Análisis de un conjunto de datos de origen biológico mediante técnicas de *machine learning* supervisadas y no supervisadas

Objetivos

El objetivo de esta actividad es implementar de forma razonada técnicas de aprendizaje supervisado y no supervisado para el análisis de un conjunto de datos de origen biológico.

Pautas de elaboración

Para esta actividad necesitaréis un ordenador con [R instalado](https://cran.r-project.org/bin/windows/base/) y los datos provistos por el profesor, que consisten en:

* Classes: contiene las clases de cada muestra.
* Column\_names: contiene los nombres de cada columna (genes)
* Gene\_expression: valores de expresión de cada gen.

Actividades

* Preparación el entorno de trabajo (instalación y carga de paquetes adecuados).
* Depurado del conjunto de datos:
  + Generar un solo *dataframe* cuyas columnas se correspondan con el nombre del gen, el nombre de las filas con el ID *(identification)* de cada muestra y donde se incluya la columna con la clase de cada muestra.
  + Imputación de los datos NA.
  + Otros métodos de procesamiento de las muestras escogidos por el estudiante después explorar los datos.
* Implementación de cuatro métodos de aprendizaje no supervisado: dos de reducción de dimensionalidad y dos de clusterización.
* Implementación de tres métodos de aprendizaje supervisado + cálculo de diferentes métricas de evaluación: matriz de confusión, precisión, sensibilidad, especificidad y score-F1.
* Preguntas de respuesta corta.

Extensión y formato

Los resultados deberán ser entregados como un *script* de R.

No hay una extensión máxima del *script.*

El código utilizado ha de estar comentado de forma que su interpretación sea sencilla.

**Preguntas sobre las actividades:**

1. Procesamiento de los datos (0,5 puntos):

¿Qué método habéis escogido para llevar a cabo la imputación de los datos? Razonad vuestra respuesta. (0,3 puntos).

¿Habéis llevado a cabo algún otro tipo de procesamiento? Razonad vuestra respuesta. (0,2 puntos).

1. Métodos no supervisados (1 punto):

¿Cuál es el motivo por el cual habéis seleccionado estas técnicas de reducción de dimensionalidad? (0,3 puntos).

¿Cuál es el motivo por el cual habéis seleccionado estas técnicas de clusterización? (0,3 puntos).

En ambos casos, ¿qué aspectos positivos y negativos tienen cada una? (0,2 puntos).

En el caso de la clusterización, ¿podéis afirmar con certeza que los clústeres generados son los mejores posibles? Razonad vuestra respuesta. (0,2 puntos).

1. Métodos supervisados (1,75 puntos):

¿Cuál es el motivo por el cual habéis seleccionado ambas técnicas de aprendizaje supervisado? ¿Cuál ha dado mejores resultados a la hora de clasificar las muestras? Razonad vuestra respuesta (1 punto).

\**Borrador:*

*Hemos elegido, por un lado, un algoritmo de análisis de la discriminante, concretamente el RDA (análisis de la discriminante regularizada) > es un método con buen rendimiento para datos con relaciones lineales y además, al ser regularizado, incluye parámetros para evitar el ajuste.*

*Hemos elegido SVM porque también tiene buen rendimiento y, además, como el RDA había salido bien, hemos optado por la forma lineal (parece que los datos si se pueden separar mediante funciones lineales).*

*Finalmente, hemos elegido Random Forest porque es una técnica más robusta que las anteriores, ya que combina el poder predictivo de muchos modelos individuales (árboles de decisión). En este caso también serviría para datos con relaciones no lineales, para ver si la precisión mejora o no.*

¿Habéis considerado oportuno implementar algún método de reducción de dimensionalidad para procesar los datos antes de implementarlos en dichas técnicas? ¿Por qué? (0,5 puntos).

*\*Borrador:*

*En vez de eso, hemos pensado que lo mejor era aplicar un modelo de regularización, concretamente LASSO, ya que teníamos muchos genes y pensamos que lo mejor para estos algoritmos era trabajar solo con los más relevantes para la clasificación.*

¿Qué aspectos positivos y negativos tienen cada una de las técnicas que habéis escogido? (0,25 puntos).

*\*Borrador:*

*-RDA: (+)buen rendimiento para relaciones lineales/evita sobreajuste. (-)asume distribución normal de los datos, no sirve para clasificación según patrones no lineales*

*-SVM: (+) efectivo en espacios de alta dimensión/tiene funciones lineales y tipo kenrel para distintos tipos de datos (lineales y no lineales) (-) coste computacional/elegir un kernel adecuado y el valor de C apropiado/*

*-Random Forest (+)combina el poder predictivo de modelos individuales para generar un modelo final mas robusto/menos sensible a datos atípicos (-) coste computacional, potencial sobreajuste/falta de interpretabilidad*

1. De estas cuatro opciones, ¿qué tipo de arquitectura de *deep learning* sería la más adecuada para procesar datos de expresión génica? Razonad vuestra respuesta (0,25 puntos).

a) Red de perceptrones *(multiperceptron layers).*

b) Redes convolucionales.

c) Redes recurrentes.

d) Redes de grafos.

Rúbrica

|  |  |  |  |
| --- | --- | --- | --- |
| Análisis de un conjunto de datos de origen biológico mediante técnicas de *machine learning* supervisadas y no supervisadas | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | Preparación del conjunto de datos | 1,5 | 15 % |
| Criterio 2 | Implementación de cuatro técnicas de aprendizaje no supervisado | 2 | 20 % |
| Criterio 3 | Implementación de tres técnicas de aprendizaje supervisado + evaluación del rendimiento de cada modelo | 2 | 20 % |
| Criterio 4 | Preguntas sobre las actividades | 3,5 | 35 % |
| Criterio 5 | Orden y claridad del código | 0,5 | 5 % |
| Criterio 6 | Creatividad: se muestra un enfoque innovador o una solución creativa para resolver el problema planteado | 0,5 | 5 % |
|  |  | **10** | **100 %** |