Pràctica 2
Explorant el filtratge col·laboratiu: Personalització i recomanació de dades

- El termini d'entrega de la pràctica finalitza el **18 de desembre de 2.024 a les 23:55**.
- L'entrevista de la pràctica es farà durant la sessió de laboratori del **19 de desembre de 2.024**.
- Al Campus Virtual heu de penjar l'**informe PDF** explicant com heu resolt la pràctica, els **codis/scripts** utilitzats, i els **fitxers de resultats** obtinguts.
- La pràctica es realitza en grup de **dues o tres persones**.
- La nota d'aquesta pràctica equival a un **15%** de la nota global de l'assignatura.

Motivació i objectius

La sobrecàrrega d'opcions que tenim en el món digital actual és aclaparadora. Des de la selecció de pel·lícules en plataformes de streaming fins a la tria de productes en línia, la necessitat de recomanacions personalitzades s'ha tornat cada vegada més evident. Per abordar aquest repte, moltes organitzacions utilitzen tècniques de filtratge col·laboratiu, les quals permeten generar recomanacions basades en les preferències i comportaments d'usuaris similars.

En aquesta pràctica, treballarem habilitats pràctiques en el desenvolupament de sistemes de recomanació efectius. Més concretament, s'espera que els alumnes puguin:

- Comprendre els principis bàsics del filtratge col·laboratiu i la seva importància en la generació de recomanacions personalitzades.
- Explorar diferents tècniques i algoritmes utilitzats en el filtratge col·laboratiu, incloent el filtratge basat en usuaris i en ítems.
- Implementar mètodes per tal d'inferir/predir valoracions d'usuaris a ítems sobre un conjunt de dades.
- Avaluar l'efectivitat del sistema de recomanació desenvolupat mitjançant mètriques d'avaluació adequades.

Metodologia

Conjunt de dades

Es proporciona un conjunt de dades $\mathcal{D}$ que conté informació sobre la valoració que han fet usuaris respecte una colla de restaurants. Aquest conjunt de dades s'entén com una matriu de 73,421 usuaris (files) i 100 restaurants (columnes). La intersecció de cada fila i columna indica la valoració v_{ij} d'un usuari i a un restaurant j. Les valoracions van entre -10 i 10

(en nombres reals). Els valors desconeguts es representen amb el valor 99. La densitat del conjunt de dades $\mathcal{D}$ és del 55.8%, és a dir, coneixem el 55.8% de les valoracions, però un 44.2% dels valors són desconeguts. Trobareu el conjunt de dades $\mathcal{D}$ emmagatzemat en format CSV (separat per ";") en el fitxer `recommendation_dataset.csv`. Aquest conjunt de dades té el següent estil:

	Restaurant1	**Restaurant2**	**...**	**Restaurant100**
User1	-7.82	8.79	...	99
User2	99	-0.29	...	1.07
...	...	...	⋱	...
User73421	99	99	...	99

Cada fila correspon a les valoracions que ha fet un determinat usuari (vector usuari), i cada columna correspon a les valoracions que han fet tots els usuaris sobre un restaurant (vector ítem).

Mètode de prediccions

Tenint això clar, haureu d'inferir/estimar els valors desconeguts que hi ha al conjunt de dades $\mathcal{D}$. Mitjançant mètodes de filtratge col·laboratiu, calcularem els valors desconeguts i tindreu un conjunt de dades $\mathcal{D}'$. Alguns mètodes que podeu utilitzar són els següents:

- Imputació per valor central: Es prediu que tots els valors desconeguts són el mateix valor, entès com la mitjana de totes les valoracions que tenim.
- Imputació per mitjanes d'usuaris/ítems: Es prediu que tots els valors desconeguts de cada vector usuari/ítem són el mateix valor, entès com la mitjana de totes les valoracions que tenim a cada vector usuari/ítem.
- Imputació per distàncies: S'identifiquen els k vectors usuaris/ítems més propers de cada vector usuari/ítem i s'agreguen els valors coneguts per fer les prediccions. Podeu fer servir l'algorisme k-NN[1] utilitzant diverses mètriques de distància (per exemple, Euclídea, Manhattan, Mahalanobis...).
- Algorismes de clustering com, per exemple, k-means[2].
- Xarxes neuronals.
- Combinació de varis mètodes anteriors.
- ...

