cop d'ull, la dispersió en la franja alta els valors no sembla tan acusada com en el cas anterior. Observem que el gruix dels usuaris han presentat, en segons, una durada total de l'activitat en l'interval $[10^2, 10^{5.5}]$[7], amb un màxim al voltant de 10^5 (2.7 h). Ara bé, aquesta aparença és causada en efecte per la transformació

[7] Noti's que aquest interval és, igualment, molt gran.

logarítmica que hem introduït en l'escala de la variable independent. En la gràfica de la figura 3.16, en què l'escala és lineal, amb prou feines es pot apreciar a simple vista l'àrea compresa sota la línia que hi apareix.

Val a dir que en aquest cas la causa de la gran disparitat observable recau en major mesura en els valors extrems, que són els que provoquen que la cua de la dreta sigui tan llarga. La informació present en les taules 3.7 i 3.8 dona fe d'aquest fet.

Això no obstant, la conclusió útil que podem extreure dels trets del comportament d'aquestes dues variables és quelcom prou concret. Les distribucion respectives dels valors d'aquests dos atributs són una raó suficient per a no descartar la hipòtesi segons la qual la proporció d'usuaris que no corresponen a estudiants matriculats en un curs ordinari de la UOC és considerable. Observem sinó el següent:

(1) És raonable pensar que un estudiant es connectarà 5 vegades o més al CV pel cap baix. Cal tenir present que, per a seguir satisfactòriament el curs com a mínim a de realitzar tasques com llegir el pla docent a l'inici del semestre, presentar tres o més PACs, consultar els horaris dels exàmens, etc.

(2) També és raonable considerar que la suma del temps necessari per a dur a terme les accions que s'enumeren en el punt anterior és com a mínim d'una hora.

Noti's que hi ha gairebé 1/4 dels usuaris que no satisfan les dues condicions que s'acaben d'apuntar. Ara bé, no gosaré intentar estimar quina és la proporció d'usuaris que corresponen a estudiants, ja que entenc que, amb les dades de què disposem, aquesta tasca no és factible dins d'uns intervals de confiança raonables. En el capítol següent (4) donaré més raons per a defensar aquesta tesi. Limitem-nos de moment a corroborar la nostra sospita sobre la diversitat d'usuaris que s'amaquen rere els identificadors de l'atribut `session.user_id`.

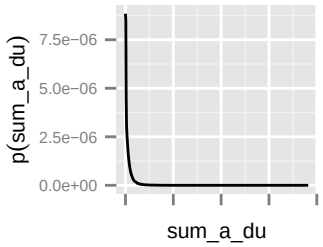


Figura 3.16: Estimació de la PDF per a `sum_activity_duration`.

`sum_a_du` $\leq k$	n_k	f_k
0	234	0.3 %
1	374	0.5 %
30	796	1 %
60	1362	1.8 %
600	8506	11.3 %
1800	14558	19.3 %
3600	18763	24.8 %

Taula 3.7: Distribució dels certs valors de `sum_a_du`. Com es pot observar, la freqüència acumulada (f) s'expressa en percentatges.

i	x_i
⋮	⋮
$n-4$	3346
$n-3$	3698
$n-2$	3730
$n-1$	5408
n	65160

Taula 3.8: Els 5 valors més grans de `sum_a_du`.

Relació entre la primera i darrera sessions

Tancarem aquesta secció tot posant la nostra atenció en la informació que podem obtenir a partir de l'estudi de les relacions entre els dos punts temporals en què tenen lloc la primera i la darrera sessions de cada usuari (atributs `f_sess` i `l_sess`).

En primer lloc, notem que l'interval temporal comprès entre `f_sess` i `l_sess` és una bona representació del període en què l'usuari ha estat actiu pel que fa la seva relació amb el CV de la UOC. Hi ha, com a mínim, dues dimensions interessants d'aquest interval:

(1) La *mida* de l'interval, és a dir:

$$l([\texttt{f\_sess}, \texttt{l\_sess}])$$

(2) La *seva posició* en el desenvolupament temporal del semestre. És a dir, allà on caigui el valor central del periode d'activitat, que

podem definir

$$(\texttt{f\_sess} + l([\texttt{f\_sess}, \texttt{l\_sess}]))/2$$

Una bona manera de mostrar la distribució dels usuaris en funció d'aquests dos paràmetres és l'elaboració d'un diagrama de dispersió com el que es mostra en la figura 3.17. Noti's que cada punt representa un usuari, en què la seva posició en l'eix de les abscisses correspon al valor de `l_sess` i la del de les ordenades al de `f_sess`.

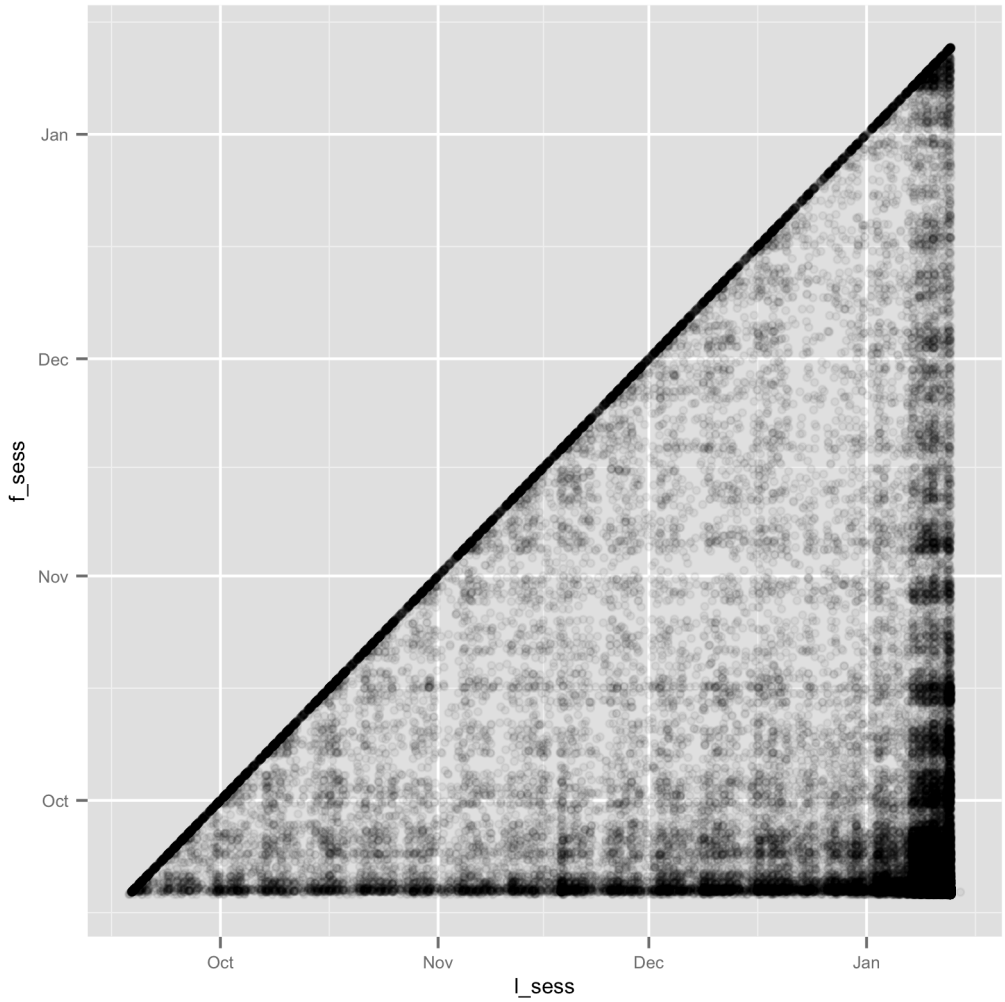


Figura 3.17: Diagrama de dispersió que relaciona els valors de `f_sess` i `l_sess`

Aquesta gràfica proporciona molta informació útil per als nostres objectius. En primer lloc, notis que la diagonal *perpedicular* a la recta $y = x$ és una bona magnitud per a representar la mida de l'interval d'activitat de cada usuari. És a dir, com més lluny es trobi un punt de la recta $y = x$, més gran és la distància temporal que hi ha entre les seves primera i darrera sessions.

Per que fa la distribució dels casos, observem el següent:

(1) La gran acumulació de casos en la cantonada inferior dreta de la gràfica correspon a aquells usuaris que començaren la seva

activitat al CV d'hora, en el desenvolupament del semestre i l'acabaren tard. És raonable incloure aquí que tots aquells usuaris que són estudiants i que han seguit un curs ordinari en la seva totalitat. Tot i això, és important tenir present que aquesta premissa no implica que en siguin la majoria.

(2) La zona fosca en forma triangular que trobem en la cantonada superior dreta està composta per tots aquells usuaris que es connectaren tard per primera vegada. Sembla que el seu pes no és menyspreable.

(3) La gran concentració de casos que ressegueixen la recta $y = x$ correspon al considerable grup d'usuaris als quals només els correspon una sessió o que, si més no, presenten un interval d'activitat semestral extremadament curt.

(4) La franja vertical de punts que es dibuixa a cavall del primer i segon quarts del més de gener en l'eix de les abscisses correspon a aquells usuaris que van connectar-se per primera vegada passat el primer d'octubre, però el període d'activitat del qual abasta fins després de les vacances de Nadal. Aquesta columna presenta una densitat decreixent en sentit vertical, cosa que és completament coherent amb el fet que per a aquells estudiants que encara no han començat a dedicar-se al curs al qual s'han matriculat, com més temps passa, més difícil és que puguin superar-lo i, per tant, menys probable és que ho facin. No obstant això, cal no oblidar en cap moment que això tampoc no indica res pel que fa la proporció d'usuaris inclosos en aquesta franja que corresponen a estudiants. Per altra banda, notis que s'hi dóna un patro d'alternança. Això encaixa completament amb les davallades d'activitat que s'observen durant el cap de setmana, tal com es pot veure en els resultats que s'exposen en la secció anterior.

(5) En darrer terme, observi's que també es pot apreciar una franja horitzontal, en aquest cas més estreta, a l'alçada de la primera setmana del curs de l'eix de les ordenades. Noti's que presenta una gradació creixent en la seva densitat. No és forassenyat suposar que aquesta correspon a un altre comportament que a què apunta el sentit comú pel que fa els estudiants d'un curs ordinari, a saber, que a mesura que passa el temps n'augmenta el nombre que l'abandona.

Aquests resultats són els primers que ens poden fer pensar que hi ha la possibilitat de descobrir tipologies d'usuaris per mitjà de mètodes d'agregació (*clustering*) que siguin útils per als nostres propòsits, així com també en la utilitat de l'aplicació d'algoritmes d'aprenentatge computacional supervisat per tal d'obtenir alguna classificació reeixida i, si s'escau, predir-ne certes característiques. Fins a quin grau és possible fer això i quines serien les transformacions addicionals de les dades necessàries per a realitzar aquestes tasques és el que s'exposa en el capítol següent.

4
Consideració sobre la densitat semàntica de les variables

Després d'haver fet una primera ullada a les característiques dels valors d'alguns dels atributs dels objectes bàsics que s'han definit en la secció 1.2, és un bon moment ara per a mirar d'escatir fins a quin punt aquests poden esdevenir una font de coneixement que ens sigui útil per la consecució dels nostres objectius.

4.1 Utilitat dels atributs emprats

Tinguem present que, com s'ha enunciat en la secció 1.2, la finalitat, en aquest estudi, de l'aplicació del mètodes propis del KDD i l'ML en les dades rebudes ha d'ésser la de proporcionar-nos un coneixement que sigui útil per a, en darrera instància, millorar els processos d'aprenentatge en general i, en concret, aquells que es donen en el marc del CV de la UOC. Així doncs, no és sobrer preguntar-se fins a quin punt poden ajudar-nos en aquesta direcció els valors dels atributs dels objectes construïts a partir de l'abstracció elemental que s'ha fet fins ara. Com em sembla haver provat en la segona part de la secció que ara s'enceta, la resposta a aquesta qüestió és que, sense cap transformació ulterior, l'ajuda que ens brinden és virtualment nul·la.

Això no obstant, com a part del desenvolupament del meu argument, caldrà primer aturar-se en la descripció d'una carcaterística de les dades rebudes que encara no s'ha esmentat i que suposaria un obstacle addicional per l'assoliment dels nostres objectius si no introduíssim les transformacions que proposo a continuació.

