Pràctica opcional de Haskell: algorisme *k* Nearest Neighbors

Presentació i objectius

L'algorisme del *k* Nearest Neightbors (*k*NN) o *k* veïns més propers és un dels algorismes bàsics de l'aprenentatge automàtic o del reconeixement de patrons. Aquest algorisme ens permet abordar problemes de classificació i de regressió.

La classificació és una de les tasques de reconeixements de patrons en la que volem etiquetar un individu a partir de certes propietats que el caracteritzen; entenen com a individu una entitat de qualsevol tipus. A continuació es mostra l'exemple que haureu de tractar en aquesta pràctica:

- Classificació de flors (Font: problema 'iris' del repositori de la UCI (Frank y Asunción, 2010)). Suposem que volem realitzar una aplicació per a un magatzem de flors, on arriben a diari milers de productes. Disposem de un sistema làser que ens subministra una sèrie de mesures sobres les flors i ens demanen que el sistema les classifiqui automàticament per a transportar-les mitjançant un robot a les diferents estanteries del magatzem. Les mesures que envia el sistema són: la longitud i amplada del sèpal i el pètal de cada flor.

La Figura 1 mostra un exemple d'aquest conjunt de dades. La primera columna correspon a la classe (el tipus de flor) i la resta als atributs (representació dels objectes). Cada fila correspon a un objecte, també anomenat típicament exemple.

Figura 1: Conjunt d'entrenament

class	sepal-length	sepal-width	petal-length	petal-width
setosa	5,1	3,5	1,4	0,2
setosa	4,9	3,0	1,4	0,2
versicolor	6,1	2,9	4,7	1,4
versicolor	5,6	2,9	3,6	1,3
virginica	7,6	3,0	6,6	2,1
virginica	4,9	2,5	4,5	1,7

Fuente: Problema "iris" del repositorio UCI (Frank y Asunción, 2010).

Un dels algorismes clàssics de classificació és el *k*NN (*k nearest neighbors*). En aquesta pràctica treballarem en la implementació del mateix i la seva aplicació sobre el conjunt de dades anterior.

Per avaluar un algorisme de classificació sobre un problema es pot dividir el conjunt de dades que tenim en dos: conjunts d'entrenament (arxiu adjunt 'iris.train.txt') i test (arxiu adjunt 'iris.test.txt') respectivament. El conjunt d'entrenament s'utilitza com a model de classificació en el *k*NN. El conjunt de test el farem servir per avaluar que tal de bé (o malament) ho fa el classificador.

En aquest algorisme la classificació de nous exemples (o exemples del conjunt de test) es realitza buscant el conjunt dels *k* exemples més propers de entre el conjunt d'entrenament etiquetats prèviament i seleccionant la classe majoritària de entre les seves etiquetes.

Exemple d'aplicació

La Figura 1 mostra un conjunt d'entrenament en el que hem de classificar flors en funció de les seves propietats. En aquest exemple aplicarem el *k*NN amb valors de *k* de 1 i 3, utilitzant la distància euclidea i el mètode de votació bàsic.

Per classificar l'exemple que mostra la Figura 2, hem de calcular les distàncies entre aquest nou exemple i tots els del conjunt d'entrenament (els de la Figura 1). La Figura 3 mostra aquestes distàncies.

Figura 2: Exemple de test

class	sepal-length	sepal-width	petal-length	petal-width
setosa	4,9	3,1	1,5	0,1

Fuente: Problema "iris" del repositorio UCI (Frank y Asunción, 2010).

Figura 3: Distàncies euclidees

0,5	0,2	3,7	2,5	6,1	3,5

Per a l'1NN escollim la classe del més proper que coincideix amb el segon exemple (distància 0,2) que té per classe *setosa*. Per al 3NN escollim els tres exemples més propers: primer, segon i quart; amb distàncies respectives: 0,5, 0,2 i 2,5. Les seves classes corresponen a *setosa*, *setosa* i *versicolor*. En aquest cas assignarem també la classe *setosa* per ser la majoritària. En tots dos casos el resultat és correcte.

Paràmetres del *k*NN

Hem vist que un dels paràmetres de l'algorisme és el nombre de veïns involucrats en el procés. Un segon paràmetre a poder canviar és la funció de distància. En aquesta pràctica treballarem amb les següents:

- Euclidea:

$$d_{euclidea}(x, y) = \sqrt{\sum_{i=1}^{4} (x_i - y_i)^2}$$

- Manhattan:

$$d_{manhattan}(x, y) = \sum_{i=1}^{4} |x_i - y_i|$$

Un tercer paràmetre correspon al mecanisme de votació:

- Simple: totes les distància tenen el mateix pes i cerquem la classe majoritària en el mecanisme de votació (és la que s'ha utilitzat en el exemple anterior)
- Ponderat: en el mecanisme de votació assignem un pes a cada exemple inversament proporcional a la distància involucrada:

$$pes = \frac{1}{distancia}$$

Mesures d'avaluació

Per avaluar la bondat del classificador es solen aplicar mesures d'avaluació comparant les prediccions del classificador amb les classes reals del conjunt de test.

- *Accuracy* o exactitud: dividim les prediccions correctes pel nombre d'exemples del conjunt de test
- *Lost* o error: dividim les prediccions incorrectes pel nombre d'exemples del conjunt de test

Noteu que les dues mesures anteriors són complementàries. És a dir, la suma de les dues ha de ser 1.

Descripció de la pràctica

Se us demana que feu un programa Haskell que contingui almenys el que es descriu en els apartats següents:

1. Declaració d'una estructura de tipus de dades capaç d'emmagatzemar les dades dels arxius d'entrenament i test adjunts pertanyents al problema 'Iris'. Funció per carregar aquests arxius en l'estructura de dades dissenyada.
2. Escriviu les següents funcions de distància, de votació, i d'avaluació:
    a. de distància euclidià,
    b. de distància de Manhattan,
    c. de votació bàsica,
    d. de votació ponderada,
    e. d'avaluació d''accuracy',
    f. d'avaluació de 'lost' o 'error'.
3. Implementeu una funció d'ordre superior per a l'aplicació del kNN que rebi les dades anteriors, la k, la funció de distància i la funció de votació i calculi les prediccions sobre el conjunt de test.

4. Implementeu una funció d'ordre superior que donades les prediccions i les classes del test i una llista de funcions d'avaluació, doni l'aplicació de cadascuna de les funcions d'avaluació per comparar les prediccions i les classes.

Heu de lliurar la vostra solució com un únic fitxer practica.hs amb els comentaris que hi cregueu adients a través del lliurament de pràctiques del Racó.

Recursos

Adjunt a l'enunciat d'aquesta pràctica teniu a la vostra disposició els arxius següents:

- iris.names.txt: descripció de les dades del problema Iris de la UCI
- iris.train.txt: conjunt d'entrenament
- iris.test.txt: conjunt de test