Podeu utilitzar els mètodes anteriors o proposar-ne d'altres (inventats o no). Podeu utilitzar els llenguatges de programació (Python, R, Java,...), les llibreries (NumPy, scikit-learn, TensorFlow,...), i programari (Weka,...) que preferiu.

[1] `https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm`
[2] `https://en.wikipedia.org/wiki/K-means_clustering`

Algunes recomanacions:

- Si el vostre ordinador té recursos limitats (especialment memòria RAM), degut a la mida del conjunt de dades $\mathcal{D}$, és possible que s'hagi de fer servir una base de dades SQL per emmagatzemar la matriu de distàncies, en cas d'implementar imputació per distàncies o bé algun mètode de clustering.
- Intenteu fer servir floats (generalment no cal precisió double) i així estalviareu memòria.
- Intenteu evitar càlculs redundants per reduir el temps de computació.

Test de resultats

L'objectiu del vostre sistema de recomanació és aconseguir predir amb la màxima precisió possible el que votaran els usuaris. Per aquest motiu, el test consistirà en comparar alguns dels valors predits $\mathcal{P}$ (extrets de $\mathcal{D}'$) amb els seus respectius valors originals $\mathcal{O}$ (extrets de $\mathcal{D}$).

Del conjunt de dades $\mathcal{D}$, s'ha extret un 1% de valors (40,000 valoracions). A la matriu $\mathcal{D}$, aquestes valoracions extretes estan indicades amb el valor 99. Les valoracions originals ($\mathcal{O}$) que vosaltres haureu de predir venen donades en el fitxer `target_recommendations.csv` (que no s'ha de modificar!) amb el següent estil:

User25287;Restaurant31;-1.84
User45110;Restaurant50;6.02
...
User43204;Restaurant38;-9.17

Això vol dir que, a la matriu original, a la casella de l'usuari 25,287 i restaurant 31, hi havia un vot amb el valor -1.84. Vosaltres, dins de $\mathcal{D}$, trobareu un 99 en aquella cel·la.

Les vostres prediccions $\mathcal{P}$ hauran d'anotar-se en un fitxer, segons el format del fitxer `template.csv` (veureu que té el mateix format que `target_recommendations.csv`):

User25287;Restaurant31;p_1
User45110;Restaurant50;p_2
...
User43204;Restaurant38;p_n

, on els valors p_1, p_2, ..., p_n seran les prediccions que vosaltres heu fet (després d'haver-hi aplicat un mètode).

Un cop tingueu les vostres prediccions guardades en un fitxer (per exemple, anomenat `resultats.csv`), calcularem el MAE (Mean Absolute Error) entre els valors de les vostres prediccions (`resultats.csv`) amb els valors originals (`target_recommendations.csv`). El MAE (calculat mitjançant l'Equació 1) indica la mitjana de l'error absolut: com més petit sigui, menor és l'error i, per tant, millors són les prediccions. El vostre objectiu és obtenir el **mínim valor de MAE possible**.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|o_i - p_i| \tag{1}$$

, on n indica el total d'observacions (40,000), o_i indica el valor original, i p_i indica el valor predit.

Podeu repetir/modificar els vostres mètodes i veure com el MAE millora o empitjora amb cada configuració (per exemple, variant la cardinalitat dels grups si feu servir un algorisme de clustering, o bé incrementant la quantitat de veïns quan apliqueu k-NN, etc.).

Per a facilitar-vos el càlcul del MAE, se us ha proporcionat un script Python, anomenat `compute_mae.py`, que ja retorna el valor del MAE de les vostres prediccions. Aquest script, que no heu de modificar, s'executa de la manera següent:

```
python compute_mae.py path/to/target_recommendations.csv path/to/resultats.csv
```

Consideracions:

- A les recomanacions que guardeu al fitxer `resultats.csv`, utilitzeu el punt (.) com a separador decimal. Rebreu un error si utilitzeu la coma (,).
- L'script Python us retornarà un error si detecta algun error de format, per exemple, les prediccions línia a línia no coincideixen.

