

Pràctica 2 d'IA1: Classificació automàtica basada en característiques.

Objectiu: Aquesta pràctica té un doble objectiu

- 1) Mostrar la representació de coneixement basada en característiques.
- 2) Introduir-nos en els fonaments de l'aprenentatge no supervisat mitjançant la implementació de l'algorisme de núvols dinàmics.

Fonaments teòrics:

En qualsevol problema en que es tractin objectes del món real, es requereix una representació computacional dels mateixos o del coneixement associat a ells. Molts cops no podem definir una correspondència directa entre l'objecte en si i la representació del mateix, ja sigui perquè aquest és de difícil descripció o perquè les valoracions que es fan sobre ell són subjectives. En aquests casos, es pot intentar buscar una sèrie de característiques per tal de representar-lo.

Per exemple, si volem descriure una lletra escrita, en funció d'una sèrie de característiques podríem escollir:

- amplada
- àrea ocupada
- creuaments de línies

Amb aquestes característiques podríem distingir perfectament algunes lletres entre elles com per exemple: A té 2 creuaments, E té 1 creuament, I no en té cap, O no en té cap, U no en té cap. La I té menor amplada que la resta. En aquest cas, intentariem representar una lletra com a un punt en l'espai de 3 dimensions on cadascun dels eixos correspon a cada característica.

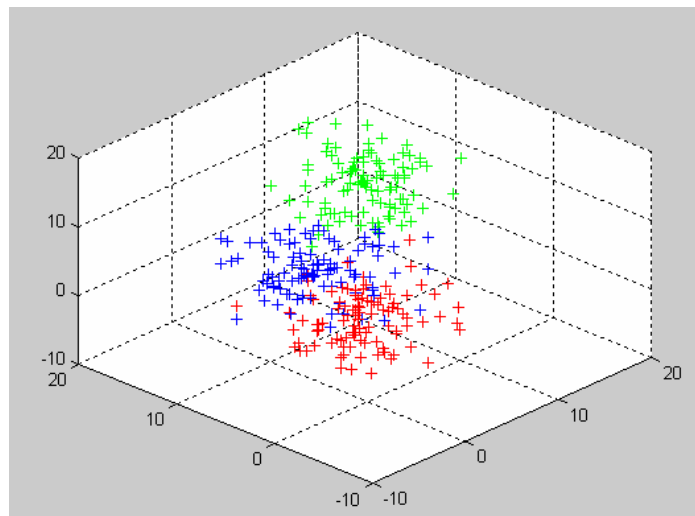


fig.1 Exemple de prendre diferents mostres de 3 classes diferents. Cada punt és la representació d'una mostra concreta descrita segons les seves característiques.

Observem que el fet d'escollir bé les característiques és fonamental i és un dels grans problemes en el món de la classificació i l'aprenentatge. En el nostre exemple les característiques escollides no són suficientment representatives i discriminants per poder distingir les vocals entre elles, de fet l'àrea ocupada ens dóna molt poca informació addicional per tal de distingir-les.

Aquesta pràctica tracta sobre *clustering o aprenentatge no supervisat*. L'aprenentatge és el pas previ a la classificació. Aprenem a partir d'un *conjunt d'aprenentatge* (mostres que fem servir per entrenar). La classificació és podria definir com el procés d'assignació de classe, on cada classe representa un objecte o entitat que s'ha de distingir de la resta.

Aprenentatge no supervisat ens indica que no incorporem informació *a priori* sobre quina classe pertany cadascuna de les mostres del conjunt d'aprenentatge. Això el que ens diu és que tenim un núvol de punts que representen diverses classes però que nosaltres no tenim o no volem introduir coneixement sobre aquests punts i es desitja que sigui el mètode el que decideixi on es troben les classes localitzades.

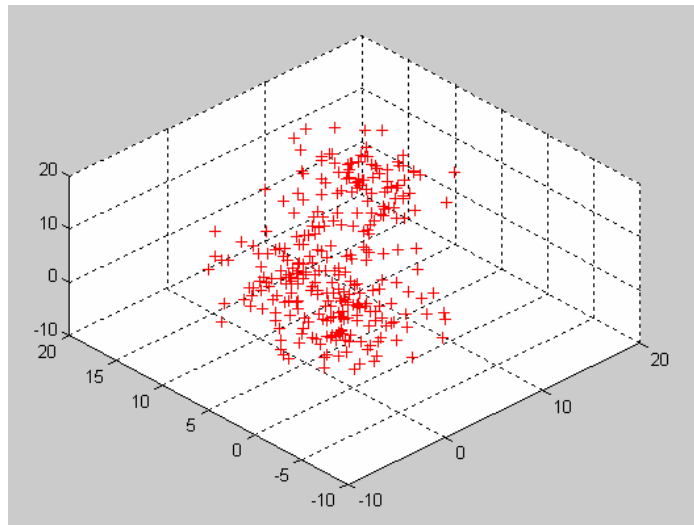


fig 2. El mateix cas d'abans però sense incorporar informació sobre quina mostra pertany a cada classe. Deixarem que sigui el procés d'aprenentatge que decideixi que pertany a cada classe

El procés d'aprenentatge tindrà la missió de realitzar una partició de l'espai.