Solapament de sessions

Un estudi una mica més exhaustiu de les dades que s'han rebut ens permet descobrir que hi ha sessions distintes corresponents a un mateix `user_id` els respectius intervals `activity_duration` de les quals es solapen. Resseguim l'operació en àlgebra relacional [**?**] que ens permet obtenir aquest resultat. En primer lloc, definim, per a cada usuari `user_id` $\in$ `users` les relacions següents:

$$\texttt{sessions}_i := \pi_{(\texttt{id,session\_start,last\_request})}(\sigma_{(\texttt{user\_id}=\texttt{id}_i)}(\texttt{sessions}))$$
$$\texttt{sessions\_0}_i := \rho_{(\texttt{id\_0/id,session\_start\_0/session\_start,last\_request\_0/last\_request})}(\texttt{user\_sessions}_i)$$
$$\texttt{sessions\_1}_i := \rho_{(\texttt{id\_1/id\_0,session\_start\_1/session\_start\_0,last\_request\_1/last\_request\_0})}(\texttt{user\_sessions\_0}_i)$$

Emprant els noms dels atributs de les dues darreres noves relacions, que són la mateixa duplicada, definim ara la condició c de la manera següent:

$$c := (\texttt{id\_0} < \texttt{id\_1}) \wedge (\texttt{session\_start\_0} \leq \texttt{last\_request\_1}) \wedge (\texttt{session\_start\_1} \leq \texttt{last\_request\_0})$$

Noti's que els dos darrers termes d'aquesta conjunció conformen la condició necessària per a poder afirmar que els dos intervals

$$[\texttt{session\_start\_0,last\_request\_0}]$$
$$[\texttt{session\_start\_1,last\_request\_1}]$$

es solapen. Per altra banda, el primer, $(\texttt{id\_0} < \texttt{id\_1})$, té la finalitat d'evitar que en la relació resultant de la $(<)$-combinació que s'indica a continuació es compti un mateix solapament més d'un cop o s'hi inclogui el d'una sessió amb ella mateixa.

En darrer lloc, obtenim la relació `overlappings` tot definint-la així:

$$\texttt{overlappings} := \bigcup_i (\texttt{user\_sessions\_0}_i \bowtie_c \texttt{user\_sessions\_1}_i)$$

És a dir, mitjançant la unió de totes les relacions que s'han obtingut per a cada un dels usuaris. Aquesta relació conté tots els solapaments entre sessions presents en les dades originals. La seva cardinalitat és

$$|\texttt{overlappings}| = 110838$$

i el nombre de sessions implicades és

$$|\{x : x = \texttt{overlappings.id\_0}\} \cup \{y : y = \texttt{overlappings.id\_1}\}| = 179332$$

això és, un 2.02%.

Per consegüent, la imatge que obtenim és que, per a cada $\texttt{user\_id}_i$, si bé el conjunt de valors de cada un dels atributs de $\texttt{sessions}_i$, pres per si sol, és ordenat, no podem considerar les sessions d'un usuari com una seqüència pròpiament dita, sinó que més aviat correspon a un esquema com el que es mostra en la figura 4.1.

Una vegada identificades les sessions que es solapen pot ser interessant realitzar el seu aplanament. És a dir, per a cada instant t, si hi havia dues sessions v_0, v_1 tals que

$$t \in v_0.\texttt{activity\_duration} \wedge t \in v_1.\texttt{activity\_duration}$$

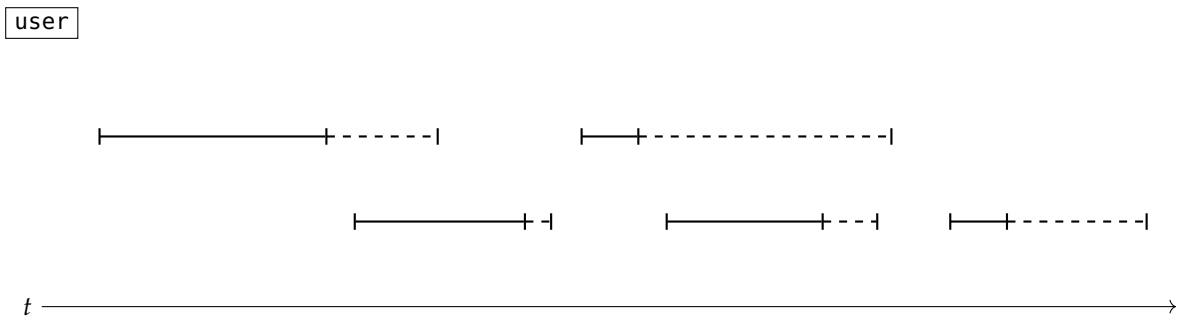


Figura 4.1: Esquema que il·lustra les relacions temporals entre les sessions que pertanyen a un mateix `user_id`.

aleshores les dues sessions es fusionen en una de sola.

Aquesta operació d'aplanament dóna lloc a una nova relació amb 8776818 instàncies, només un 0.6% més petita. Ara bé, encara que llur nombre no es redueixi sensiblement, això no significa que el canvi que aquest nou conjunt de dades introduiria en els estadístics que hem emprat per a fer els primers estudis sigui negligible, especialment si els volem prendre com a base per a les operacions d'agregació i classificació. Tanmateix, com exposo a continuació, no ens caldrà refer-los en aquesta direcció, ja que podem extraure informació de les dades que és més valuosa que aquella que hem emprat fins ara sense fer-ho.

Significat dels atributs pel que fa el procés d'aprenentatge

Fins ara hem centrat la nostra atenció fonamentalment en la relació mútua del 3 primers dels 4 atributs presents en cada una de les entrades del fitxer CSV rebut. Per una banda, ens hem fixat en la mida i la posició, en el desenvolupament temporal del semestre, de l'interval temporal `activity_duration`. Per l'altra, hem mirat de donar una caracterització de cada usuari a partir de la descripció de la distribució dels valors d'aquests dos atributs en els grups de sessions que li corresponen. La mida d'aquest interval d'activitat l'hem representada per mitjà del valor, en segons, de $l($`activity_duration`$)$, i la seva posició temporal per mitjà de `central_activity_point`.

Malgrat tot, cal preguntar-se fins a quin punt aquestes dues variables descriuen prou bé la relació d'un usuari amb el CV en el transcurs d'una sessió. Pel que fa això, notem el següent:

(1) Tant la marca temporal `session_start` com `last_request` ens indiquen poca cosa més que el fet que l'usuari (en cas que es tractés d'un humà) *tenia la seva atenció posada* en el CV en aquests dos moments. Ara bé, com és obvi, això no significa que, en cas que es tracti d'un estudiant, hagi dedicat la totalitat del temps comprès entre aquests dos instants a la realització de tasques relacionades amb el procés d'aprenentatge, ni molt menys. Com a estudiant de la UOC, l'experiència m'indica que no són rars els casos en què hom inicia la sessió al CV, dedica una estona a l'estudi, centra la seva activitat en quelcom d'altre durant un període de longitud considerable i després hi retorna sense que

hagi caducat la sessió. Per altra banda, tampoc són estranyes les situacions en què l'estudiant està en efecte dedicant temps al seguiment del curs de la UOC (per exemple, redactant un treball o estudiant dels apunts) i això no es tradueix en la seva presència com a usuari identificat en el CV. Per posar un altre exemple basat en l'experiència pròpia, en el moment en què estic escrivint aquesta oració ja he dedicat prop de 18 hores a l'elaboració d'aquesta memòria sense haver-me connectat al CV durant un període de 3 dies. Evidentment, no podem pretendre saber que el meu cas particular és prou representatiu del comportament del conjunt dels estudiants de la UOC, però bé hem d'admetre que *les dades de què disposem no ens permeten tampoc afirmar si no ho és*.

(2) Per altra banda, val a dir que el nombre de sessions establertes per part d'un usuari en particular també és un és una representació força pobre pel que fa la seva relació amb el procés d'aprenentatge. Altra vegada, tampoc no és forassenyat suposar que hi ha casos en què un estudiant s'identifica en el CV per mer hàbit encara que no s'hagi de dedicar pròpiament al seguiment del curs de la UOC en aquell precís moment.

(3) El gran nombre de casos amb valors estranyament baixos i les llargues cues que s'extenen pel tram superior de llurs distribucions en aquestes les dues variables analitzades, així com en l'atribut `n_sess`, ens han de fer dubtar sobre la possibilitat d'assumir que el comportament general de les instàncies de `user` és representatiu d'aquells que són estudiants. Per a poder estar en condicions d'emmarcar el nostre estudi en la EDM&LA és imprescindible que poguem obtenir coneixement útil en relació als actors que prenen part en el procés d'aprenentatge, en aquest cas, professors, consultors i preferentment, alumnes. Això sembla difícil d'aconseguir si és que ens hem d'atendre a aquestes variables.

(4) En darrer lloc, la presència de sessions que es solapen ens recorda que tenim un desconeixement absolut de quin és el procés que ha donat lloc a la generació de les dades que conté el fitxer `20131.con.txt`. És ben possible que aquests solapaments es deguin al fet que un mateix usuari s'ha identificat des de dos o més clients diferents (és a dir, dos o més navegadors o aplicacions mòbils diferents, s'executin o no en amfitrions distints). Però també poden haver estat causats per la mena de procés que agrega les dades, que sigui tal que hagi donat lloc a la pèrdua de llur consistència. De totes maneres, sobre això tampoc en sabem res.

La constatació d'aquests quatre fets és ja suficient per a concloure que els valors concrets que presenten els atributs de `sessions` són un indicador molt poc fiable per a extreure conclusions sobre el comportament dels estudiants pel que fa llur relació amb el CV, i encara ho

són menys si hem d'intentar modelar el d'altres actors, com poden ser professors o consultors. Això no obstant, en la secció següent pretenc haver trobat una mètrica que sí que ens proporciona informació útil pel que fa aquest aspecte.

4.2 Obtenció de mètriques més útils

Així doncs, davant la pobresa significativa de les dades que hem rebut, ens hem de veure obligats a renunciar a la missió de mirar d'obtenir uns resultats suficientment valuosos en el marc de la EDM&LA? La resposta és que això només és així si mirem de fonamentar-los en el detall de les variacions dels valors dels atributs dels objectes obtinguts per mitjà d'aquesta primera abstracció bàsica. Afortunadament, aquest no és el cas, i, com es pot veure a continuació, podem construir nous objectes derivats la significació dels valors dels atributs dels quals sigui prou laxa com per poder obtenir resultats prou precisos córrer el risc que siguin greument esbiaixades.

La noció de presència *en el CV*

Una de les conseqüències més rellevants dels resultats obtinguts fruit de l'estudi dels trets dels usuaris, els quals es mostren en la secció 3.3, és el reconeixement de l'especial valor significatiu que tenen *la primera* i *la darrera* sessions de cada usuari, representades per part dels valors dels atributs `f_sess` i `l_sess`, respectivament. Independentment de les limitacions que ens imposen els fets constatats en el capítol 2 aquestes dues variables *sí que ens donen una informació fiable* sobre quin és el periode durant el qual l'usuari manté una presència en el CV, independentment del seu grau d'intensitat. Un bon exemple d'aquest fenomen és la gran quantitat d'informació significativa que hem pogut obtenir tot relacionant els valors d'aquestes dues variables i construint un diagrama de dispersió com el de la figura 3.17.

En concret, podem afirmar que `l_less` marca el moment temporal a partir del qual l'usuari deixa de tenir cap relació amb el CV fins el final del semestre. Si l'usuari és un estudiant, com més d'hora es situï el valor d'aquest atribut, més probable serà que hagi abandonat el curs en la totalitat de les assignatures a què s'havia matriculat. Aquest atribut és doncs un bon candidat a generar una marca de classe en vistes a un procés d'aprenentatge supervisat que ens proporcioni informació útil pel que fa el risc d'abandonament dels usuaris i, per tant, dels estudiants.

Per altra banda, podem prescindir del detall dels valors particulars de $l($`activity_duration`$)$ i `central_activity_point` i preguntarnos solament per la presència o absència de cada usuari concret en el marc dels intervals definits a partir d'un cicle temporal determinat. Tot prenent la relació `users`, podem construir-ne una de nova que presenti l'esquema següent.

Per a cada tupla $\texttt{user}_i$ que hi pertany, n'obtindrem una derivada

tal que

$$\texttt{user}'_i = \langle \texttt{user\_id}_i, a_{i,0}, \ldots, a_{i,m} \rangle$$

on el valor de cada atribut $a_{i,j}$ ($j \in [0, m]$) i llur nombre ($m + 1$) els definirem de la manera següent.

