PAC 1

Presentació

La nostra empresa ha creat una aplicació mòbil amb la qual els usuaris poden demanar suggeriments de restaurants. Un cop han visitat el restaurant, els usuaris fan una valoració del seu preu i qualitat que s'afegeix a una base de dades que ajuda el sistema a generar nous suggeriments pels usuaris, comparant usuaris amb preferències semblants.

En aquesta PAC 1 repassarem aquest conceptes bàsics de recomanació seguint el fil d'aquest exemple.

Competències

En aquest enunciat es treballen en un determinat grau les següents competències general de màster:

- Capacitat per a projectar, calcular i dissenyar productes, processos i instal·lacions en tots els àmbits de l'enginyeria en informàtica.
- Capacitat per al modelat matemàtic, càlcul i simulació en centres tecnològics i d'enginyeria d'empresa, particularment en tasques d'investigació, desenvolupament i innovació en tots els àmbits relacionats amb l'enginyeria en informàtica
- Capacitat per a l'aplicació dels coneixments adquirits i de solucionar problemes en entorns nous o poc coneguts dins de contextes més amplis i multidisciplinars, essent capaços d'integrar aquests coneixements.
- Posseir habilitats per a l'aprenentatge continuat, autodirigit i autònom.
- Capacitat per a modelar, dissenyar, definir l'arquitectura, implantar, gestionar, operar, administrar y mantenir aplicacions, xarxes, sistemes, serveis i continguts informàtics.
- Capacitat per assegurar, gestionar, auditar i certificar la qualitat dels desenvolupaments, processos, sistemes, serveis, aplicacions i productes informàtics.

Les competències específiques d'aquesta assignatura que es treballen són:

- Entendre que és l'aprenentatge automàtic en el context de la Intel·ligència Artificial
- Distingir entre els diferents tipus i mètodes d'aprenentatge
- Aplicar les tècniques estudiades a un cas concret

Objectius

En aquest PAC es practicaran els conceptes del temari relacionat amb recomanació i clustering, en una vertent pràctica amb un cas concret.

Descripció de la PAC a realitzar

Dades

La web triahotel.com ens dóna inicialment un fitxer de dades amb les opinions de 100 clients sobre els 20 hotels registrats a la web. Teniu en compte que els clients no han visitat tots els hotels, per la qual cosa no hi trobareu totes les combinacions hotel/client.

Cada usuari s'identifica amb un número de l'1 al 100. Cada hotel s'identifica amb un número de l'1 al 20. Cada usuari emet 8 opinions sobre cada hotel que visita, que corresponen a valoració general (valors 1..10), habitacions (1..5), situació (1..5), neteja (1..5), preu (1..10), recepció (1..5), restaurant (1..10) i altres serveis (1..5).

Aleshores cada línia del fitxer **hotels.data** és una valoració d'un hotel per un usuari i té el següent format (els valors separats per espais):

```
idHotel idUsuari valoració1 valoració2 ... valoració8
```

Els idHotels i idUsuaris estan desordenats.

Per l'apartat 4 hi ha un altre fitxer, **favorits.data**, amb l'hotel favorit de cada usuari. Aquesta parella usuari/hotel no surt al fitxer anterior.

```
idUsuari idHotel
```

Activitats

Per resoldre aquesta PAC heu de fer servir el codi dels programes 2.2-2.7 amb les modificacions escaients per treballar amb els nous fitxers de dades. També fareu servir el codi 4.4 (k-means) i algunes funcions de la biblioteca **sklearn**.

A totes les activitats cal que justifiqueu raonadament la vostra resposta.

Activitat 1

El pas previ a qualsevol anàlisi de dades és efectuar un tractament per tal d'adequar els seus valors. Entre les operacions habituals trobem la normalització o estandardització de les dades (veieu per exemple http://ca.wikipedia.org/wiki/Distribuci%C3%B3_normal).

Primer de tot es demana, doncs, que realitzeu el tractament previ de les dades que considereu necessari (o pot ser no cal fer-ne cap tractament previ).

