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

**EQUIPO 2 LOTE 10**

Gálvez Robleño, Carlos

Rayo Morales, Mario Andrés

Benítez Fernández, Sergio

Salas Ramos, Maria

Encarnación Alejandro, Freyda Luisa

**Preguntas sobre las actividades:**

1. Procesamiento de los datos:

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

El método seleccionado para la imputación de los datos es el siguiente:

* **Imputación por la mediana** para las variables numéricas, que ayuda a evitar distorsiones causadas por valores extremos.
* **Imputación por la moda** para las variables categóricas.

¿Habéis llevado a cabo algún otro tipo de procesamiento? Razonad vuestra respuesta.

Sí, se han llevado a cabo varios otros procesos de procesamiento:

* **Eliminación de filas no relevantes** (como "sample\_0").
* **Escalado de las variables numéricas** para garantizar que todas tengan la misma importancia en los modelos.
* **Selección de características** utilizando Lasso para reducir la dimensionalidad y seleccionar las variables más relevantes.
* **División de los datos** en entrenamiento y prueba para evaluar el rendimiento de los modelos.
* **Entrenamiento y evaluación de modelos supervisados**.
* **Evaluación de modelos no supervisados.**

1. Métodos no supervisados:

¿Cuál es el motivo por el cual habéis seleccionado estas técnicas de reducción de dimensionalidad?

* ***Técnica t-SNE:***

Elegimos el algoritmo T-distributed stochastic Neighbor Embedding (t-SNE) debido a que es más eficiente que otras técnicas en el manejo de datos no lineales. Esto se debe a que identifica muy bien las relaciones complejas entre las variables, preservando la localidad de los datos y revelando parte de la estructura global en el mapeo a baja dimensionalidad, siendo eficiente para identificar y estructurar datos. Aunque es costoso computacionalmente, para el tamaño muestral de nuestros datos resulta eficiente. Además, las 2 primeras dimensiones explicaban la mayor parte de la varianza por lo que el número de dimensiones no era un factor que fuese a afectar en su ejecución.

¿Cuál es el motivo por el cual habéis seleccionado estas técnicas de clusterización?

* ***Técnica: Ward***

Se trata de un método de clusterización que permite identificar patrones complejos y agrupar los datos en función de su semejanza (ejemplo: agrupación de pacientes con síntomas similares, respuestas similares a tratamientos, expresión similar de genes, etc). De esta forma genera clústeres homogéneos, facilitando el análisis y la interpretación de los datos, y, por tanto, minimiza la varianza dentro de los clústeres, lo que asegura que los datos dentro de los clústeres sean lo más similares posible.

En ambos casos, ¿qué aspectos positivos y negativos tienen cada una?

* ***Técnica: t-SNE***

Entre los aspectos positivos, identifica muy bien las **relaciones complejas entre las variables**, **preserva la localidad de los datos** y r**evela parte de la estructura global** en el mapeo a baja dimensionalidad, además de ser eficiente para identificar y estructurar datos. Entre sus aspectos negativos se encuentran su **comportamiento estocástico** (tiene naturaleza aleatoria, esto significa que los resultados pueden variar cada vez que se ejecuta el algoritmo), lo que lo hace no determinista (es necesario utilizar semillas, *seed*, para que nos reproduzca el mismo resultado); y **su complejidad cuadrática** en tiempo y espacio, que puede hacerlo muy lento con muestras grandes (aunque se puede usar multiprocesamiento). Además, generalmente solo se puede utilizar hasta una tercera dimensión, y más allá de esto, los datos pueden no quedar claros y las implementaciones del algoritmo pueden no permitir más dimensiones.

* ***Técnica Ward:***

Entre sus aspectos positivos están **la creación de grupos compactos** ya que tiende a crear clústeres de tamaño uniforme y consistentes, lo que puede ser útil en muchas aplicaciones (como en el manejo de datos multivariantes, comunes en estudios de salud donde se recopilan múltiples variables por paciente); la **minimización de Varianza dentro de los clústeres**, lo que asegura que los puntos dentro de un mismo clúster sean lo más similares posible**;** y **su capacidad de manejar datos complejos**, lo cual lo hace especialmente idóneo para disciplinas como la biología. Entre sus aspectos negativos se encuentran que **computacionalmente es complejo** y por tanto costoso, por lo que no es lo más eficiente para grandes conjuntos de datos, haciendo que el proceso sea lento; es **sensible a la escala de los datos**, por lo que es importante normalizar los datos antes de aplicar el algoritmo. Además, es **no determinista**, por lo que lo que los resultados pueden variar con diferentes ejecuciones (es necesario utilizar semillas, *seed*, para que nos reproduzca el mismo resultado).

|  |  |  |
| --- | --- | --- |
| **MÉTODOS** | **PROS** | **CONTRAS** |
| **ISOMAP** | Reduce dimensionalidad no lineal. | Costoso computacionalmente, sensible. |
| **T-SNE** | Bueno para visualización, captura relaciones locales. | Lento, difícil de interpretar, no adecuado para clusters grandes |
| **K-MEANS** | Simple, eficiente, escalable, fácil de entender | Requiere predefinir k, sensible a inicialización y valores atípicos |
| **WARD** | No requiere predefinir k, maneja clusters irregulares | Computacionalmente costoso, sensible al ruido y menos escalable |

**Tabla 1.** Tabla resumen de los principales aspectos positivos y negativos de los métodos de análisis no supervisado escogidos.