Sigui $d \in \mathbb{N} \times \mathrm{s}$ una possible durada expressada en segons. Prenguem t_0 tal que $\min(\texttt{session\_start}) \in [t_0, t_0 + d]$. Definim ara una discretització que inclogui la durada total del semestre en $m + 1$ intervals consecutius I_j ($j \in [0, m]$) de tal manera que:

$$I_j = [t_0 + dj, t_0 + d(j + 1)]$$

En tercer llocm, el tipus de cada atribut $a_{i,j}$ serà booleà ($\{0, 1\}$) i li assignarem el valor que li correspon de la manera següent:

$$a_{i,j} = \begin{cases} 1 \text{ si } \{v \in \texttt{sessions}_i : (v.\texttt{session\_start} \leq t_0 + d(j+1)) \wedge (t_0 + dj \leq v.\texttt{last\_request})\} \neq \emptyset \\ 0 \text{ altrament} \end{cases}$$

En altres paraules, per a qualsevol usuari i-èssim, l'atribut $a_{i,j}$ valdrà 1 si i només si existeix alguna de les sessions de $\texttt{user}_i$ tal que el seu `activity_interval` es solapi amb l'interval I_j. Altrament valdrà 0. Aquest atribut expressa si l'usuari i ha estat actiu al CV, *per poc que sigui*, durant l'interval I_j. Em prenc la llibertat d'anomenar *presència* (i, recíprocament, *absència*), el tret booleà que mesura cada un d'aquests atributs.

D'aquesta manera, en funció de la mida d dels intervals I_j que escollim, obtindrem relacions derivades diferents amb llurs respectius graus de granularitat a l'hora d'expressar la presència o absència de cada usuari en el CV durant el transcurs del desenvolupament temporal del semestre.

Per als estudis que segueixen he construït dues noves relacions tot assignant dos valors diferents a d:

(1) En primer lloc, `user_days` amb $d = 86\,400\,\mathrm{s} = 1$ dia, de tal manera que $m = 119$ i el seu esquema presenta, per tant, 121 atributs.

(2) Per altra banda, `user_weeks` amb $d = 604\,800\,\mathrm{s} = 1$ setmana, de tal manera que $m = 17$ i el seu esquema presenta, per tant, 19 atributs.

Per cloure aquesta subsecció notem que no he triat cap altre valor de d per les dues raons següents:

(1) No n'he triat un d'inferior a un dia (per exemple, períodes de 6 h, que podrien modelar correctament el cicle [matinada, matí, tarda, vespre]) per la impossibilitat de saber quina és exactament l'hora local del client per a cada una de les sessions.

(2) No n'he triat cap entre el dia i la setmana, o superior a aquest darrer, ja que no es donen altres cicles temporals clars (com a

mínim pel que fa la vida quotiada de la majoria dels humans integrants de societats occidentals) que no depassin el marc d'un semestre (per exemple, l'any).

Reformulació i concreció dels objectius

Noti's que podem interpretar els valors de cada un dels atributs $a_0, \ldots, a_m$ dels elements les noves relacions `user_days` i `user_weeks` com una seqüència binària de longitud $m + 1$ que caracteritza l'evolució de la presència (o absència) al CV de cada un dels usuaris identificats. Essent aquesta mètrica més resistent a la incertesa introduïda per part dels fets que s'enumeren en la secció 4.1, ja estem en condicions, doncs, de formular un objectiu més concret per a la nostra tasca d'EDM&LA.

Els objectius d'aquest estudi es redueixen a la caracterització de l'evolució de la presència dels usuaris en el CV en funció de les dues discretitzacions $d \in \{1 \text{ dia}, 1 \text{ setmana}\}$ per als dos aspectes concrets següents:

(1) En primer lloc, aplicarem mètodes d'agregació per a identificar diferents arquetips pel que fa l'evolució de la presència dels usuaris al CV.

(2) En segon lloc, assajarem la construcció d'un model predictiu pel que fa la probabilitat que un usuari qualsevol hagi *abandonat* el CV en un interval I_j. Entendrem abandó com l'absència definitiva fins a final de semestre.

5
Models d'agregació

Disposem-nos doncs a aplicar mètodes d'agregació amb la intenció d'obtenir una tipologia de l'evolució de la presència dels usuaris al CV.

*5.1 El mètode k-*means *sobre els valors dels atributs* $a_0, \ldots, a_m$

Com és ben sabut, podem interpretar la seqüència de valors booleans dels atributs $a_0, \ldots, a_m$ com la posició de cada instància de `user` en un espai de $m + 1$ dimensions. Aquesta posició, és clar, caracteritza unívocament l'evolució concreta que ha seguit l'usuari en qüestió pel que fa la seva presència en el CV de la UOC. En aquesta secció mostro els resultats de l'aplicació del popular mètode d'agregació (*clustering*) *k-means* [**?**] sobre els valors d'aquests atributs.

Anotem-ne el funcionament a mode de recordatori. Per a un conjunt de casos $R = \{v_1, \ldots, v_n\}$ on cada $v_i = \langle x_{i,1}, \ldots, x_{i,m} \rangle$, el mètode *k-means* cerca una partició de R, $S = \{s_1, \ldots, s_k\} \subset \mathcal{P}(R)$, en k agregacions ($k \leq n$) de tal manera que es minimitzi el valor θ_k resultant de l'expressió següent,

$$\theta_k = \sum_{i=1}^{k} \sum_{r \in s_i} \|r - \mu_i\|^2$$

on μ_i és el punt mitjà o centroide de l'agregació s_i. En altres paraules, θ_k és el valor resultant de la suma de les sumes dels quadrats de les distàncies euclidianes entre cada cas i el centroide de l'agregació al qual ha estat assignat. La primera i més simple versió de l'algoritme que troba aquest mínim és la que rep el nom d' *algoritme de Lloyd* [**?**], i és la que he emprat en aquest estudi[1].

D'altra banda, aquest algoritme no obté per si sol el valor òptim per a k. A l'hora de triar-lo haurem de fonamentar-nos, doncs, en criteris externs. Un dels més simples consisteix en trobar una solució de compromís entre, per una banda, la minimització del nombre k d'agregacions i, per l'altra, la de la suma θ_k dels quadrats de llurs distàncies internes. En efecte, com més petit és k, més sintètica és la tipologia obtinguda, però, per contra, com més gran és θ_k, menys informació aporta la carcateritzacio de cada grup sobre els elements que conté.

[1] L'obtenció dels resultats de l'aplicació d'aquest algoritme s'ha aconseguit tot emprant la implementació que se'n fa en la funció `kmeans` del paquet `stats` del programari R. A part de la indicació d'emprar aquest algoritme, també s'hi ha passat com a paràmetre que s'extengui el màxim d'iteracions a 1000 amb la finalitat de fer créixer la probabilitat d'assolir la convergència.

Disposem tot seguit en una gràfica la relació dels valors d'aquestes dues variables. Això és el que es mostra en les figures 5.1 i 5.2.

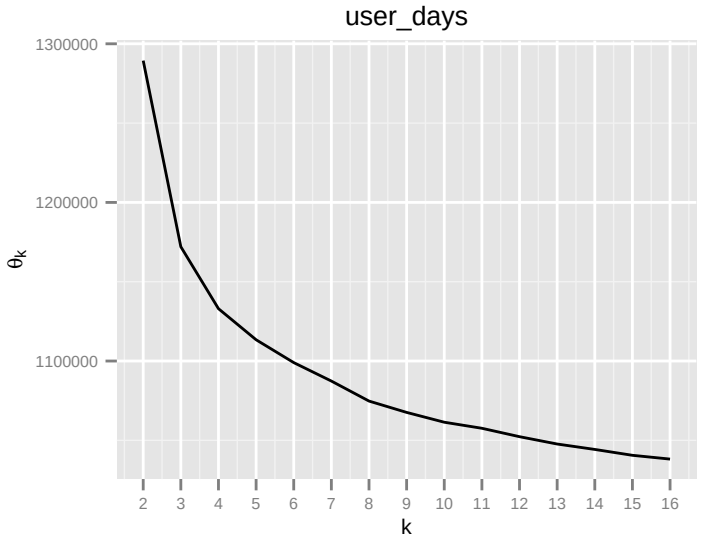


Figura 5.1: Relació entre k i θ_k per al conjunt `user_days`.

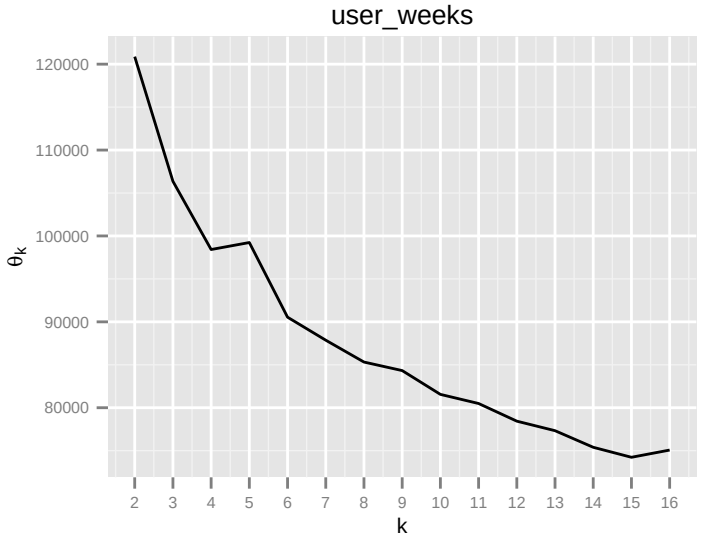


Figura 5.2: Relació entre k i θ_k per al conjunt `user_weeks`.

Observi's que el que cal és trobar un punt a partir del qual el valor absolut del pendent de la línia canviï tan poc, per a cada increment de k en proporció a la mesura en què ho fa en cada pas anterior, que ja no pagui la pena afegir noves classes a S. Com es pot apreciar, a primer cop d'ull sembla que els valors que són millors candidats per ésser assignats a k són 4, 8 i 10 per a `user_days` i 4, 6, 8 i 10 per a `user_weeks`.

Això no obstant, cal tenir molt present que θ_k és tan sols una variància i, per consegüent, consisteix solament en un resum de la dis-

paritat interna entre els casos assignats a cada s_i, però no ens diu res pel que fa l'estructura de la relació entre els valors de llurs atributs. Una eina que pot proporcionar-nos una ajuda addicional a la mera suma dels quadrats de les distàncies internes és sotmetre els centroides $\mu_1, \ldots, \mu_k$ a un procés d'agregació jeràrquica (*hierarchical clustering*) [**?** , p. 520-28]. D'aquesta manera, veurem en quina mesura s'assemblen els centres de cada *cluster* i hi podrem identificar redundàncies, si s'escau, així com reconèixer-hi supertipus que les agrupin. En aquest estudi, l'algoritme triat per a l'agregació jeràrquica ha estat també un de molt comú, a saber, l'*algoritme d'enllaç complet* o *del veí més llunyà*[2].

[2] La implementació d'aquest algoritme que s'ha emprat és la que proporciona la funció `hclust` del programari R

Com és ben sabut, tot algoritme d'agregació jeràrquica comença tot construint una agrupació diferent per a cada instància. Tot seguit, en cada una de les seves iteracions fusionarà les dues agrupacions s_i i s_j que estiguin més a prop segons la funció distància $D(s_i, s_j)$ que calgui aplicar, fins a obtenir un sol *cluster* general. Com indica el seu nom, en el cas de l'enllaç complet D es defineix:

$$D(s_i, s_j) = \max_{v_n \in s_i, v_m \in s_j} d(v_n, v_m)$$

On, en aquest cas, $d(v_n, v_m)$ és la distància euclidiana:

$$d(v_n, v_m) = \|v_n - v_m\|$$

En les subseccions que hi ha a continuació veurem alguns exemples de la seva aplicació.

Ara bé, tot i l'ajuda que ens proporcioni aquesta segona operació, val a dir que si no volem passar per alt trets interessants de les agrupacions obtingudes, no podrem eludir el procés d'examinar, una per una, les representacions concretes de les caracteritzacions de llurs centroides per a cada valor possible de $k \in [2, 16]$. Podem sistematitzar el procés a seguir per a executar aquesta tasca tot estipulant que, per a cada partició S_k cal dur a terme els dos passos següents:

(1) Examinar les característiques dels centroides de cada una de les agregacions per a comprovar si n'hi ha dues o més que siguin molt similars i, a la vegada, semblants també respecte alguna de la partició S_{k-1}. En aquest cas tenim raons per a triar $k - 1$.

