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).

Mótivo de Elección del PCA

He seleccionado el Análisis de Componentes Principales (PCA) buscando una combinacion lineal de las variables objeto de estudio maximizando la varianza de los datos, y que minimice la correlacion entre ellas. Además esta técnica se ordena en función de su varianza, es decir, que se ordena en funcion de cuánto se separan o desvían los valores individuales de un conjunto respecto a su media. La PCA para este estudio ha logrado clasificar 5 grupos de genes que contribuyen significativamente a la varianza global, sin embargo, en este caso particular, las dos primeras componentes principales (PC1 y PC2) explican solo el 29.43% de la variabilidad total, lo que indica que no capturan una proporción suficiente de la varianza para identificar patrones claros. Se necesitaría 155 componentes para poder explicar el 90% de los datos de conjunto. Esto se refleja en las nubes de puntos generadas, que muestran una superposición significativa entre las clases, dificultando la separación entre grupos, salvo en el caso de AGH, que parece estar mejor definido.

Mótivo de Elección del t-SNE

Elegí el t-SNE (t-Distributed Stochastic Neighbor Embedding) porque es muy eficaz para reducir la dimensionalidad de los datos y, al mismo tiempo, mantener las relaciones locales entre los puntos. En el caso de los genes en cáncer de mama, las interacciones son bastante complejas y no siempre siguen un patrón lineal.

Los resultado muestran una separacuion bastabnte buena entre los 4 tipos de canceres, bien delimitado y bien separados entre ellos. Estos resultados indican que cada tipo de cáncer tiene un patrón único de expresión génica, como si cada uno tuviera su "huella digital molecular". Esto refleja que los procesos y vías biológicas que los caracterizan son diferentes entre sí, como el modo en que crecen, se desarrollan o interactúan con su entorno. La separación clara entre los grupos sugiere que estos genes podrían servir para entender mejor lo que hace único a cada tipo de cáncer.

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

Clustering Jerárquico (Método DIANA)

El código implementa el clustering divisivo utilizando el algoritmo DIANA con dos métricas de distancia distintas: euclidiana y Manhattan. En la primera línea, se utiliza la distancia euclidiana, que mide la "distancia directa" entre puntos en un espacio multidimensional. En la segunda línea, se utiliza la distancia Manhattan, que calcula las diferencias absolutas entre las coordenadas de los puntos en cada dimensión.

Mótivo de Elección K-means

K-means permite dividir los datos en un número específico de grupos (k), lo que es útil cuando se tiene una idea previa de la cantidad de clústeres esperados. Es eficiente para conjuntos de datos grandes y produce particiones claras, lo que facilita el análisis.

K-means es fácil de entender y rápido de ejecutar, incluso con grandes conjuntos de datos. Esto lo convierte en una opción popular para explorar grandes volúmenes de datos de expresión génica.

- A diferencia del clustering jerárquico, K-means permite predefinir el número de clústeres a formar. Esto puede ser útil cuando se tiene una idea previa del número de subgrupos que se desean identificar.

- El algoritmo es eficiente desde el punto de vista computacional, lo que permite aplicarlo en grandes conjuntos de datos sin que se vuelva un cuello de botella en el proceso.

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

**Metodo Diana**

Aspectos positivos del clustering jerárquico (DIANA):

- A diferencia de métodos como K-means, no es necesario predefinir el número de grupos, lo que lo hace más flexible.

- La representación gráfica mediante dendrogramas facilita la interpretación de los resultados, ya que muestra cómo los clústeres se agrupan a medida que se dividen.

Aspectos negativos del clustering jerárquico (DIANA):

- Este método puede ser computacionalmente costoso para grandes conjuntos de datos, lo que puede dificultar su aplicación en estudios de expresión génica a gran escala.

- El clustering jerárquico puede verse afectado por valores atípicos o ruido en los datos, lo que puede hacer que los clústeres no reflejen correctamente las relaciones biológicas subyacentes.

- Aunque no se debe predefinir el número de clústeres, puede ser complicado determinar el nivel adecuado de corte en el dendrograma para obtener una partición óptima de los clústeres.

Aspectos positivos del clustering no jerárquico (K-means):

- K-means es fácil de entender y rápido de ejecutar, incluso con grandes conjuntos de datos. Esto lo convierte en una opción popular para explorar grandes volúmenes de datos de expresión génica.

- A diferencia del clustering jerárquico, K-means permite predefinir el número de clústeres a formar. Esto puede ser útil cuando se tiene una idea previa del número de subgrupos que se desean identificar.

- El algoritmo es eficiente desde el punto de vista computacional, lo que permite aplicarlo en grandes conjuntos de datos sin que se vuelva un cuello de botella en el proceso.

Limitaciones clustering no jerárquico (K-means):

- K-means requiere que el número de clústeres se especifique de antemano. Elegir el valor correcto para K puede ser complicado y no siempre refleja la estructura subyacente de los datos.

- K-means es sensible a los valores atípicos (outliers), que pueden distorsionar los centroides y afectar los resultados del agrupamiento.

- K-means funciona mejor cuando los clústeres son de forma esférica y tienen varianzas similares. Si los clústeres tienen formas complejas o distribuciones muy diferentes, los resultados pueden no ser satisfactorios.

**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 se puede afirmar con certeza que los clústeres generados sean los mejores posibles, ya que la calidad de la clusterización depende de varios factores, como el algoritmo utilizado, los parámetros seleccionados (por ejemplo, el número de clústeres en k-means) y las características inherentes de los datos. Además, algunos algoritmos, como k-means, asumen que los clústeres son esféricos y de tamaño similar, lo cual puede no ajustarse a la distribución real de los datos. Aunque se han utilizan métricas como el método del codo o el índice de silueta para evaluar la calidad, estas no garantizan que los resultados sean óptimos ni que representen la estructura subyacente de los datos de manera perfecta.

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 %** |