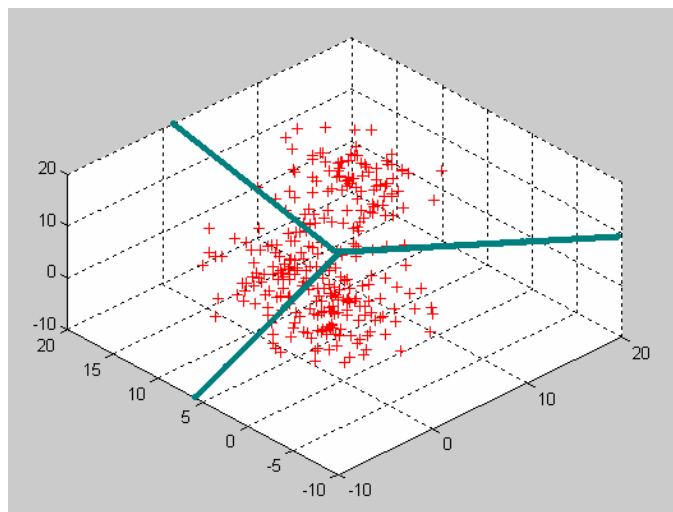


fig 3. Partició de l'espai associat a cada classe.

Un dels algorismes d'aprenentatge no supervisat és el de **núvols dinàmics**. Aquest algorisme es basa en raons purament geomètriques per trobar la millor partició de l'espai. Partim de k representants (tants representants com classes desitjades), aquests representants indicaran el centre del núvol d'una determinada classe.

Pseudocodi:

```

Begin initialize  $k, M_1, M_2, \dots, M_k$ 
      Do    {
            Classify samples according to nearest  $M_i$ 
            Recompute  $M_i$ 
          }
      Until no change in  $M_i$ 
      Return  $M_1, \dots, M_k$ 
End

```

on M_i és el representant i -èssim.

L'algorisme és fonamentalment iteratiu i necessita de la definició d'algun tipus de distància per tal de classificar segons el **més proper** i de alguna forma de **recalcular** els representants. Per exemple, podem fer servir la distància euclídea, manhattan, ... i per tal de recalculer els representants podríem fer la mitja de les mostres de les classes trobades en la classificació.

La inicialització del procés acostuma a fer-se agafant els k primers punts del conjunt.

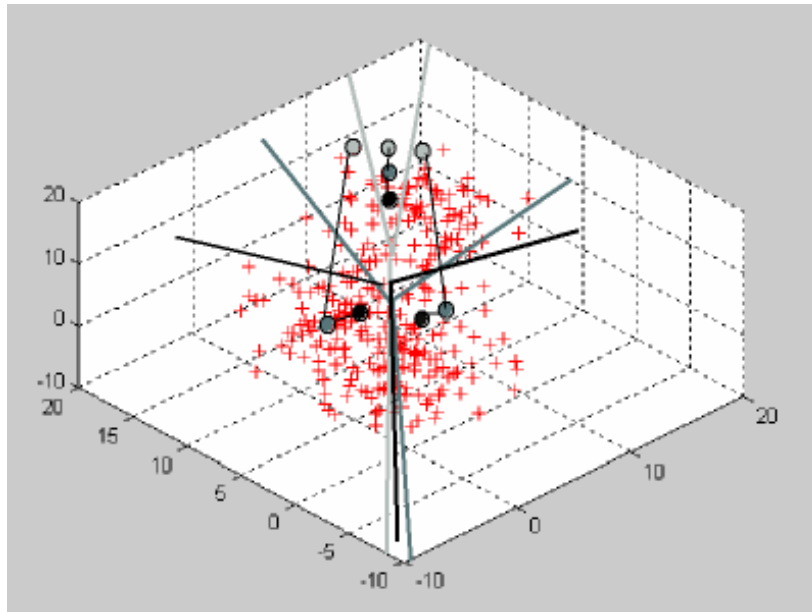


fig 4. Exemple de l'evolució del k-means.