(2) De manera recíproca, examinar les característiques dels centroides de la partició S_{k+1} i constatar si és el cas que hi apareix alguna agrupació s nova tal que no n'hi ha cap a S_k que se li assembli. Òbviament, si això és així, tindrem raons per a decantar-nos per S_{k+1}.

És tot trobant un punt de compromís entre aquestes dues tendències que obtindrem un valor òptim de k.

Passem ara a exposar els resultats concrets obtinguts per a les dues relacions `user_days` i `user_weeks`.

Resultats per a `user_days`

Pel que fa a `user_days`, el nombre d'agrupacions més adient ha estat $k = 8$. En la figura 5.5 es mostra la representació gràfica dels centroides μ_i ($i \in [1,8]$) corresponents a cada una de les agrupacions que hem obtingut. Noti's que per a cada una d'elles tenim una gràfica de línia que fa correspondre, a cada dia j-èssim del semestre, la mitjana aritmètica dels valors que pren l'atribut a_j en els casos que hi pertanyen. Addicionalment, s'hi ha superposat la corba de l'estimació LOESS per a fer palesa la tendència general dels valors tot obviant les oscilacions provocades pel part del cicle setmanal.

El repartiment dels casos entre els *clusters* és representat a l'histograma de la figura 5.3. Observi-s'hi com la gran majoria dels casos han estat assignats a l'agrupació 2, que correspon a la d'aquells usuaris que han establert un nombre molt baix de sessions al CV.

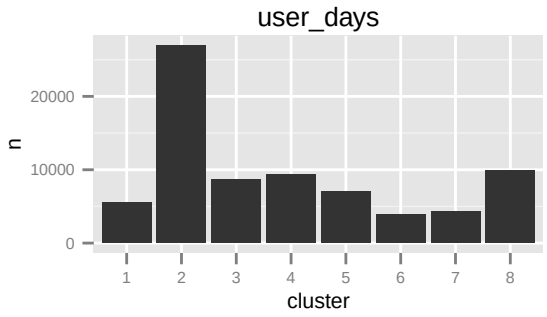


Figura 5.3: Proporció d'usuaris assignats a cada agregació resultant del 8-*means* aplicat a `user_days`.

En tercer lloc, en la figura 5.4 podem apreciar el dendrograma construït a partir de l'aplicació de l'agregació jeràrquica sobre els 8 centroides obtinguts. S'hi dibuixa una divisió clara entre dos grans grups d'usuaris de mida similar $A = \{1,2,5\}$ (47.8%) i $B = \{3,4,6,7,8\}$ (52.2%).

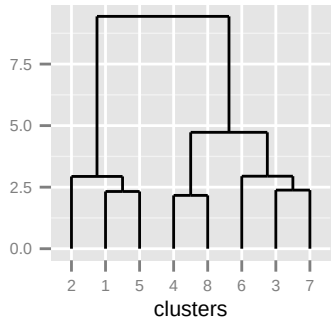


Figura 5.4: Dendrograma per al mètode 8-*means* en `user_days`

No és massa arriscat afirmar que aquesta jerarquia de particions ens subministra una tipologia d'usuaris del CV, com a mínim pel que fa l'evolució de llurs presències o absències en funció del dia del semestre. Intentem ara d'elaborar un esbós de caracterització dels casos que pertanyen a cada un dels *clusters*. Aquesta es mostra a continuació tot seguint la jerarquia que s'expressa en el dendrograma de la figura 5.4. Com es pot suposar, els noms he posat als usuaris que pertanyen a cada un dels *clusters* són solament una proposta, resultat d'aquesta primera aproximació. No tinc la intenció, de cap manera, que aquesta nomenclatura esdevingui definitiva.

A. **Absents** $\{1,2,5\}$ (52.2%)

   Aquesta és la primera de les dues agregacions principals i la que conté més usuaris. Com es pot apreciar, consisteix en el grup d'aquells usuaris que, de mitjana, no són presents al CV de la UOC. Aquesta absència, no obstant, presenta les següents varietats.

   a. **Esporàdics** $\{2\}$

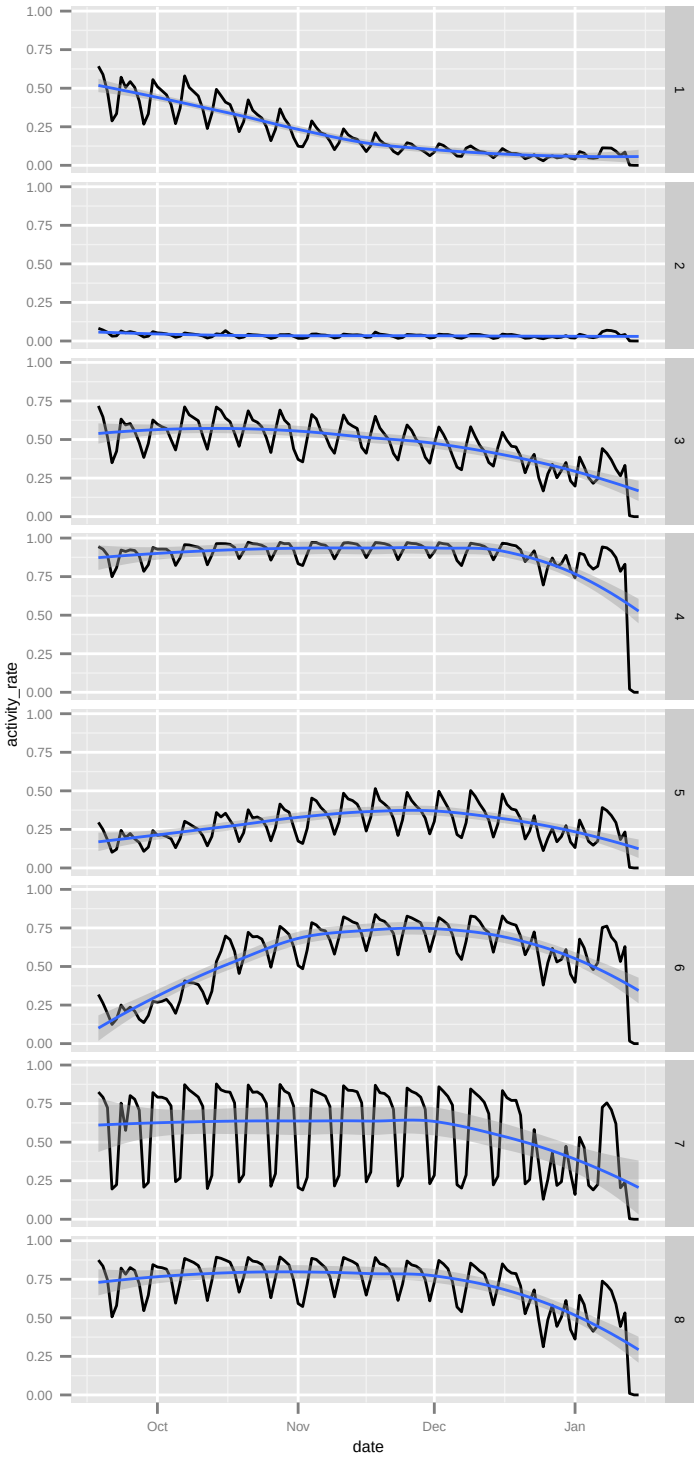


Figura 5.5: Caracterització dels centroides corresponents a les agregacions construïdes per mitjà del mètode 8-*means* en `user_days`.

En primer lloc, el *cluster* que aplega, de llarg, més usuaris (35.6%). Aquí hi trobem aquell gran grup d'usuaris la presència dels quals al CV és merament nominal. De fet, la mitjana aritmètica del valor `n_sessions` dels que hi pertanyen és 6.2.

b. **Capbussadors** $\{1\}$

Amb aquest nom he batejat aquells usuaris que comencen el curs amb una presència prou regular però que ja a partir del primer de novembre presenta una tendència negativa i acaba pràcticament anul·lant-se. Noti's que és probable que aquest comportament correspongui a aquells usuaris que són estudiants matriculats a algun curs, però que n'abandonen el seguiment a mitjans de semestre.

c. **Tímids** $\{5\}$

En aquesta tercera agrupació hi trobem els usuaris l'evolució de la presència dels quals presenta un perfil que encaixa amb el que podriem esperar d'aquells que estan seguint el curs satisfactòriament, però en què, no obstant, el valor dels índexs és sensiblement menor a l'esperat. És a dir, tot i que hi identifiquem els tres trams que ja han aparegut en la gràfica general present en la figura 3.12, la *intensitat* de la seva presència és inferior a la mitjana.

B. **Presents** $\{3, 4, 6, 7, 8\}$ (47.8%)

Aquesta és la segona de les dues agregacions principals. El tret bàsic de l'evolució presencial dels usuaris que hi pertanyen és que aquesta es situa per sobre la mitjana.

a. **Endarrerits** $\{6\}$

Amb aquesta etiqueta em refereixo a aquells que han acabat el semestre haven establert una presència sòlida al CV, però que malgrat tot l'havien començat sense ser-hi gaire sovint. L'experiència ens diu que aquest patró de comportament també és coherent amb el que podríem esperar d'usuaris que corresponen a estudiants.

b. **Fatigats** $\{3\}$

Els patrons de comportament dels usuaris que pertanyen a aquest *cluster* presentarien, en general, una recta de regressió lineal de pendent negatiu pel que fa els valors de llurs respectius atributs $a_1, \ldots, a_{119}$. És a dir, tot i que exhibeixen una mitjana de presència superior a 0.5, aquesta segueix una evolució lleugerament descendent.

c. **Setmanaris** $\{7\}$

Amb aquest nom de fortuna dubtosa em refereixo a aquells usuaris que, tot i tenir una presència consistent al CV durant la totalitat del semestre, aquesta es veu fortament marcada pel cicle temporal setmanal, més concretament, per la dicotomia entre dies laborals i de cap de setmana. Es podria al·legar que aquesta agrupació hauria de subsumir-se a la següent (els

persistents), ja que els usuaris que hi pertanyen són clarament presents al CV d'una manera notablement regular fins a l'inici de les festes de Nadal. Això no obstant, els valors que arriba assolir la corba de regressió polinòmica local (LOESS) són sensiblement menors, i això m'ha semblat una raó suficient per a no fer-ho.

d. **Persistents** $\{3, 7\}$

   En darrer lloc tenim els dos *clusters* que inclouen conjuntament aquells usuaris que, de mitjana, han estat presents al CV de la UOC de manera sostinguda durant tot el desplegament temporal del semestre. És aquest tret principal el que m'ha portat a batejar-los amb aquest nom. La divisió que ha introduït entre ells el procés d'agregació 8-*means* es fonamenta solament en el valor numèric mitjà dels seus atributs. D'entre els usuaris que pertanyen a aquestes dues agregacions, els de la 8 presenten valors inferiors als de la 4. Els he batejat així:

   i. **Dedicats** $\{8\}$

   ii. **Obsessionats** $\{4\}$

Resultats per a `user_weeks`

L'agregació els resultats de la qual es mostren en la subsecció anterior, feta a partir d'atributs que corresponen a la presència o absència dels usuaris al CV en funció del dia del semestre té en compte, òbviament, el patró de comportament dels usuaris en el transcurs de la setmana com a cicle temporal. Per contra, el fet de prendre $d = 1$ setmana, com es mostra en la que s'acaba d'iniciar, es limita a l'evolució de la tendència d'aquesta mètrica durant el transcurs del semestre sense tenir-la en compte.

Sembla que en aquest cas el nombre adequat d'agrupacions és $k = 6$. En la figura 5.7 es mostren les gràfiques corresponents als centroides de cada una de les 6 agregacions obtingudes per a aquest nou valor de d.

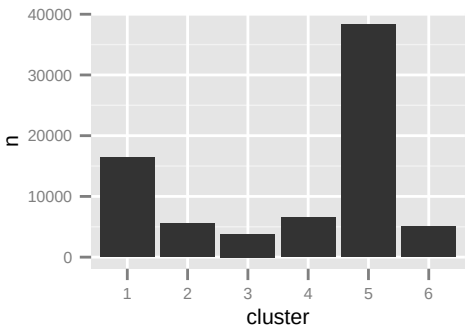


Figura 5.6: Proporció d'usuaris assignats a cada agregació resultant del 6-*means* en `user_weeks`

La proporció del nombre de casos assignats és representada en l'histograma de la figura 5.6. Com s'hi fa evident, en aquest cas la distribució d'usuaris s'acumula al *cluster* 5 (50.6%), que no correspon

a aquelles amb una presència en el CV generalment baixa, com en el cas anterior, sinó justament al contrari.