Activitat 2

Feu un agrupament **k-means** (amb **K=4**) dels hotels tenint en compte com a vector de característiques de cada hotel les valoracions mitjanes de cada aspecte del hotel (és a dir valor mitjà assignat a valoració general, a habitacions, a situació, a neteja, etc.).

Feu servir una funció de distància adient per comparar hotels.

Activitat 3

Feu servir la mesura de qualitat d'agrupament **Adjusted Rand Index** per avaluar l'agrupament de l'apartat anterior. Aquest índex compara dos agrupaments per veure com de semblants són; pot donar un valor entre -1 i 1, on un 1 significa que dos agrupaments són idèntics i un valor negatiu significa que són molt diferents.

Aquest índex es pot calcular fent servir una funció de **sklearn**, com per exemple:

```
>>> from sklearn import metrics
>>> labels_true = [0, 0, 0, 1, 1, 1]
>>> labels_pred = [0, 0, 1, 1, 2, 2]

>>> metrics.adjusted_rand_score(labels_true, labels_pred)
0.24...
```

Calculeu el **Adjusted Rand Index** sabent que l'agrupament real dels hotels hauria de ser [1..5], [6..11], [12..16] i [17..20].

En el vostre exemple, **labels_true** hauria de ser el agrupament que heu obtingut per k-means i **labels_pred** l'agrupament real dels hotels.

Activitat 4

Genereu un recomanador ponderat basat en memòria i, per cada usuari, trobeu l'hotel més recomanable. Coincideixen les recomanacions amb l'hotel favorit de cada usuari llistat al fitxer **favorits.data**?

Recursos

Aquest PAC requereix dels següents recursos:

Bàsics: Fitxers de dades adjunts a l'enunciat

Complementaris:

Manual de teoria de l'assignatura. En especial les taules de codi 2.2-2.7 i 4.4.

Criteris de valoració

Els exercicis tindran la següent valoració associada:

Activitat 1: 1 punt

Activitat 2: 3 punts

Activitat 3: 2.5 punts

Activitat 4: 3.5 punts

Raoneu la resposta en tots els exercicis. Les respostes sense justificació no rebran puntuació.

Format i data de lliurament

La PAC s'ha de lliurar abans del proper 3 d'Abril (abans de les 24h).

La solució a entregar consisteix en un informe en format PDF fent servir la plantilla penjada al tauler de l'assignatura més els fitxers de codi (*.py) que heu fet servir per resoldre la prova. Aquests fitxers s'han de comprimir en un fitxer ZIP.

Adjunteu el fitxer a un missatge a l'apartat de **Lliurament i Registre d'AC (RAC).** El nom del fitxer ha de ser *CognomsNom*_IA_PAC1 amb l'extensió *.zip.*

Per a dubtes i aclariments sobre l'enunciat, adreceu-vos al consultor responsable de la vostra aula.

Nota: **Propietat intel·lectual**
Sovint és inevitable, en produir una obra multimèdia, fer ús de recursos creats per terceres persones. És per tant comprensible fer-ho en el marc d'una pràctica dels estudis del Màster en Informàtica, sempre i això es documenti clarament i no suposi plagi en la pràctica.
Per tant, en presentar una pràctica que faci ús de recursos aliens, s'ha de presentar juntament amb ella un document en què es detallin tots ells, especificant el nom de cada recurs, el seu autor, el lloc on es va obtenir i el seu estatus legal: si l'obra està protegida pel copyright o s'acull a alguna altra llicència d'ús (Creative Commons, llicència GNU, GPL ...). L'estudiant haurà d'assegurar-se que la llicència que sigui no impedeix específicament seu ús en el marc de la pràctica. En cas de no trobar la informació corresponent haurà d'assumir que l'obra està protegida pel copyright.
Hauran, a més, adjuntar els fitxers originals quan les obres utilitzades siguin digitals, i el seu codi font si correspon.