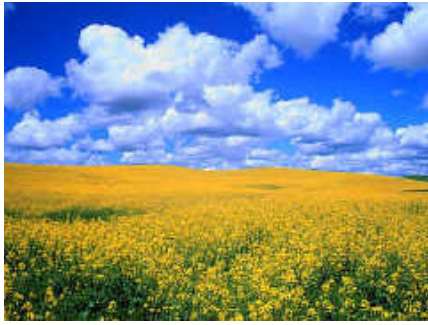
Enunciat de la pràctica:

Aplicarem el coneixement obtingut per realitzar un filtratge clàssic de color en una imatge. Volem representar una imatge amb un cert nombre fix de colors, els que més s'adeqüin a la imatge.

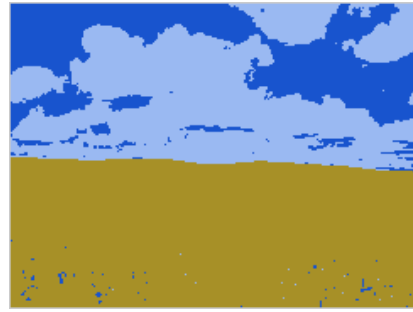
Primer de tot cal representar el coneixement en funció d'unes característiques, en aquest cas, volem representar colors. Una representació clàssica a la que ja estem força habituats es la representació per RGB (quantitat de vermell, quantitat de verd i quantitat de blau); podriem fer-ho escollint altres representacions, com per exemple LHS, LUV.

El problema que ens trobem és que no sabem a priori quins poden ser aquests colors que representin la imatge, tot i que si sabem quans colors volem que hi apareguin. Aquest es un clar exemple de la necessitat d'aprenentatge o classificació no supervisats.

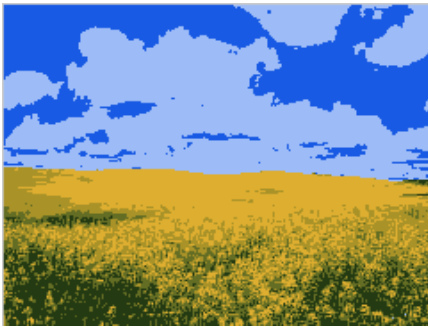
Mostrem a continuació uns exemples



(a)

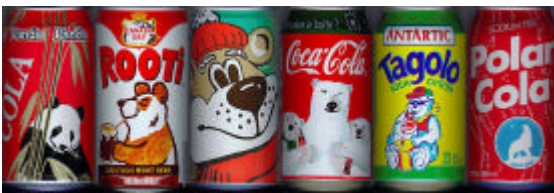


(b)



(c)

fig. 5 (a) Imatge original. (b) Imatge amb 3 classes, cadascuna de les classes ha estat representada amb el seu representant. (c) Imatge amb 6 classes.



(a)



(b)

fig. 6 (a) Imatge original. (b) Imatge amb 6 classes. Observem que la imatge amb 6 classes no representa bé els colors groc i verd, això és un efecte de l'espai de característiques RGB. Observem que no sempre els representants que ens semblen perceptualment més correctes són els que ens trobarà l'algorisme, ja que aquest només atén a distribucions geomètriques del conjunt de test.

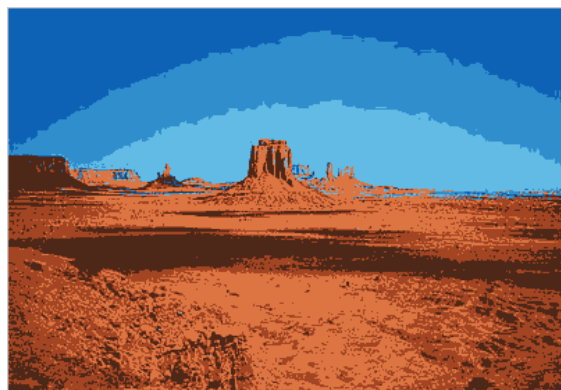


fig 7. Més exemples de *clustering* amb 6 classes (dreta)

Què cal fer a la pràctica?

a.- Implementeu l'algorisme de núvols dinàmics en LISP.

Format:

Nuvols-dinamics (n_clusters dades_entrada)

El format de les dades d'entrada és :

((<dada_multidimensional_1>) (<dada_multidimensional_2>) ...
...(<dada_multidimensional_N>))

El format de la sortida és :

((<representant_1>) ... (<representant_k>))

b.- Cal adjuntar una funció que donats els representants i les dades de test, classifiqui aquestes indicant si pertany a la classe 1 (estan més aprop del representant_1), ...