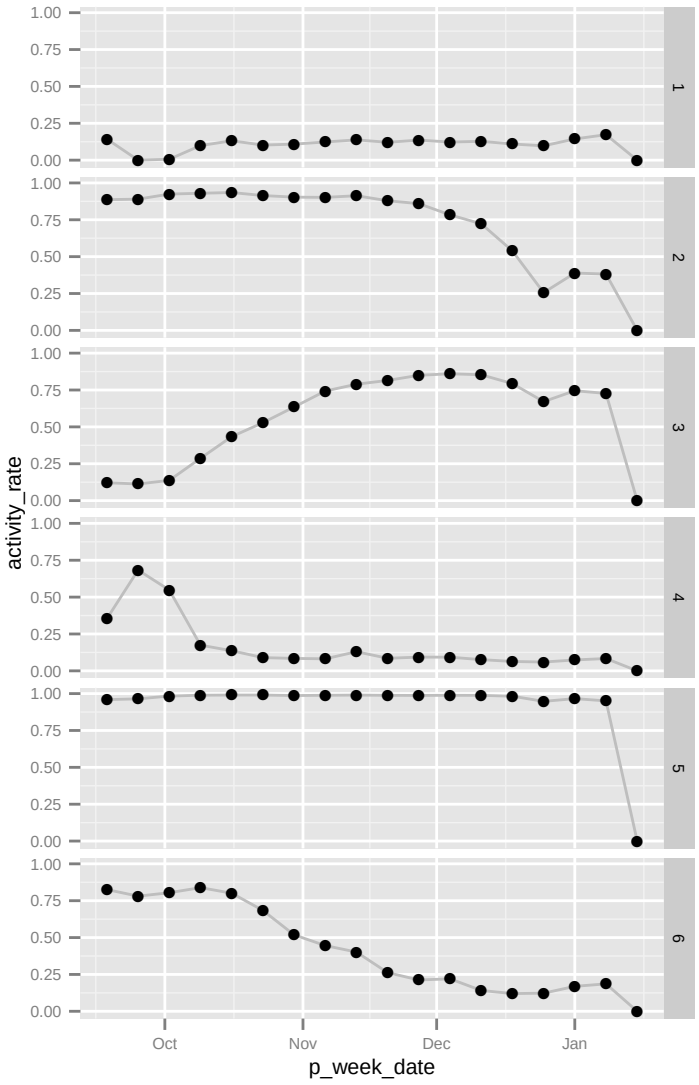


Figura 5.7: Proporció d'usuaris assignats a cada agregació resultant del 6-*means* en `week_days`

Per altra banda, el dendrograma que representa la proximitat relativa entre cada un dels centroides μ_i ($i \in [1,6]$) és el que es pot apreciar en la figura 5.7. També en aquest cas l'agregació jeràrquica també ha produït, com era d'esperar, una primera divisió clara entre aquells usuaris que de mitjana han estat presents al CV durant tot el semestre i aquells que no. Això no obstant, aquesta primera divisió no ha estat tant equilibrada com en el cas anterior i, hem obtingut dues agrupacions $A = \{1,4,6\}$ (37%) i $B = \{2,3,5\}$ (63%). A més, com bé ens indica el dendrograma obtingut, la proporció que s'estableix entre, per una banda, la diferència existent entre els casos pertanyents respectivament a cada una de les dues agrupacions de la divisió base i, per l'altra, aquella que es dóna entre les subagrupacions que inclouen és menor que la s'exhibia per al *k-means* efectuat sobre `user_days`[3].

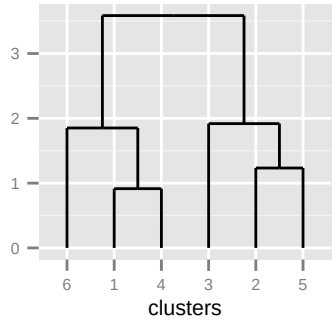


Figura 5.8: Dendrograma per al mètode 6-*means* en `user_week`

[3] Noti's tanmateix que les distàncies entre els centroides dels clusters que es mostren en els dendrogrames corresponents a les agregacions de `user_days` i `user_weeks`, respectivament, no són comparables sense fer cap transformació prèvia. En efecte, la distància màxima possible entre dos punts d'uns d'un hipercub de $m+1$ dimensions, de costat de mida 1, que és la regió en què situa cada cas la seqüència de valors dels atributs $a_0, \ldots, a_m$, és igual a $\sqrt{m+1}$.

Esbossem, com en el cas anterior, una possible tipologia a extraure dels resultats d'aquest procés d'agregació. Com es pot veure, a continuació també es subministra una jeraquia d'agregacions en consonància amb la que es pot apreciar en el dendrograma de la figura 5.8. He intentat emprar una nomenclatura similar a la que he assignat als *clusters* construïts sobre `user_days`. Noti's que només es proporciona una explicació en els casos en què s'ha introduït alguna divergència respecte d'aquells.

A. **Absents** $\{1, 4, 6\}$ (37%)

   a. **Esporàdics** $\{1\}$

   b. **Capbussadors** $\{6\}$

   c. **Desdits** $\{4\}$

      Aquest és un tipus que no apareixia en la jeraquia d'agregacions construïdes a partir de `user_days`. Observi's que consisteix en el grup d'usuaris que, de mitjana, apareixen pel CV durant les tres primeres setmanes del sementre per a no tornar-hi virtualment més durant el temps restant. En aquest sentit, podríem dir que són uns *capbussadors exagerats*. Aquest cluster és relativament petit (6.7%) i, tot i que pot incloure aquells estudiants que van matricular-se al curs ordinari de la UOC però just començar el semestre *van desdir-se'n*, és clar que aquí també volem situar usuaris que, simplement, no són estudiants.

B. **Presents** $\{2, 3, 5\}$ (63%)

   a. **Endarrerits** $\{3\}$

   b. **Fatigats** $\{2\}$

   c. **Persistents** $\{5\}$

      Òbviament, atesa la naturalesa de les dades recollides en `user_weeks` no s'ha pogut reconèixer el tipus d'usuaris *setmanaris* com en el cas anterior. Per altra banda, noti's com aquest *cluster* ha estat ara convertit en una fulla de la jerarquia tot absorvint les dues que subsumia.

5.2 *Utilitat dels resultats de l'agregació sobre* $a_1, \ldots, a_m$

Tot tenint present sempre la limitació imposada per la simplificació que ens proporciona una variable resum com és la mitjana aritmètica, afegida al fet que és especialment sensible als casos extrems, bé hem d'acceptar que l'aplicació del mètode *k-means* als valors de les dues seqüències d'atributs $a_0, \ldots, a_m$ que hem construït proporcionen una primera tipologia d'estudiants en funció de l'evolució llur *presència* en el CV durant el transcurs del semestre. Malgrat tot, si la identificació d'aquests arquetips ens ha proporcionat cap coneixement sobre el domini `users`, val a dir que aquest és completament *a posteriori*, és a dir, s'ha obtingut una vegada s'ha conclòs semestre.

No es pot negar que aquest coneixement pugui ser de certa utilitat per a intentar millorar, en darrera instància, l'eficàcia dels processos que tenen lloc en l'entorn d'aprenentatge del CV de la UOC. Si el comportament general de la població d'usuaris del campus durant el semsetre de tardor varia poc d'any en any, l'educador encarregat de la planificació dels cursos que hi tenen lloc pot emprar un refinament ulterior de resultats com els que s'acaben d'obtenir per a saber *quin tipus d'usuaris pot esperar-se trobar* en ocasions futures, i actuar en conseqüència.

A més, l'addició de noves dades sobre la població estudiada pot introduir molta més llum en la imatge que dibuixa aquesta jerarquia d'agregacions. Cal recordar que, com he notat en la secció 4.1, les dades de què he disposat per a fer aquest estudi són poc significatives en la mesura que detallen molt poc la mena d'informació que proporcionen. Si, per exemple, haguéssim sabut quins valors de `user_id` corresponen a estudiants matriculats, és d'esperar que els resultats d'un estudi de *clustering* com el que s'acaba de dur a terme haurien proporcionat fruits molt més esclaridors. Per altra banda, si a més d'haver pogut refinar les nostres estimacions haguéssim disposat de dades que fossin indicadors més o menys fiables del grau d'èxit de cada un dels estudiants pel que fa la consecució del procés d'aprenentatge (les qualificacions acadèmiques obtingudes, posem per cas), l'estudi de llur correlació amb la pertinença a un o altre *cluster* fins i tot assentaria fonaments per a poder establir directrius d'actuació en la millora dels processos educatius. Però de ben poc serveix a hores d'ara fabular sobre les delícies d'una hipotètica situació més folgada. Centrem l'esforç, per contra, en mirar d'exprémer al màxim la precària significació de les dades de què en realitat disposem.

Per a fer això interntarem ara un assaig de classificació sobre els resultats obtinguts en funció d'alguna atribut que ens permeti predir informació útil per als nostres propòsits.

Possibilitat de prediccions a partir del resultat

Un dels aspectes de la informació que aporten les dades rebudes que, com he dit en la secció 3.3, poden fer servei als objectius particulars de la EDM&LA, és el fet que coneguem en quins moments tingueren lloc tant la primera com la darrera sessió de cada un dels usuaris. En concret, el fet de conèixer-ne la última ens diu quelcom de fins a quin punt l'usuari ha abandonat prematurament la seva relació amb el CV de la UOC i, per tant, sobre el seu abandonament i el consegüent fracàs acadèmic en el cas que es tracti d'un estudiant.

Mirarem ara d'estudiar si hi ha cap relació entre l'abandonament prematur i la pertinència a alguna agregació concreta. Començarem la operació tot definint una variable booleana κ per a cada element i-èssim de `users` de la manera següent:

$$\kappa_i = \begin{cases} 1 \text{ si } \texttt{user}_i.\texttt{l\_sess} < q_{50\%}(\texttt{l\_sess}) \\ 0 \text{ altrament} \end{cases}$$

És a dir $1 = \kappa_i \in \{0,1\}$ si i només si l'usuari i-èssim presenta un valor de l'atribut `l_less` més petit que la mediana d'aquest atribut per a la totalitat de casos de `users`. Evidentment, si $\kappa_i = 1$ ha de significar que l'usuari en qüestió és classificat com havent abandonat prematurament, aleshores, per definició, seran considerats així un nombre molt proper a la meitat.

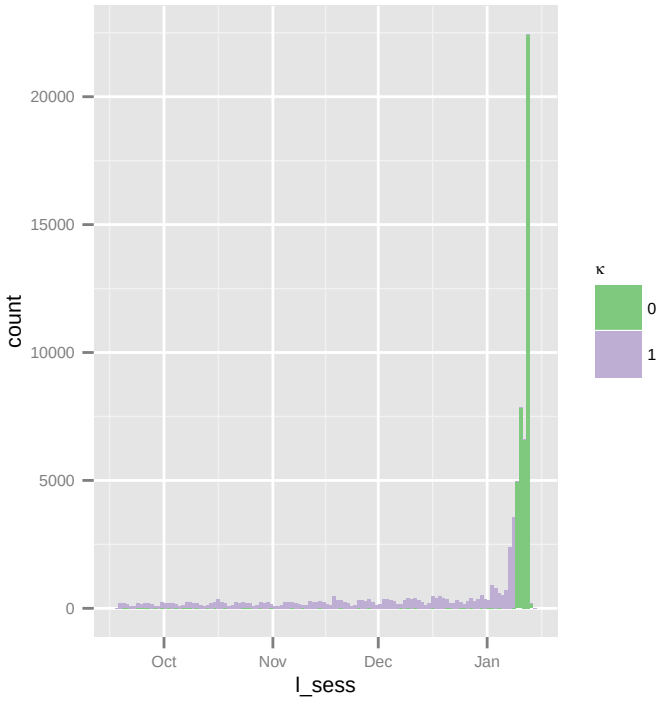


Figura 5.9: Distribució del nombre d'usuaris en funció del valor de `l_sess`. El color indica el valor de κ.

Efectivament, en la mostra de què disposem $q_{50\%}(\texttt{l\_sess}) = \texttt{2014-01-09}$ i aquesta sembla una data massa tardana com per a reflectir fidelment la noció d'*haver-ho deixat abans d'hora*. Tot i l'aparent arbitrarietat de la tria d'aquesta línia divisòria, en realitat ja ens serà prou útil com per a il·lustrar el que s'exposa a continuació.

