Práctica 3: Microarray de expresión genética

Conjunto de datos: AML-ALL-normalized

- AML: tipo de leucemia
- Clases: 2
- Atributos: 7130
- Instancias: 72

- Estimar, mediante validación cruzada de 10 particiones, el error de las hipótesis generadas por los siguientes clasificadores:
  - J48, NB, IBK1, Regresión Logística, MLP (H10), SVM(lineal)

Selección de atributos: filtro

- Seleccionar sucesivamente 4, 8, 16 y 32 atributos mediante los siguientes métodos de filtro
  - incertidumbre simétrica, ReliefF, eliminación recursiva con SVM y CFsubsetEval (correlación atributos-atributos/clase)

- Examinar los atributos seleccionados por cada método.

- Estimar, mediante validación cruzada de 10 particiones, el error de las hipótesis generadas por los siguientes clasificadores:
  - J48, NB, IBK1, Regresión Logística, MLP (H10), SVM(lineal)

Discusión de resultados

- Para cada uno de los algoritmos utilizados, crear una tabla de dos entradas (número de atributos y método de selección) anotando la tasa de error estimada

- Discutir los resultados, primero para cada tabla y luego en conjunto.

Selección de atributos: envoltorio

- Utilizar un método de envoltorio y selección hacia adelante para seleccionar los atributos con los siguientes métodos
  - J48, NB, IBK1, Regresión Logística, MLP (H10), SVM(lineal)

- Examinar la selección de atributos para cada algoritmo

- Estimar, mediante validación cruzada de 10 particiones, el error de las hipótesis generadas por los siguientes clasificadores:
  - J48, NB, IBK1, Regresión Logística, MLP (H10), SVM(lineal)

Discusión de resultados

- Crear una tabla que permita discutir fácilmente los resultados obtenidos por el método de envoltorio.

- Comparar también los resultados con los métodos de filtro, tanto las tasas de error obtenidas como los atributos seleccionados

PCA

- Sea M1 el método que obtiene una menor tasa de error entre los métodos anteriores
- Denominar *n* al número de atributos seleccionado por el método M1
- Intentar realizar un análisis de componentes principales del conjunto de datos original
  - Si el tiempo de cómputo supera las 24 horas**, no** realice los dos siguientes ejercicios
- Entrenar el algoritmo utilizado por M1 con los *n* primeros componentes principales. Estimar la tasa de error mediante validación cruzada de 10 particiones
- Entrenar el algoritmo utilizado por M1 con los 2*n* primeros componentes principales. Estimar la tasa de error mediante validación cruzada de 10 particiones
- Discutir los resultados

Conclusiones

- Resumir, en menos de 300 palabras, las conclusiones obtenidas.