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).**

Para asegurar la calidad y completitud de los datos antes del análisis, se revisaron los datos para identificar la presencia de valores nulos (NA) y valores infinitos. El proceso incluyó las siguientes verificaciones:

**Identificación de valores nulos:** Utilizamos la función any(is.na(datos\_para\_limpiar)) para verificar si había valores nulos en el conjunto de datos. En este caso específico, se encontró que no había valores nulos ya que la función devolvió FALSE. Esto indica que todos los valores presentes en el conjunto de datos eran no nulos.

**Identificación de valores infinitos**: De manera similar, se revisaron los datos para detectar valores infinitos utilizando sapply(datos\_para\_limpiar, function(x) sum(is.infinite(x))). La suma total de valores infinitos se calculó y se verificó que no había valores infinitos presentes en el conjunto de datos.

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

Sí, se llevaron a cabo varios tipos adicionales de procesamiento para preparar los datos:

Transformación de datos: Se convirtieron los valores que estaban almacenados como texto a datos numéricos utilizando la función parse\_number. Esto fue necesario para asegurar que todas las columnas contuvieran datos numéricos que pudieran ser utilizados en el análisis estadístico y los modelos de machine learning.

Normalización: Se escalaron los datos para que cada columna tuviera una media de aproximadamente 0 y una desviación estándar de aproximadamente 1 utilizando la función scale. La normalización es un paso crucial, especialmente cuando se aplican técnicas de reducción de dimensionalidad y algoritmos de machine learning que son sensibles a las escalas de los datos.

Eliminación de columnas irrelevantes: Se eliminaron las columnas muestra y clase del conjunto de datos principal (datos\_para\_limpiar) para facilitar el análisis centrado en las características relevantes. Posteriormente, se volvieron a añadir estas columnas después de la normalización y otros preprocesamientos necesarios.

1. **Métodos no supervisados (1 punto):**

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

Seleccionamos PCA porque es una técnica robusta y ampliamente utilizada para reducir la dimensionalidad de los datos. PCA proyecta los datos en un espacio de menor dimensión, capturando la mayor parte de la variabilidad presente en los datos originales.

Elegimos MDS porque es una técnica efectiva para visualizar relaciones entre datos en un espacio de menor dimensión, preservando tanto como sea posible las distancias euclídeas originales.

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

Seleccionamos K-Means por su simplicidad en la agrupación de datos. K-Means es una técnica popular para particionar datos en un número predefinido de clusters, minimizando la variación dentro de cada cluster. Además, es fácil de interpretar y visualizar.

Elegimos DBSCAN porque es una técnica de clusterización basada en densidad que puede identificar clusters de cualquier forma y manejar el ruido en los datos, además esta técnica es ideal cuando los datos tienen una estructura compleja.

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

**PCA:**

Positivos: Reduce la dimensionalidad y mejora la interpretabilidad.

Negativos: No siempre capta la estructura no lineal de los datos.

**MDS:**

Positivos: Útil para preservar distancias originales.

Negativos: Puede ser computacionalmente costoso para grandes datasets.

**K-Means:**

Positivos: Rápido y fácil de implementar.

Negativos: Asume clusters esféricos y necesita especificar K.

**DBSCAN:**

Positivos: Identifica clusters arbitrarios y maneja ruido.

Negativos: La elección de parámetros eps y minPts puede ser sensible y afectada por la densidad de los datos.

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

No podemos afirmar con certeza que los clusters generados son los mejores posibles porque la calidad de los clusters depende de los parámetros seleccionados (como el número de clusters en K-Means o los valores de eps y minPts en DBSCAN). Cambios en estos parámetros pueden producir diferentes resultados.

Por otro lado, diferentes técnicas de clusterización pueden producir resultados diferentes. Por lo tanto, es importante probar varios métodos y configuraciones para comparar y determinar cuál proporciona los clusters más representativos y útiles para el análisis específico.

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).**

Seleccionamos SVM, SVM Gaussiano, Random Forest y Naive Bayes debido a sus diferentes enfoques para manejar datos complejos:

**SVM**: Es eficaz para clasificar datos con alta precisión al encontrar el hiperplano óptimo que separa las clases. Es robusto frente a problemas de sobreajuste.

**SVM Gaussiano**: Extiende SVM con un kernel radial, permitiendo manejar relaciones no lineales en los datos.

**Random Forest**: Es robusto frente al ruido y evita el sobreajuste al combinar múltiples árboles de decisión. También proporciona una medida de importancia de variables.

**Naive Bayes**: Es simple y rápido, adecuado para datos con una distribución normal, y funciona bien con grandes conjuntos de datos.

**¿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).**

Sí, implementamos PCA para reducir la dimensionalidad antes de aplicar las técnicas de machine learning supervisado. Esto ayuda a eliminar características redundantes, reduce el ruido, mejora la eficiencia computacional y previene el sobreajuste, permitiendo a los modelos generalizar mejor.

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

**SVM:**

Positivos: Tiene alta precisión y es robusto frente al sobreajuste.

Negativos: Computacionalmente es intensivo y sensible a la elección de parámetros.

**SVM Gaussiano:**

Positivos: Maneja relaciones no lineales y ofrece una alta precisión en datos complejos.

Negativos: Es más complejo de ajustar en comparación con SVM y también computacionalmente más intensivo.

**Random Forest:**

Positivos: Maneja bien datos con ruido, evita el sobreajuste y proporciona importancia de variables.

Negativos: Menos interpretativo y más demandante computacionalmente.

**Naive Bayes:**

Positivos: Es simple, rápido i funciona bien con grandes conjuntos de datos.

Negativos: Asume independencia entre características y por lo tanto, es menos preciso en datos complejos no normalmente distribuidos.

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.

Las redes recurrentes, ya que son ideales para datos de expresión génica porque pueden manejar secuencias y capturar dependencias temporales. Los datos de expresión génica tienen una naturaleza secuencial donde la expresión de un gen puede influir en otros. Las redes recurrentes son capaces de recordar información previa en la secuencia, lo que es crucial para entender estas relaciones complejas. A diferencia de las redes de perceptrones y convolucionales, que no pueden manejar dependencias a largo plazo, las redes recurrentes están diseñadas específicamente para este tipo de tareas, haciéndolas la mejor opción para analizar datos de expresión génica.

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 %** |