Definim tot seguit la mitjana mostral $\overline{\kappa}$ per a cada *cluster* de la partició.

$$\overline{\kappa}_s = \frac{1}{|s|} \sum_{\texttt{user}_i \in s} \kappa_i$$

Amb això ja som en condicions de mesurar la precisió d'una partició qualsevol del conjunt `users` pel que fa la predicció del valor κ de cada un dels elements que inclou. Per definició, la probabilitat que $\kappa_i = 1$ d'un element $\texttt{user}_i$ pres aleatòriament és molt propera a 0.5. Així, definirem la funció que mesura la precisió d'una partició S qualsevol tot calculant la mitjana aritmètica següent:

$$\overline{\texttt{acc}}_k = \frac{1}{|S|} \sum_{s \in S} \texttt{acc}(s_i)$$

On `acc` és una funció sobre l'espai de d'agregacions definida com s'indica a continuació.

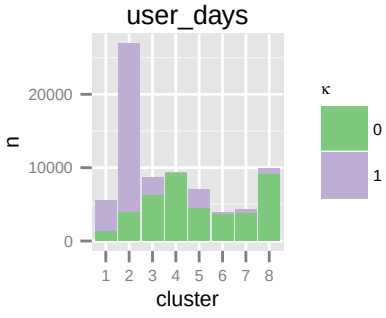


Figura 5.10: Distribució del nombre d'elements de `user_days` en funció del *cluster* a què han estat assignats amb el mètode 8-*means*. El color indica el valor de κ.

$$\texttt{acc}(s) = \begin{cases} 2 \cdot \overline{\kappa}_s - 1, \text{ si } \overline{\kappa}_s > 0.5 \\ 1 - 2 \cdot \overline{\kappa}_s, \text{ altrament} \end{cases}$$

És a dir, si per a més de la meitat dels elements de s és el cas que $\kappa = 1$, aleshores $\texttt{acc}(s)$ mesura la proporció en què llur nombre supera el 50% del total. Si no és així, mesura en quina proporció ho fa el nombre de casos en què $\kappa = 0$. Així doncs, $\overline{\texttt{acc}}$ és una mesura de fins a quin punt la partició S és una bona representació del valor de κ per a cada un dels elements de `users`.

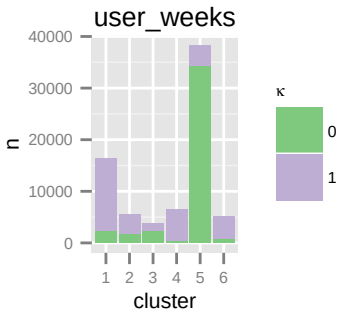


Figura 5.11: Distribució del nombre d'elements de `user_weeks` en funció del *cluster* a què han estat assignats amb el mètode 6-*means*. El color indica el valor de κ.

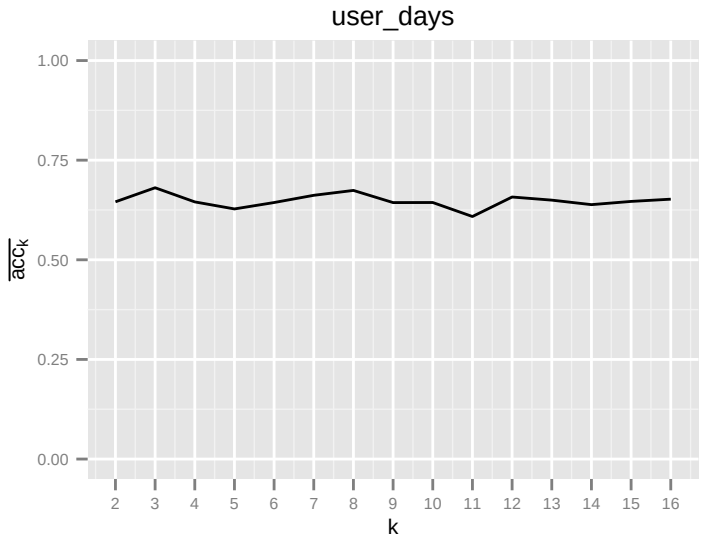


Figura 5.12: Proporció d'usuaris assignats a cada agregació resultant del 6-*means* en `week_days`

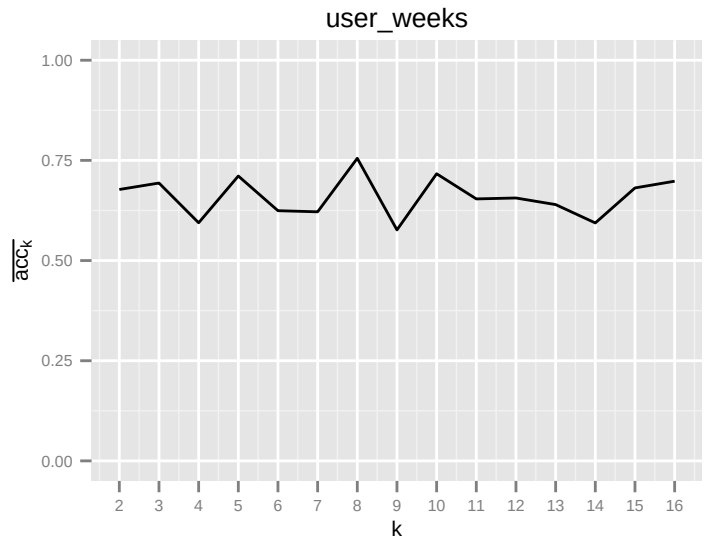


Figura 5.13: Proporció d'usuaris assignats a cada agregació resultant del 6-*means* en `week_days`

Ara bé, l'he construïda precisament per a mostrar que *no* consisteix en un bon predictor. A tall d'exemple, en la figura 5.10 es mostra

quina és la proporció d'usuaris de cada *cluster* en funció del valor de κ en la partició obtinguda per a `user_days` i en la 5.11 s'hi mostra el mateix pel que fa `user_weeks`. Si bé en algunes agregacions sembla que el valor de κ és un bon descriptor dels usuaris que hi han estat assignats, en realitat no proporciona prou solidesa com per a prendre'l com a tal en general. De fet, els valors de $\overline{\texttt{acc}}(S)$ són força modestos per a tots els de k que hem contemplat fins ara. Observi-se'n l'evolució en funció del nombre de *clusters* emprat que es mostra en les gràfiques de les figures 5.12 i 5.13, respectivament.

La conclusió útil a extraure d'això és que, si bé ens ha proporcionat certa intuició de quina és l'estructura interna del conjunt d'usuaris presents en les dades dades que hem rebut, el mètode d'agregació *k-means* no ens servirà per a fer prediccions útils per als nostres propòsits a partir només del que elles soles proporcionen.

6
Model de predicció d'abandonament del CV

Malgrat les dificultats vistes fins ara, hi ha un altre plantejament del problema del qual potser podrem obtenir resultats més fructífers. Recordem que, en la transformació de les dades introduïda en la secció 4.2, hem construït, per a cada element de `user_days` i `user_weeks`, un conjunt de $m+1$ nous atributs $a_0, \ldots, a_m$ amb domini $\{0,1\}$, que indiquen si l'usuari en qüestió s'ha connectat al CV de la UOC en el període $[t_0 + dj, t_0 + d(j+1)]$ per a certa durada d. Ara bé, aquest conjunt de nous atributs són en realitat una seqüència de valors booleans que, ordenats cronològicament, descriuen l'evolució de la presència al campus de l'usuari a què pertanyen. Així doncs, podem considerar que a cada usuari i-èssim li correspon una seqüència binària de dígits de longitud $m+1$ que en descriuen el comportament.

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

Mirarem ara d'obtenir informació útil a partir d'aquest nou enfocament.

6.1 *La seqüència de valors de $a_0, \ldots, a_m$ com una cadena de Markov*

En primer lloc, modelarem el procés estocàstic representat per part de les variables $a_0, \ldots, a_m$ com una *cadena de Markov de temps discret* (DTMC) [**?** , p. 1-47]. És a dir, considerarem cada $a_{i,j}$ com una variable aleatòria que representa l'estat de l'usuari i-èssim durant l'interval I_j i segueix una distribució de Bernoulli. Com és obvi, l'espai d'estats d'aquest objecte serà booleà, això és, el conjunt de valors $\{0,1\}$.

Addicionalment, prendrem una cadena de grau 1, és a dir, serà un model en què l'estat que hagi pres una variable corresponent a l'interval I_j depèn solament del que hagi pres aquella que representa l'interval I_{j-1} immediatament precedent i en cap mesura de les

anteriors a aquesta.

Més formalment, si $x \in \{0,1\}$ i $P(a_j = x_j, \ldots, a_0 = x_0) > 0$, aleshores és el cas que:

$$P(a_{j+1} = x | a_j = x_j, \ldots, a_0 = x_0) = P(a_{j+1} = x | a_j = x_j)$$

A més, suposarem que es tracta d'un model homogeni en el temps, és a dir, que els valors de les respectives probabilitats condicionals del pas d'un estat a un altre no varien en funció de l'interval temporal a què corresponguin les variables que estiguem considerant.

Totes aquestes restriccions permeten representar el comportament de la nostra cadena de Markov com una màquina d'estats finits com la que es mostra en la figura 6.1.

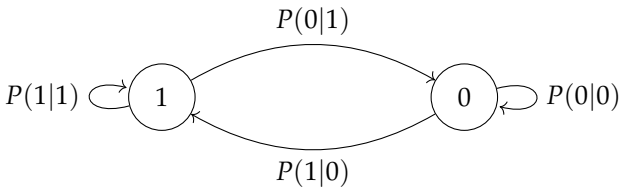


Figura 6.1: Representació del patró de la cadena de Markov de grau 1 i homogènia en temps discret a emprar per a modelar l'evolució de la presència de cada un dels usuaris al CV de la UOC.

Molt sovint es sintetitza el seu comportament en una matriu de transicions

$$\begin{pmatrix} P(0|0) & P(1|0) \\ P(0|1) & P(1|1) \end{pmatrix}$$

en què les files corresponen als possibles valors de a_j i les columnes als de a_{j+1}. Ara bé, per la regla de la suma, en un espai d'estats binari com el nostre, l'obtenció de la matriu de transicions es redueix a l'estimació de dos paràmetres α i β de la manera següent:

$$\begin{pmatrix} 1-\alpha & \alpha \\ 1-\beta & \beta \end{pmatrix}$$

Estimació de models

Noti's que els valors concrets que puguin prendre el parell de paràmetres α i β es poden interpretar com una caracterització del comportament d'un usuari pel que fa la seva presència al CV de la UOC. És a dir, podem reduir les tuples que pertanyen, respectivament, a `user_days` i `user_weeks` als atributs `user_id`, `alpha` $= \alpha$ i `beta` $= \beta$. Per a cada usuari i-èssim, mirarem d'estimar quins són els valors de α i β que defineixen la DTMC de grau 1 homogènia que és més probable que hagi generat la seqüència de valors booleans presents en $a_{i,0}, \ldots, a_{i,m}$ que li corresponen.

En primer lloc, val a dir que, per què l'estimació tingui sentit, cal prescindir d'aquells usuaris que presentin el valor 1 per tot a_j ($j \in [0, m]$)[1], el nombre dels quals ha resultat ser menyspreable (en el cas de `user_days` no n'hi ha hagut cap, i solament 2 en `user_weeks`).

[1] Noti's que, per la naturalesa mateixa de les dades sobre què estem treballant, és impossible que hi hagi cap usuari que presenti el valor 0 per a tot a_j

Fet això, per a estimar α i β s'ha emprat el mètode de la màxima versemblança (MLE). En general, si S és el conjunt d'estats possibles i $x, y, z \in S$, s'estima $P(y|x)$ tot calculant el paràmetre $\widehat{p}_{x,y}$ de la manera següent [**?**][2]:

$$\widehat{p}_{x,y} = \frac{n_{x,y}}{\sum_{z \in \{0,1\}} n_{x,z}}$$

[2] Així és com es realitza l'estimació en la funció `markovchainFit` del paquet `markovchain` de R que he emprat.

On $n_{x,y}$ indica el nombre transicions de x a y que s'han observat. Per tant, α i β s'han obtingut de la manera següent:

$$\alpha = \frac{n_{0,1}}{\sum_{z \in \{0,1\}} n_{0,z}} \qquad \beta = \frac{n_{1,1}}{\sum_{z \in \{0,1\}} n_{1,z}}$$

Dispersió dels resultats

Passem tot seguit a mostrar els resultats de l'estimació de α i β per a cada element de `user_days` i `user_weeks`. A la gràfica de la figura 6.2 es mostra un diagrama de dispersió on es situa cada cas de `user_days` en funció dels valors de α i β. A la de la figura 6.3 es fa el mateix per als elements de `user_weeks`. Observi's que en tots dos casos, tal com s'havia fet en el diagrama de la figura 3.17, els punts presenten certa transparència (el canal *alpha* és reduït al 10%) per poder apreciar-ne millor les aglomeracions.

