Técnicas de Inteligencia Artificial
Ingeniería Electrónica

Práctica – Aprendizaje Automatizado

1) En el conjunto de datos del ejemplo PlayTennis, calcular la entropía del subconjunto de datos de Play Tennis en el cual Humedad=Hight. Cuál sería la ganancia de información si se considera esa propiedad como primer nodo?

2) Considerando el siguiente ejemplo de los Simpson (F Sancho Caparrini, Univ de Sevilla)

Nombre	Long Pelo	Peso	Edad	**Género**
Homero	0”	250	36	**H**
Marge	10”	150	34	**M**
Bart	2”	90	10	**H**
Lisa	6”	78	8	**M**
Maggie	4”	20	1	**M**
Abe	1”	170	70	**H**
Selma	8”	160	41	**M**
Otto	10”	180	48	**H**
Crusty	6”	200	45	**H**
Comic	8”	290	38	**???**

Puede desarrollar un árbol de decisión que utilice sólo dos variables para determinar el género de un personaje en ese contexto? Qué valores de corte propondría para esas dos variables?
Resolver en forma intuitiva primero y luego fundamentar con ganancia de información.

3) Diseñar una función en Matlab

`[E_prom, E_min, k, tr] = miCvLoss(t)`, donde:

`t` es un árbol de clasificación
`E_prom` es el error promedio por validación cruzada para 10-folds,
`E_min` es el error mínimo por validación cruzada para 10-folds,
`k` es la partición que obtuvo el mínimo error y
`tr` es el mejor árbol obtenido entre los 10.

4) a) Obtener del dataset `espirales.mat` un conjunto de 1200 datos para entrenamiento y un conjunto de 300 datos para test.

b) Utilizando sólo el conjunto de entrenamiento del ítem anterior, conseguir los 4 mejores árboles usando 200, 300, 400 y 600 datos (6-fold, 4-fold, 3-fold y 2-fold respectivamente).

c) Clasificar los datos de test del ítem a con el mejor árbol del ítem b y graficar las espirales anidadas obtenidas. Ayuda:

```
rho = 0:.01:1;
theta1 = 0:(2*pi/100):2*pi;
theta2 = -pi:(2*pi/100):pi;
polar(theta1, rho);
hold on;
polar(theta2, rho);
```

```
resultado = best(X);
gscatter(X(:, 1), X(:, 2), resultado, 'rb', 'oo')
```

5) a) Construir árboles de clasificación para el dataset `ionosphere` para luego realizar una gráfica del error de validación (`cvLoss`) en función del mínimo número de observaciones por hojas (`MinLeaf`).
b) Considerando el árbol de clasificación del conjunto `ionosphere`, realizar una gráfica del error de validación en función de los niveles de poda de `0` a `8`.

6) El dataset `puntos.mat` está formado por las coordenadas `x,y` de puntos del plano cercanos al origen. El conjunto de puntos está clasificado en cuatro sectores determinados por dos rectas.

a) Cargar el conjunto de datos y observar la gráfica mediante los siguientes comandos (el archivo `puntos.mat` debe estar almacenado en el directorio actual de Matlab):

```
>> load puntos
>> gscatter(puntos(1,:), puntos(2,:), targetsPuntos)
```

b) Obtener un árbol de decisión para el dataset `puntos` utilizando Statistics Toolbox – Classification Trees de Matlab (se deben transponer las matrices de entrada y se debe utilizar alguna estrategia para conseguir un árbol de clasificación y no uno de regresión). Interpretar el resultado.

c) Calcular la proporción de clasificados correctamente sobre el mismo conjunto de datos utilizado para generar el árbol. Investigar si este resultado se puede mejorar mediante una poda de nivel 1.

d) Particionar el conjunto de entrada para poder calcular la proporción de clasificados correctamente sobre un conjunto de datos distinto del utilizado para generar el árbol (conjuntos de entrenamiento y test).

e) Repetir el ítem anterior para otro tamaño de conjuntos de entrenamiento y test.

f) Graficar la clasificación obtenida.

7) Adaptar el problema anterior para puntos en el espacio (tres dimensiones). Generar valores aleatorios para la componente z de cada punto (z en el [-5, 5]) y adaptar la matriz de targets mediante una función. El espacio se divide ahora en ocho sectores según el signo de la tercera componente.

Ayuda:

```
puntos3D = vertcat(puntos, 10 * rand(1, 400) - 5);
targetsPuntos3D = targetsPuntos;
N = length(puntos3D); % 400
for i = 1:N
    if puntos3D(3,i) < 0, targetsPuntos3D(1, i) = targetsPuntos(1, i) + 4; end
end
```

8) i) Calcular la salida de una neurona de dos entradas (vector P), con pesos W, umbral θ y función transferencia *f* en cada caso:

a) P = (1, 0), W = (1, 2), θ = 1, f lineal (purelin).
b) P = (1, 1), W = (-2, 1), θ = 2, f escalón (hardlim).
c) P = (1, 0), W = (1, 2), θ = 1, f sigmoide (logsig).

ii) Verificar los resultados con Matlab:

- Mediante una sucesión de comandos.

- Ejecutando `nnd2n2`.

9) a) Generar el vector ordenado `x` con 50 valores en el `[0, 5]` (conjunto de entrenamiento) aleatoriamente con distribución uniforme. Obtener $y = sen^3(x)$. Graficar la función $y = f(x)$.

b) Entrenar una red neuronal con 20 neuronas en la capa intermedia para fitear la función.

c) Generar 50 nuevos valores para x (conjunto de test), aleatoriamente con distribución uniforme y obtener el resultado de simular la red obtenida para dichos valores.

d) Realizar una gráfica donde se comparen ambas curvas.

e) Ensayar estrategias para obtener mejores resultados.

10) Obtener redes neuronales para las espirales anidadas que se describen en el dataset `espirales.mat` tomando 600, 900, 1200 y 1500 datos y variando la cantidad de neuronas en la capa intermedia en 2, 6 y 10. Realizar un cuadro con los errores obtenidos en cada caso y analizar los resultados.

11) El dataset `heladas.mat` contiene datos reales, correspondientes a observaciones meteorológicas en la zona de Zavalla (50 Km de Rosario). El objetivo es predecir la ocurrencia o no de una helada en la mañana siguiente. Los datos disponibles son mediciones de variables meteorológicas en distintos horarios. La matriz `testHeladas` contiene datos reales, pero la columna correspondiente a la clase se debe predecir.

Descripción de los datos:

Clases: 1 hay helada, 0 no
Atributos (continuos, normalizados en escala): mínima de ayer, mínima de anteayer, t bulbo húmedo 20hs ayer, t bulbo seco 20hs ayer, nubosidad, día del año.
Aplicar las herramientas para aprendizaje automatizado (AD y RNA) a dicho dataset tratando de mejorar los resultados mediante las posibles variaciones. Realizar un análisis comparativo de dichos resultados obtenidos.