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).

Por un lado, decidimos usar un método de análisis de la discriminante, concretamente el RDA (análisis de la discriminante regularizada). Este método lo que hace es, a partir de un conjunto de variables predictoras, generar una función discriminante, que permite clasificar las observaciones según las clases establecidas por la variable categórica (variable dependiente). Concretamente, el RDA es un método muy adecuado cuando los datos tienen relaciones lineales y se quiere evitar el sobreajuste, ya que tiene dos parámetros de regularización (gamma y lambda) que penaliza las variables con colinealidad.

También decidimos usar el método Support Vector Machine (SVM). En este caso, para la clasificación, este algoritmo se basa en el cálculo de hiperplanos, que permiten la separación entre clases. Estos hiperplanos pueden constituirse a partir de funciones lineales o funciones de tipo kernel. Como el RDA daba una buena precisión, consideramos que los datos se podían separar con funciones lineales y por ello decidimos usar para el SVM su forma lineal.

Finalmente, usamos Random Forest, que es una técnica más robusta que las anteriores, ya que combina el poder predictivo de muchos modelos individuales (árboles de decisión). Consideramos apropiado comparar la capacidad predictiva de modelos más simples (RDA y SVM) con un modelo más complejo como el Random Forest. Además, Random Forest también contempla relaciones no lineales de los datos, un aspecto que no tuvimos en cuenta con los dos modelos anteriores.

Los resultados fueron para los tres algoritmos, una precisión bastante elevada (0,99 para el RDA, 0,84 para el SVM y 0,99 para el Random Forest). A parte de la precisión, también calculamos el área bajo la curva. Aunque realmente no se podían generar ni curvas ROC ni curvas PR ya que se trataba de un problema multiclase, usamos la función multiclass.roc(), que calcula una curva ROC para cada nivel de la variable categórica y luego hace un promedio. Esta función no permite obtener un resultado gráfico, pero sí un valor de área bajo la curva (AUC) promedio. Para los tres modelos, este área estaba muy próxima a 1, por lo que los tres son muy eficientes a la hora de hacer la clasificación.

¿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).

Consideramos tanto la posibilidad de usar los resultados de los algoritmos de reducción de la dimensionalidad como pre-procesado para los algoritmos de aprendizaje supervisado, como el uso de métodos de regularización (LASSO, RIDGE, Elastic Net). Llegamos a la conclusión de, en muchos casos, los resultados de los métodos de aprendizaje no supervisado son difícilmente interpretables, ya que, en muchos casos, al proyectar los datos de un espacio de mayor dimensión a uno de menor dimensión, el resultado son un conjunto de variables, diferentes a las originales, y difíciles de comprender y extraer un significado biológico. Por ello, en su lugar, decidimos usar un método de regularización, para quedarnos solo con un subconjunto de todas las variables originales. De entre los métodos de regularización que existen, decidimos usar LASSO, puesto que aplica una penalización más intensa y con ella conseguimos seleccionar 104 variables, con respecto a los 500 originales, reduciendo considerablemente la dimensionalidad y quedándonos solo con las variables más relevantes a la hora de hacer la clasificación.

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

|  |  |  |
| --- | --- | --- |
| **Modelo** | **Ventajas** | **Inconvenientes** |
| **RDA** | -Buen rendimiento para datos con relaciones lineales.  -Previene sobreajuste y problemas de multicolinealidad. | -Asume una distribución normal de los datos.  -No es adecuado para problemas de clasificación con patrones no lineales. |
| **SVM** | -Efectivo en espacios de alta dimensión.  -Admite funciones lineales y funciones tipo kernel. | -Coste computacional.  -Puede ser difícil elegir la función óptima para cada caso.  -Puede ser difícil determinar el hiperparámetro C más adecuado. |
| **Random Forest** | -Combina el poder predictivo de modelos individuales para generar un modelo final más robusto.  -Menos sensible a la presencia de 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 %** |