UNIVERSIDADE DA CORUÑA

ICBio 20/21

P1&P2

Autores: Sergio Rodríguez Nieto
David Maseda Neira
Fecha:*Coruña,a 6 de abril de 2021*

Índice

1. Introducción

En esta memoria incremental se presentan el procedimiento y los resultados de las prácticas en Inteligencia Computacional para la Bioinformática.

2. Descripción de los *datasets*

2.1. *Breast Cancer Wisconsin Dataset*

Consiste en 683 ejemplos, con 10 características por ejemplo:

- Radio
- Textura
- Perímetro
- Área
- Suavidad
- Compacidad
- Concavidad
- Puntos de concavidad
- Simetría
- Dimensión fractal

Los ejemplos se clasifican como malignos(1) o benignos(2).

2.2. *Iris Flower Dataset*

Consiste en 150 ejemplos, 50 para cada una de las 3 clases (Setosa, Versicolor y Virginica) Contiene 4 características por ejemplo:

- Longitud del sépalo (cm)
- Anchura del sépalo (cm)
- Longitud del pétalo (cm)
- Anchura del pétalo (cm)

3. Preparación del *dataset*

3.1. Preparación del *Breast Cancer Wisconsin Dataset*

Los datos del *Breast Cancer Wisconsin Dataset* se obtienen en un CSV. Durante esta práctica, no se realiza extracción de características, por lo que la carga de los datos del CSV se realiza directamente. Estos datos se guardan en un fichero MAT de matlab, para no volver a parsear el csv en posteriores ejecuciones. De igual modo, de realizarse preprocesado, este solo sería necesario la primera vez, al construir el dataset.

3.2. Preparación del *Iris Flower Dataset*

De igual manera que con el *dataset* anterior, no se realiza preprocesado, y los datos se guardan en un fichero de matlab para posterior acceso.

4. Metodología y desarrollo

El código está organizado de la siguiente manera:
Existen 3 módulos:

- lineal.m
- quadratic.m
- Tree.m

Cada uno de ellos define una función, que construye y entrena el modelo según los datos proporcionados. La firma de las funciones es la siguiente:

```
#quadratic.m
quadratic(dataset) -> [trainACC,testACC]

#linear.m
linear(dataset) -> [trainACC,testACC]

#Tree.m
Tree(dataset,PredictorNames,ResponseName,
MinLeafSize = 1,MinParentSize = 10,
debug = false) -> [trainACC,testACC]
```

El parámetro debug en *Tree.m* sirve para ocultar las figuras de la estructura de los árboles, por simplicidad al ejecutar. Si se quiere consultar la estructura de los árboles, se fija el parámetro debug a true.

Para cada uno de los modelos, la función que define ejecuta secuencialmente lo siguiente:

1. Carga los datos del dataset provisto en la firma de la función.
2. Configura los parámetros para el modelo según lo definido en la firma de la función.
3. Entrena el modelo y lo valida con el *split* de test.
4. Muestra las métricas por clase y globales.
5. La función devuelve el accuracy, tanto en train como en test.

4.1. Ejecución

Se provee un archivo main.m, que construye y ejecuta todos los modelos para ambos *datasets*. **IMPORTANTE**: Se ha utilizado el bloque **arguments** de MATLAB, que sólamente está disponible de la versión 2019b en adelante, por lo cual si se ejecuta en una versión anterior, muestra un error de sintaxis.

5. Modelos y entrenamiento

Se prueban 5 modelos diferentes: Un discriminante lineal, un discriminante cuadrático y 3 variaciones de árboles de clasificación, con variaciones en los parámetros. Todos los modelos se entrenan con un *10-fold*. Se recogen las siguientes métricas de rendimiento:

- Recall
- Precisión
- Especificidad
- VPN
- Accuracy
- F1

En esta sección, se presentan los modelos, sus parámetros, y los resultados para el conjunto de test.

5.1. Discriminante lineal

	Setosa	Versicolor	Virginica
Recall	1.00	0.96	0.98
Precision	1.00	0.98	0.96
Specificity	1.00	0.96	0.98
VPN	1.00	0.98	0.99
Accuracy	1.00	0.98	0.98
F1	1.00	0.97	0.97

Cuadro 1: Iris(lineal)

	Maligno	Benigno
Recall	0.98	0.92
Precision	0.96	0.97
Specificity	0.98	0.92
VPN	0.97	0.97
Accuracy	0.96	0.96
F1	0.97	0.94

Cuadro 2: Cáncer(lineal)

5.2. Discriminante cuadrático

	Setosa	Versicolor	Virginica
Recall	1.00	0.92	0.98
Precision	1.00	0.98	0.94
Specificity	1.00	0.92	0.98
VPN	1.00	0.97	0.99
Accuracy	1.00	0.97	0.97
F1	1.00	0.94	0.96

Cuadro 3: Iris(Quad)

