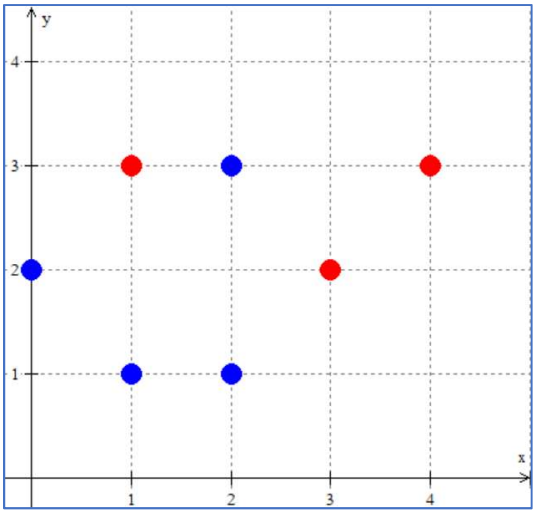
# Tarea Sistemas Inteligentes

## Árboles de Decisión

- La siguiente Tarea se puede hacer de manera individual o en equipos de máximo 3 integrantes.
- En esta Tarea se requieren los siguientes libros que puedes accesar de la página de los autores:
    - **Texto A**: The Elements of Statistical Learning : Data Mining, Inference, and Prediction. 2nd Edition. Hastie/Tibshirani/Friedman. Springer. 2017. https://web.stanford.edu/~hastie/ElemStatLearn/
    - **Texto B**: An Introduction to Statistical Learning with applications in R. 2014. Hastie/Tibshirani/James/Witten. Springer. http://faculty.marshall.usc.edu/gareth-james/ISL/

1. Del Texto B, lee la sección 8.1, pp. 303-316, y posteriormente, de la sección de ejercicios de dicho capítulo (Sección 8.4, página 332), responde los siguientes incisos:
    a. Ejercicio 1, pág. 332.
    b. Ejercicio 4, pp. 332 y 333.

2. Este ejercicio lo debes resolver con papel y lápiz. En las páginas 309 y 310 del Texto A y las páginas 311 y 312 del Texto B, se definen, para el caso de un árbol de clasificación, la Tasa del Error de Clasificación, el Índice Gini y la Entropía-Cruzada. Usaremos estas tres métricas y la siguiente figura para este ejercicio:




Supongamos que tenemos las coordenadas del conjunto de 7 puntos de entrenamiento con dos clases de puntos, rojos (R) y azules (A) como se indica en la figura. Supongamos que deseamos aplicar el método de árbol de decisiones para futuras clasificaciones de nuevos datos. Para ello, deseamos saber cómo nos

conviene empezar a construir este árbol de clasificación binario. Supongamos que se tienen las siguientes dos propuestas:

Propuesta 1: iniciar el nodo raíz del árbol de clasificación con el criterio $X < 2.5$.
Propuesta 2: iniciar el nodo raíz del árbol de clasificación con el criterio $Y > 2.5$.

Indica con cuál de las dos propuestas nos conviene iniciar si utilizamos:

a. La métrica del error de clasificación.
b. El índice Gini.
c. La entropía cruzada.

3. Para el siguiente ejercicio usaremos la base de datos de la UCI: German Credit Data Set: http://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data) . En particular trabajaremos con el archivo "german.data".
    a. A partir del conjunto de datos inicial, realiza un preprocesamiento de los datos categóricos: verifica que cada variable tiene al menos el 10% de datos en cada uno de sus niveles, de no ser así, agrúpalos de la mejor manera para que se cumpla dicho criterio. Finalmente, asigna los nombres a la variable y los niveles resultantes, siguiendo la información de la página de la UCI.
    b. Realiza un preprocesamiento análogo, ahora para las variables numéricas. En particular, para este ejercicio, aplicaremos el criterio de eliminar todos aquellos datos en los que alguna de las variables continuas tenga valores extremos menores o mayores a 2.5*IQR de Q1 y Q3, respectivamente. Es decir, Q1-2.5*IQR y Q3+2.5*IQR.
    c. Realiza una partición del conjunto de datos resultante después del preprocesamiento realizado en los incisos anteriores, en el conjunto de Entrenamiento y el de Prueba. Indica y justifica la partición que utilices.
    d. Obtener el árbol de decisión para el conjunto de entrenamiento generado y usando la métrica del índice Gini como criterio para dividir el árbol en cada nodo. NOTA: Utiliza la ayuda ?rpart en R para revisar el argumento "params" sobre como utilizar "Gini" o "Ganancia de Información".
    e. Obtener la matriz de confusión del modelo obtenido.
    f. Repite los incisos "d" y "e", pero usando ahora el criterio de Ganancia de Información. NOTA: En particular la librería rpart en R utiliza la métrica de Ganancia de Información que vimos en clase, pero que está basada precisamente en la Entropía-Cruzada que se define en los textos A y B. R llama simplemente "information" a esta métrica.