

Enunciats dels projectes de PROP **Quadrimestre de primavera, curs 06/07**

Els projectes d'aquest quadrimestre, en conjunt, constitueixen un petit laboratori del que s'anomena "Mineria de Dades" (en anglès, *Data Mining*). En el procés de mineria de dades es tracta de descobrir patrons o relacions en un conjunt de dades donat.

Per als tres projectes, les dades vindran en fitxers de text. Cada línia del fitxer s'anomenarà registre, i contindrà els valors de diversos atributs. Poden haver-hi atributs numèrics, booleans o categòrics (ex: blau, verd, vermell).

Enunciat 1: Regles d'associació

Una regla d'associació és una implicació entre atributs de l'estil

SI (practica_esport = fals) I (fruita_en_dieta < 10%) I (colesterol_en_sang > 3g/L)
LLAVORS malaltia_cardiaca = cert

Aquí, "practica_esport" i "malaltia_cardiaca" són atributs booleans mentre que "fruita_en_dieta" i "colesterol_en_sang" són atributs numèrics.

Es tracta de fer una eina que trobi les regles d'associació que en les dades superen certs llindars de rellevància (% de registres on l'antecedent de la regla és cert) i de fiabilitat (% d'aquests registres on, a més, el conseqüent és cert). Per simplificar, només es demana que l'eina funcioni amb atributs booleans.

Enunciat 2: Xarxes Neuronals

Les xarxes neuronals (*neural networks*) són un de tants mecanismes per calcular un resultat numèric a partir d'altres resultats numèrics. En particular, poden usar-se per intentar *predir* el valor d'un atribut a partir dels valors d'altres atributs numèrics. Hi ha algorismes per *entrenar* una xarxa neuronal (ajustar els seus paràmetres) perquè faci aquesta predicció amb més i més precisió.

Es tracta de fer una eina per entrenar (mitjançant algun algorisme de tipus *backpropagation*) xarxes neuronals per predir un atribut a partir d'altres atributs.

Enunciat 3: Arbres de decisió

Un arbre de decisió (*decision tree*) és un mecanisme per intentar predir el valor d'un atribut a partir dels valors dels altres atributs. En un arbre de decisió, els nodes interns estan etiquetats amb preguntes sobre alguns dels atributs (exemple: "la renda mensual és alta, mitjana o baixa?", o "feina_fixa?") i tenen tants fills com resultats possibles de la comprovació. Cada fulla conté la predicció que es fa per a aquells registres que compleixen totes les condicions que duen a aquella fulla (per exemple: "concedir el crèdit / denegar el crèdit / estudiar personalment").

Es tracta de fer una eina que ajudi a inferir arbres de decisió per predir el valor d'un dels atributs a partir dels altres.

Comentaris per a tots tres enunciats:

1. Per a tots tres enunciats, s'explicaran a classe les idees principals dels algorismes que cal aplicar per resoldre'ls, o bé es proporcionaran referències adequades.
2. En tots tres casos, seria bo que l'usuari pogués triar entre diversos nivells d'interactivitat: des del cas que el programa ho fa gairebé tot automàticament fins aquell en què l'usuari pot establir manualment tots els paràmetres de l'algorisme, seguir el procés pas a pas, modificar el resultat final, etc.
3. Es valorarà que el programa ofereixi un entorn tant còmode com sigui possible per al necessari preprocés de les dades:
 1. Es podria incloure opcionalment la inspecció visual de les dades, eliminació de registres escollits (p.ex., *outliers* o redundants), eliminació d'atributs (que podrien ser, p.ex., poc rellevants), afegir atributs nous (p.ex., calculats o derivats dels altres), etc.
 2. S'ha de incloure obligatòriament la discretització o booleanització d'atributs numèrics, o la “numerització” en el cas de la pràctica 2.
 3. Es podria opcionalment tractar el cas dels *missing values* (emplenar valors d'atributs que puguin faltar).
 4. El programa ha de donar l'opció de que l'usuari pugui guardar aquestes dades preprocessades per no haver de repetir el preprocés en treballar de nou amb el mateix joc de dades, així com, evidentment, guardar i recuperar el resultat de l'algorisme principal (regles d'associació trobades, xarxa neuronal entrenada, arbre inferit).

Dates dels lliuraments:

Primer: divendres, 16 de març

Segon: divendres, 20 d'abril (especificació de classes compartides: 13 d'abril; acceptació de classes compartides: 2 de maig)

Tercer: divendres, 25 de maig (lliuraments interactius: setmana del 28 de maig)