	Maligno	Benigno
Recall	0.94	0.98
Precision	0.99	0.90
Specificity	0.94	0.98
VPN	0.90	0.99
Accuracy	0.95	0.95
F1	0.96	0.94

Cuadro 4: Cáncer(Quad)

5.3. Árbol 1

Árbol de búsqueda con los parámetros por defecto:

- MinLeafSize: 1
- MinParentSize: 10

	Setosa	Versicolor	Virginica
Recall	1.00	0.90	0.92
Precision	1.00	0.93	0.91
Specificity	1.00	0.90	0.92
VPN	1.00	0.95	0.96
Accuracy	1.00	0.94	0.94
F1	1.00	0.91	0.91

Cuadro 5: Iris(Tree 1)

	Maligno	Benigno
Recall	0.96	0.91
Precision	0.95	0.92
Specificity	0.96	0.91
VPN	0.92	0.95
Accuracy	0.94	0.94
F1	0.95	0.91

Cuadro 6: Cáncer(Tree 1)

5.4. Árbol 2

Árbol de búsqueda con los parámetros:

- MinLeafSize: 5
- MinParentSize: 10

	Setosa	Versicolor	Virginica
Recall	1.00	0.94	0.88
Precision	1.00	0.90	0.94
Specificity	1.00	0.95	0.89
VPN	1.00	0.97	0.95
Accuracy	1.00	0.94	0.94
F1	1.00	0.92	0.90

Cuadro 7: Iris(Tree 2)

	Maligno	Benigno
Recall	0.96	0.92
Precision	0.96	0.93
Specificity	0.96	0.92
VPN	0.93	0.96
Accuracy	0.95	0.95
F1	0.96	0.92

Cuadro 8: Cáncer(Tree 2)

5.5. Árbol 3

Árbol de búsqueda con los parámetros:

- MinLeafSize: 2
- MinParentSize: 5

	Setosa	Versicolor	Virginica
Recall	1.00	0.94	0.90
Precision	1.00	0.92	0.94
Specificity	1.00	0.94	0.90
VPN	1.00	0.97	0.95
Accuracy	1.00	0.95	0.95
F1	1.00	0.92	0.92

Cuadro 9: Iris(Tree 3)

	Maligno	Benigno
Recall	0.97	0.92
Precision	0.96	0.95
Specificity	0.97	0.92
VPN	0.95	0.96
Accuracy	0.95	0.95
F1	0.97	0.93

Cuadro 10: Cáncer(Tree 3)

6. Comparativa: Medidas globales

Se utiliza la accuracy en test como métrica para comparar los modelos.

	Accuracy(Train)	Accuracy(Test)
Discriminante Lineal	0.9574	0.9605
Discriminante Cuadrático	0.9579	0.9501
Árbol 1	0.9754	0.9384
Árbol 2	0.9711	0.9546
Árbol 3	0.9793	0.9414

Cuadro 11: Medidas globales de accuracy en training y test(Cáncer)

	Accuracy(Train)	Accuracy(Test)
Discriminante Lineal	0.9847	0.9822
Discriminante Cuadrático	0.9857	0.9822
Árbol 1	0.9867	0.9689
Árbol 2	0.9798	0.9733
Árbol 3	0.9847	0.9778

Cuadro 12: Medidas globales de accuracy en training y test(Iris)

7. Significancia estadística - Boxplots

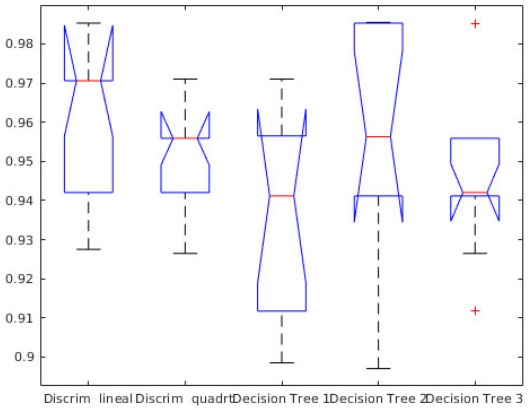

Figura 1: Cáncer

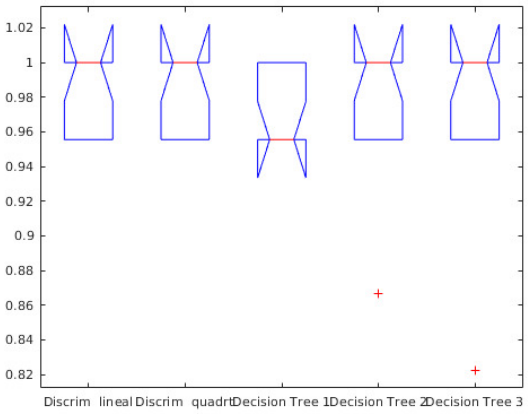

Figura 2: Iris

Índice de cuadros