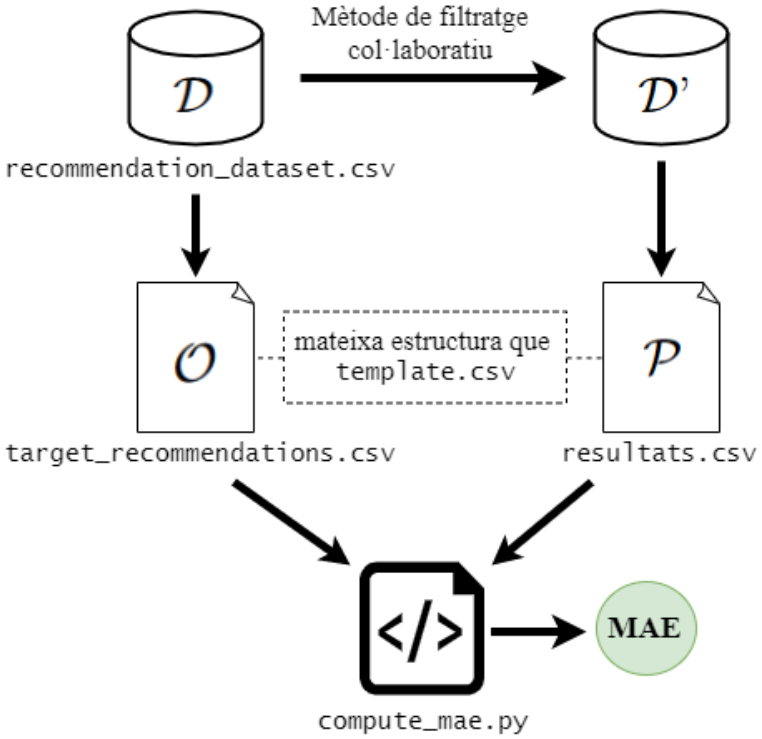

Figura 1: Resum gràfic de la metodologia de la pràctica.

Avaluació

L'avaluació de la pràctica dependrà del resultat del MAE que obtingueu. Aplicant el mètode d'imputació per valor central, el mètode més simple, s'obté un MAE de 4.43. Per aprovar la pràctica, haureu de millorar en un 10% aquest MAE, és a dir, haureu d'aconseguir un MAE inferior a 3.987. El grup que aconsegueixi el millor MAE tindrà un 10. Entre el mínim MAE exigit (3.987) i el millor MAE, es farà un escalat lineal que determinarà la vostra nota (segons l'Equació 2).

$$\text{Nota} = 5 + \frac{5 \cdot (\text{Vostre\_MAE} - 3.987)}{\text{Millor\_MAE} - 3.987} \tag{2}$$

El professor es reserva el dret de modificar la nota si detecta incoherències en el càlcul del MAE o còpies entre grups.

Lliurament

El lliurament de la pràctica es farà mitjançant la tasca habilitada al Campus Virtual fins el dia **18 de desembre de 2.024 a les 23:55**. Haureu d'entregar un fitxer ZIP que contingui:

- L'informe PDF[3] amb una breu explicació del mètode/mètodes emprats i indicar **clarament** quin MAE heu obtingut.
- El codi/script del mètode implementat.
- El fitxer `resultats.csv` amb les millors prediccions que heu aconseguit, és a dir, aquell que heu utilitzat per calcular el MAE.

Entrevista

L'entrevista de la pràctica es farà durant la sessió de laboratori del **19 de desembre de 2.024**. En aquesta entrevista, el professor farà preguntes als membres del grup per validar l'autoria de la pràctica i avaluar el nivell de coneixements adquirit pels estudiants. **És obligatori realitzar l'entrevista; en cas contrari, la pràctica no s'avaluarà.**

[3] Recomanem l'ús de LaTeX utilitzant la plataforma Overleaf: `www.overleaf.com`