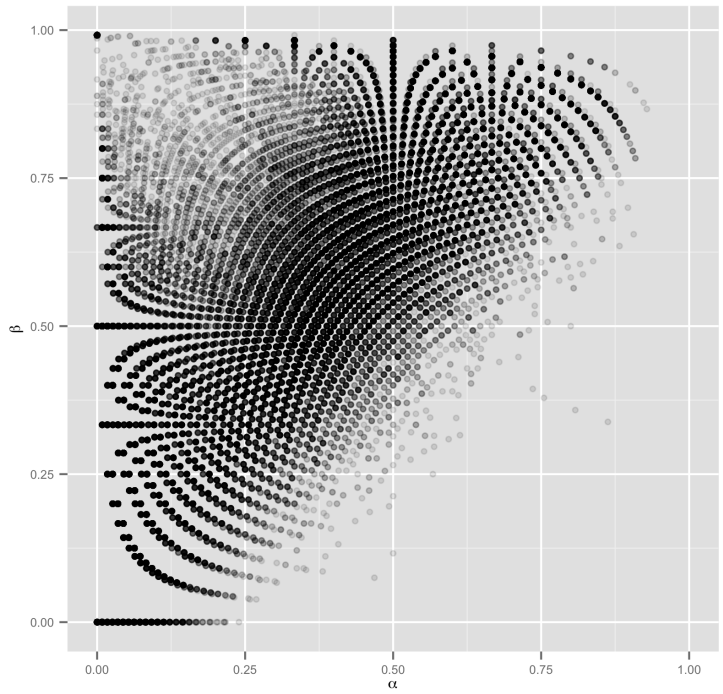


Figura 6.2: Diagrama de dispersió on es representen els valors de α i β per als elements de `user_days`.

Si bé la finalitat d'emprar el model de les cadenes de Markov era mirar de reflectir millor la naturalesa estocàstica de l'evolució

de la presència de cada usuari al CV de la UOC. A més, la simple estimació de α i β era especialment atractiva ja que permet reduir-ne la caracterització a dos camps i evitar els inconvenients d'emprar els $m+1$ atributs (recodem que $m = 119$ per a `user_days` i $m = 17$ per a `user_weeks`), com ja es pot copsar a primer cop d'ull, *els resultats no són especialment esperançadors*.

En el diagrama de la figura 6.2 els casos estan repartits d'una manera tan uniforme en el pla que es fa virtualment inviable el reconeixement de *clusters* de tal manera que aquests ens proporcionin una informació útil. I és que si bé és possible aplicar mètodes de *clustering* en una mostra com aquesta, com és obvi, les agregacions obtingudes ens seran de poca utilitat, ja que l'homogeneïtat de la distribució dels casos els resta poder explicatiu. Pel que fa els resultats sobre `user_weeks`, que es mostren a la figura 6.2, presenten algunes diferències, tot i que aconsegueixen esquivar les mateixes dificultats. En aquest cas, gran quantitat d'usuaris es superposen d'una manera quasi exacta en certs punts concrets del pla. Això és una conseqüència natural del fet que les cadenes de valors booleans a partir de les quals hem estimat els paràmetres presentin una longitud menor, ja que aquest fenomen redueix l'espai de possibles valors de $n_{x,y}$. Malgrat tot, la distribució dels casos en el pla és, en general, igualment homogènia, i un procés de reconeixement de *clusters* donaria lloc a resultats igualment infructuosos.

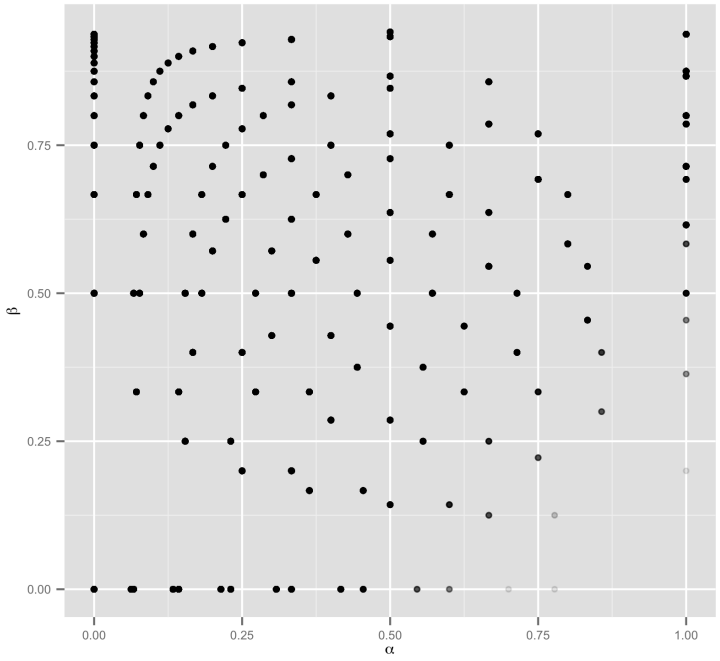


Figura 6.3: Diagrama de dispersió on es representen els valors de α i β per als elements de `user_weeks`.

Evidentment, poden ésser moltes i molt variades les raons de la inviabilitat d'aquest model, ja que, al cap hi a la fi, ja hem reconegut d'entrada la simplificació que comporta. Potser caldria augmentar el

grau de la cadena de Markov, o bé no fer-la homogènia en el temps i fer variar les probabilitats de les transicions en funció del moment del temps representat. Tot i això, he decidit no seguir en aquesta direcció, ja que em sembla haver trobat un model molt millor, que mantingui la simplicitat i, a la vegada, ofereixi prediccions útils per als nostres propòsits.

6.2 *Modelat mitjançant models ocults de Markov*

Si bé tot model és una simplificació, el que hem triat en la secció anterior pateix d'un enfocament massa ingenu. En ell estàvem fent dependre la presència o absència d'un alumne al CV en un interval I_j només de la presència o absència del mateix alumne en l'interval I_{j-1}. Com és natural, hem obtingut un perfil diferent per a cada combinació possible de valors de a_j, perfil que, a més, només hem pogut produir *a posteriori*.

Tanmateix, com apuntava en la secció 5.2, no és pas aquesta informació la que ens interessa, sinó més aviat la probabilitat que un usuari qualsevol hagi abandonat el curs per a cada interval I_j. És a dir, per a cada seqüència de valors $a_{i,0}, \ldots, a_{i,k}$, corresponents a l'usuari i-èssim, ens interessa saber quina és la probabilitat que per tot $a_{i,j} = 0$ amb $k < j \leq m$. A més, seria òptim poder obtenir aquesta informació *a priori*, és a dir, *mentre* ens trobem en el transcurs de l'interval I_k o en un moment anterior. D'aquesta manera sí que podríem predir l'abandó per part d'usuaris del campus (una part dels quals sabem que són estudiants matriculats) abans que hagi finalitzat el semestre.

Com es veurà tot seguit, pretenc haver aconseguit un model que satisfà totes aquestes condicions i que, a més, manté una simplicitat notable. El model en qüestió és un cas concret del model ocult de Markov (HMM).

Descripció del model

Com és ben sabut [**?** , p. 588-91,835], en un HMM els diferents estats concrets, és a dir, els valors que prenen cada una de les variables x_j que componen la seqüència estocàstica es suposen desconeguts. Tot i això, per a cada variable oculta x_j se n'afegeix una de nova y_j el valor que prengui (és a dir, l'estat en què es trobi) sí que serà conegut i dependrà de la probabilitat condicional $P(y_j|x_j)$. En la figura 6.4 es mostra una representació gràfica de l'evolució temporal d'aquest sistema.

Així doncs, en aquest cas l'estructura del HMM vindrà definida per:

(1) Un espai d'estats S que poden prendre les variables x_j que conformen la cadena de Markov oculta.

(2) Un espai d'estats E (noti's que és possible que $S \neq E$) que poden prendre les variables observables y_i.

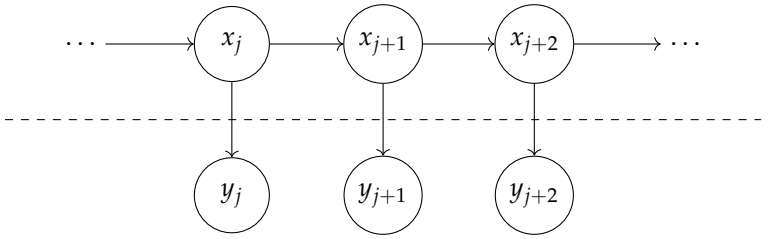


Figura 6.4: Esquema de l'evolució temporal d'un model ocult de Markov.

(3) Un estat inicial $s_0 \in S$.

(4) Una matriu de transicions dels estats de la cadena de Markov oculta.

(5) Una matriu d'emissions que expressi la probabilitat $P(y|x)$ d'observar cada emissió y condicionada per la probabilitat de l'estat x.

Tot tenint present aquest esquema general, podem modelar el cas que ens ocupa de la manera següent:

(1) Suposarem que cada usuari pot trobar-se, per a cada interval I_j, en un dels dos estats del conjunt $S = \{A, Q\}$ en funció del següent:

   i. L'usuari pot ésser *actiu*, és a dir, que *està pendent* d'allò que s'esdevé al CV. Anomenarem A aquest estat.

   ii. Per contra, també pot *haver abandonat* el CV, això és, que no tornarà a establir-hi cap sessió fins a final de semestre. Anomenarem Q aquest segon estat.

(2) Considerarem que tots els usuaris comencen el semestre essent actius, és a dir, que no n'hi ha cap que, durant el primer interval I_0 ja hagi abandonat. Per tant l'estat inicial serà A.

(3) Per altra banda, considerarem ara els valors de $a_{i,0}, \ldots, s_{i,m}$ com les emissions per part de les variables ocultes corresponents del model a què pertany a l'usuari i-èssim. Així doncs, l'espai d'estats variables observables és $E = \{0, 1\}$

(4) Per definició, si un usuari ha abandonat el CV ja no tornarà a connectar-s'hi. Per tant, considerarem que Q és un estat absorvent, és a dir, que $P(A|Q) = 0$ i $P(Q|Q) = 1$. D'aquesta manera, la matriu de transicions tindrà l'aspecte següent:

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

   On α és un dels paràmetres que determina el model a estimar.

(5) Per altra banda, també és obvi que si un usuari ha trencat la reva relació amb el CV, aleshores tampoc ja no hi establirà més

sessions. És a dir, $P(0|Q) = 1$ i $P(1|Q) = 0$. Conseqüentment, la matriu d'emissions és:

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

On β és l'altre dels paràmetres a estimar.

En la figura 6.5 pot veure's un esquema gràfic del model que acabem de descriure.

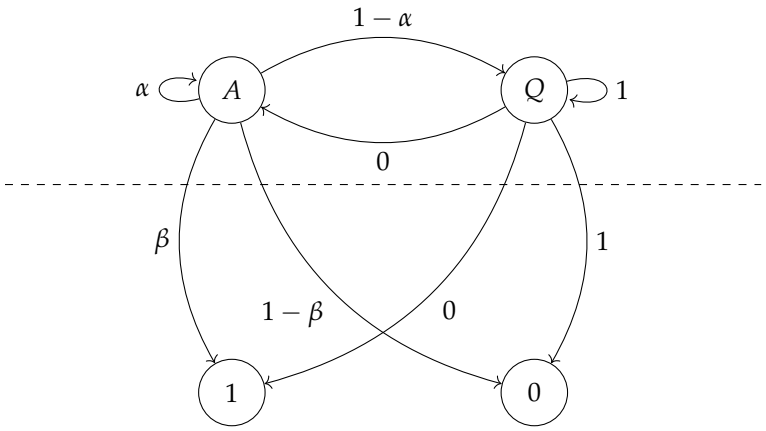


Figura 6.5: Esquema del model ocult de Markov per a predir l'abandonament, per part dels usuaris, de la relació amb el CV.

Així doncs, hem reduït la solució del problema a la tasca d'haver d'estimar els dos paràmetres α i β.

Estimació de paràmetres

Com durem a terme aquesta estimació? En vistes a una validació creuada, ho farem a partir d'una mostra que anomenarem `users_train`, independement de si és un subconjunt de `user_days` o de `user_weeks`. Per altra banda, si bé és cert que hi ha una gran varietat d'algoritmes per a estimar els paràmetres d'un HMM n'hi haurà prou emprant un mètode molt simple.

Notem en primer lloc, que α és la probabilitat que un usuari que es troba actiu durant un interval I_j, segueixi essent-ho durant l'immediatament posterior I_{j+1}. Per definició, sabem que com a mínim per a un 50% dels usuaris, aquests es trobaran actius des de l'interval I_0 fins aquell que inclou la mediana mostral $q_{50\%}(\texttt{l\_sess})$ dels valors que pren en el conjunt d'entrenament `users_train` el camp `l_sess`, el qual indica, recordem, quin és el `central_activity_point` de la darrera sessió de l'usuari. Podem aproximar aquest fet tot suposant que, si prenem un usuari qualsevol, la probabilitat que aquest encara sigui actiu en l'interval I_j que conté el valor d'aquesta mediana, és de 0.5.

Per altra banda recordem que, en una DTMC, si coneixem la distribució de probabilitats de l'estat inicial podem preveure la probabilitat de l'estat en el pas j-èssim. Per al nostre model concret, si

definim $\lambda \in [0, m]$ com l'índex de l'interval que conté $q_{50\%}(\texttt{l\_sess})$ l'operació que proporciona aquesta informació és la següent:

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

Sabem, doncs, que α^{λ} és la probabilitat de $P(A)$ en l'interval I_{λ}. Però ja hem dit que aproximaríem aquest valor a 0.5. Per tant:

$$\alpha^{\lambda} = \frac{1}{2} \Rightarrow \alpha = \left(\frac{1}{2}\right)^{\frac{1}{\lambda}}$$

Hem pres una mostra aleatòria del 50% per a confeccionar `train_users` en cada cas i, n'hem obtingut els valors següents:

1. Per a $\texttt{train\_users} \subseteq \texttt{user\_days}$

$$\lambda = 113 \Rightarrow \alpha = 0.9938847$$

2. Per a $\texttt{train\_users} \subseteq \texttt{user\_weeks}$

$$\lambda = 16 \Rightarrow \alpha = 0.9576033$$

Mirem ara d'estimar β. Noti's que aquest paràmetre representa la probabilitat que un usuari estableixi alguna sessió durant un iterval qualsevol mentre es troba actiu (estat A). Així, doncs, per a cada $\texttt{user}_i$ podem definir la subseqüència de valors booleans $a_{i,0}, \ldots, a_{i,u}$ ($u \leq m$) tal que $\texttt{user}_i.\texttt{l\_sess} \in I_u$, és a dir, la subseqüència durant la qual ha mantingut aquest estat. Així, definirem el nou camp següent per a cada usuari i-èssim,

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

que representa la *taxa de presència* al CV de la UOC que aquest usuari ha tingut mentre encara no l'ha abandonat.

El paràmetre β l'estimarem simplement prenent la mitjana mostral del valor d'aquest nou camp:

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

Els valors concrets obtinguts són:

1. Per a $\texttt{train\_users} \subseteq \texttt{user\_days}$

$$\beta = 0.4087153$$

2. Per a $\texttt{train\_users} \subseteq \texttt{user\_weeks}$

$$\beta = 0.7235815$$

Resultats de la predicció

Ja estem en condicions, doncs, de validar el nostre model. El conjunt de validació pres ha estat, simplement en complementari

$$\texttt{users\_test} = \texttt{users\_train}^C$$

en `user_days` i `user_weeks`, respectivament.

Obervi's que en realitat es fa una predicció per a cada interval de la discretització en funció de d. És a dir, per a cada interval I_j, el model predirà quin és l'estat ocult en el qual és més probable que es trobi l'usuari i-èssim en funció de la seqüència d'observacions $a_{i,0}, \ldots, a_{0,j-1}$ i dels paràmetres que acabem d'estimar. La predicció en qüestió l'he realitzada per mitjà de l'aplicació de l'*algoritme de Viterbi* [**?**] [3].

[3] Aquest és l'algoritme que segueix la funció `viterbi` inclosa en el paquet d'R que porta per nom `HMM` [**?**].

Per a cada interval I_j amb $j \in [1, m]$, doncs, podem construir una matriu de confusió com la que es mostra a continuació, els valors de la qual avaluen la utilitat del model a l'hora de predir si l'usuari ja ha abandonat el CV en aquest interval.

$$\begin{pmatrix} TP & FP \\ TN & FN \end{pmatrix}_j$$

Noti's que les files es refereixen, en l'ordre $\langle Q, A \rangle$, als valors observats que corresponen als estats de la cadena oculta de Markov, i les columnes, també en el mateix ordre, als valors observats. Les inicials dels valors indicats signifiquen, per tant:

(1) TP (*True Positives*): El nombre d'usuaris en què s'ha predit que es trobaven en l'estat Q i la predicció ha estat correcta.

(2) TN (*True Negatives*): El nombre d'usuaris en què s'ha predit que es trobaven en l'estat A i la predicció ha estat correcta.

(3) FP (*False Positives*): El nombre d'usuaris en què s'ha predit que es trobaven en l'estat Q i la predicció ha estat incorrecta.

(4) FN (*False Negatives*): El nombre d'usuaris en què s'ha predit que es trobaven en l'estat A i la predicció ha estat incorrecta.

En les figures 6.6 i 6.7 s'hi mostren, en percentatges, les evolucions respectives d'alguns d'aquests valors en funció del temps per a `user_days` i `user_weeks`.

Les variables que hi apareixen es calculen com s'indica tot seguit:

$$\texttt{accuracy} = 100\frac{TP+TN}{TP+TN+FP+FN} \qquad \texttt{false\_negatives} = 100\frac{FN}{TP+TN+FP+FN}$$

$$\texttt{false\_positives} = 100\frac{FN}{TP+TN+FP+FN} \qquad \texttt{quit\_users} = 100\frac{TP+FN}{TP+TN+FP+FN}$$

Observi's que en el primer cas la precisió del model és força pobra a partir de finals de setembre però millora a mesura que avança el semestres. Per altra banda, cosa que és més rellevant, la major part

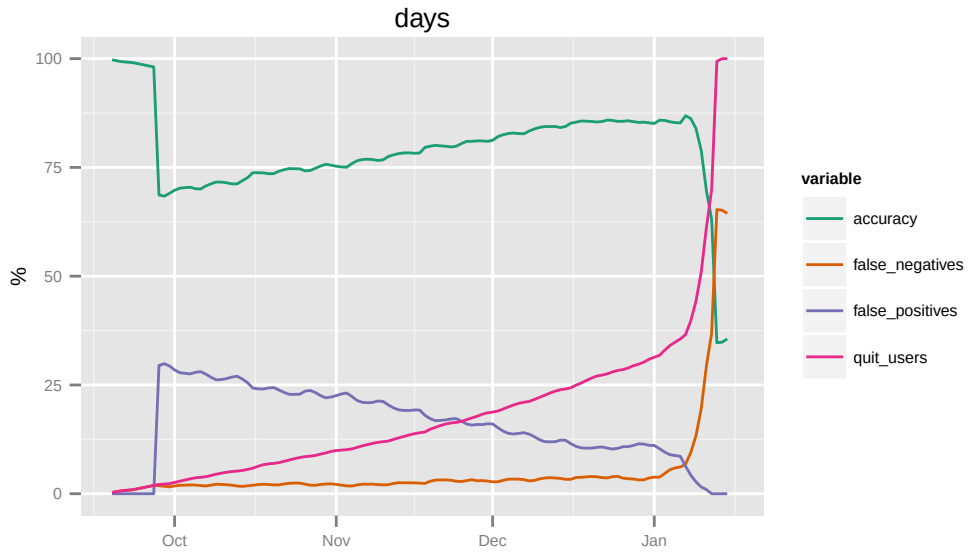


Figura 6.6: Evolució dels valors de la matriu de confusió en funció del temps per a `user_days`

de les errades consisteixen en falsos positius, és a dir, prediccions d'abandonament per part d'usuaris que en realitat seguien actius. Això és un mal menor ja que el que ens importa, en vistes a la millora del procés educatiu, és evitar que ens passin per alt abandonaments, és a dir, els falsos negatius.

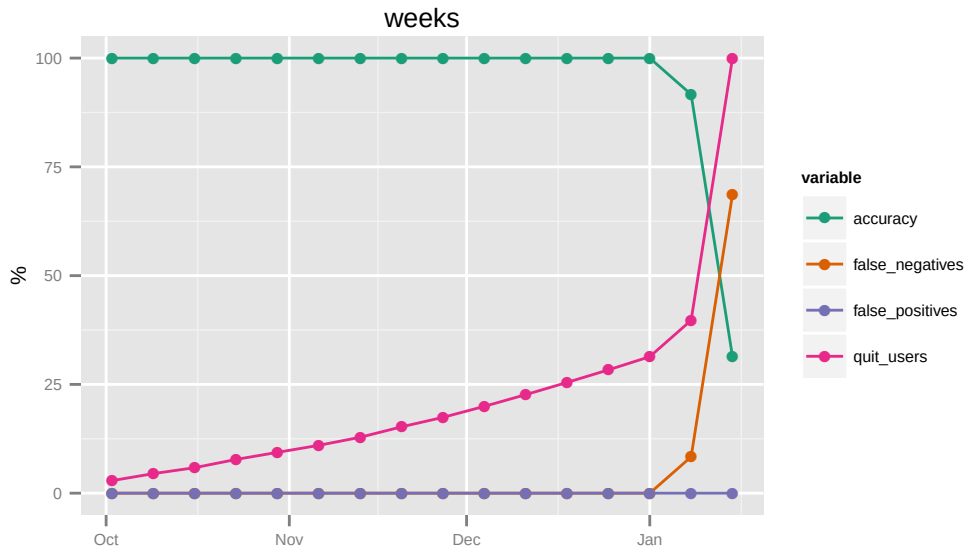


Figura 6.7: Evolució dels valors de la matriu de confusió en funció del temps per a `user_weeks`

Per acabar, notem que la fiabilitat del model en la discretització per setmanes és molt més alta. La precisió és del 100% des de l'inici del semestre fins a final d'any.

7
Conclusions

Clourem aquest treball tot fent una síntesi dels resultats més valuosos que hem pogut recollir de l'estudi que presenta.

7.1 Utilitat de la discretització

Recordem en primer lloc que les dades que hem rebut presenten, en resum, les dificultats següents:

(1) No coneixem en cap mesura el procés que ha donat lloc a llur generació. Per tant, no sabem fins a quin punt el detall dels valors dels camps que inclouen són un bon reflex dels fets que representen.

(2) Tampoc sabem si cada un dels usuaris representats per part dels identificators que apareixen en les dades corresponen o no a estudiants de la UOC, que són l'objecte que ens interessa en aquest estudi.

(3) En darrer terme, fins i tot si un identificador d'usuari pertany a un estudiant, també és ben poca la informació que cada element de `sessions` proporciona sobre el procés d'aprenentatge en què es troba immers.

Així, doncs, la mètrica presència / absència per interval de temps per a cada usuari és una manera d'extraure informació fiable a partir de les dades en brut. De fet, a partir d'elles, és ben poc el que es pot dels usuaris més enllà del que expressa aquesta variable.

*7.2 Agregació per k-*means *per a identificar tipus de comportament*

L'aplicació del mètode *k-means* a partir del resultat de la discretització ha estat reeïxida a l'hora d'identificar una jerarquia de perfils de comportament dels usuaris pel que fa llur relació amb el CV durant el transcurs del semestre. Tot i que no sapiguem quins usuaris són estudiants, noti's que aquesta tipologia és útil per als propòsits de la *EDM&LA* ja que, no només inclouen la caracterització dels que sí que ho són, sinó que estableixen un primer criteri per a discriminar-los respecte altres.

7.3 L'eficàcia predictiva dels models ocults de Markov

En darrer lloc, entenc que el resultat més valuós de l'estudi que s'exposa en aquesta memòria és el model ocult de Markov (HMM) els paràmetres dels quals són determinats d'una manera especialment simple. Pel que fa la seva precisió a l'hora de predir els abandonaments per part dels usuaris ha estat del 100% fins a finals de desembre pel cas de la discretització setmanal.

A més, aquesta predicció segueix essent útil pel que fa els nostres objectius encara que no poguem saber quins dels usuaris són estudiants, ja que l'avís d'abandó és igualment eficaç per a tots ells, i és d'esperar que una aplicació pràctica d'aquest model ja faci aquesta distinció.