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

* ***Técnica: Ward***

Consideramos que sí, debido a que con una k=5 ha generado 5 clústeres muy definidos, en los cuales ninguno de ellos está sobrerrepresentado (lo que indicaría una mala segregación o discriminación de los datos, al incluirse en su mayoría dentro del mismo clúster) y tampoco infrarrepresentado (lo que podría estar indicando la generación de un grupo artificial). Por tanto, los 5 clústeres generados parecen consistentes.

1. Métodos supervisados:

¿Cuál es el motivo por el cual habéis seleccionado ambas técnicas de aprendizaje supervisado?

Las tres técnicas de aprendizaje supervisado que hemos decidido aplicar son el SVM lineal, el RDA y el LDA debido a que consideramos que serían efectivos para clasificar datos de expresión génica, algunos de estos motivos son:

* El SVM lineal es apropiado para problemas con alta dimensión de datos, como la expresión genética, pues busca un hiperplano ideal que diferencie las clases de forma eficaz.
* El LDA es un método que se fundamenta en disminuir la dimensionalidad y potenciar la separabilidad entre clases, lo que la convierte en eficaz para datos de buena distribución.
* El RDA es una versión normalizada de LDA que modifica la matriz de covarianza con el fin de optimizar la clasificación cuando los datos poseen una estructura complicada, que podría ofrecer información más precisa.

¿Cuál ha dado mejores resultados a la hora de clasificar las muestras? Razonad vuestra respuesta.

Al analizar los 3 distintos métodos, lo primero que se ve es que las matrices de confusión del RDA y LDA son idénticas, ambas con un 99.37% de precisión, y mostrando la misma sensibilidad y especificidad en todas las clases. Esto indica que, al final la regularización adicional del RDA no ha aportado ninguna mejora significativa con respecto al LDA. Por ello, solo compararemos entre el SVM lineal y el conjunto (LDA/RDA).

A la hora de clasificar datos con el conjunto de datos utilizado, podemos ver que el SVM lineal y los métodos LDA/RDA sí han presentado diferencias, y podemos comparar los resultados para ver cuál ha rendido más a la hora de clasificar las muestras de expresión génica.

El SVM lineal tuvo una precisión del 90.57% y un coeficiente Kappa de 0.8736, lo que indica una buena capacidad de clasificación, pero inferior a los otros dos métodos. La sensibilidad obtenida entre clases es variada, con valores bajos en CHC (0.7037) y HPB (0.5333), lo que significa que tuvo dificultades para identificar correctamente estas categorías. Por otro lado, LDA/RDA dieron un 99.37% de precisión y un Kappa de 0.9917, lo que refleja una clasificación casi perfecta. Además, mostraron sensibilidad y especificidad de 1.0 en la mayoría de las clases, asegurando que las muestras fueran correctamente clasificadas sin falsos positivos o negativos.

¿Habéis considerado oportuno implementar algún método de reducción de dimensionalidad para procesar los datos antes de implementarlos en dichas técnicas? ¿Por qué?

Para mejorar el rendimiento de la posterior implementación de métodos supervisados, se ha utilizado el método LASSO, el cual permite reducir la dimensionalidad de la base de datos.

Se ha implementado este método de reducción de dimensionalidad al obtenerse una base de datos, posterior al procesamiento de los datos iniciales, con una alta dimensionalidad (800 observaciones y 500 variables).

Este gran número de variables, con un número de observaciones insuficientemente mayor, puede afectar la eficacia y dificultar la implementación y aplicación de las diferentes técnicas de aprendizaje utilizadas. Por lo tanto, el método LASSO permite reducir el número de variables (en nuestro caso a 100 variables) regularizando y reduciendo la influencia de variables irrelevantes al seleccionar las variables con mayor relevancia.

¿Qué aspectos positivos y negativos tienen cada una de las técnicas que habéis escogido?

* ***Técnica: RDA***

RDA es adecuado cuando tienes un número grande de genes en relación con el número de muestras, y cuando los datos pueden tener multicolinealidad o distribución no normal.

**Aspectos positivos:** Maneja bien la alta dimensionalidad y la colinealidad, previene sobre ajuste.

**Aspectos negativos:** Requiere ajuste de hiperparámetros, interpretación compleja.

* ***Técnica: LDA***

LDA puede ser útil si tus clases están bien separadas y los datos siguen distribuciones normales. Es más sencillo de interpretar y rápido de ejecutar, pero limitado si los datos son complejos.

**Aspectos positivos:** Simple y fácil de interpretar, eficiente computacionalmente, buen rendimiento con separaciones lineales.

**Aspectos negativos:** Supone normalidad y covarianza homogénea, limitado en separaciones no lineales.

* ***Técnica: SVM Lineal***

SVM lineal es útil si tus datos tienen muchas características (genes) y es probable que no sean linealmente separables, pero puede ser más costoso computacionalmente y más difícil de interpretar.

**Aspectos positivos:** Bueno para alta dimensionalidad, robusto frente a sobreajuste, puede manejar bien problemas no lineales.

**Aspectos negativos:** Requiere ajuste de parámetros, sensible al ruido, computacionalmente costoso.

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.

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

b) Redes convolucionales.

c) Redes recurrentes.

d) Redes de grafos.

Las redes de perceptrones multicapa (MLP, por sus siglas en inglés) son una de las arquitecturas de redes neuronales más comunes y efectivas para tareas de clasificación en datos tabulares, como los datos de expresión génica. A continuación, te explico por qué esta arquitectura es la más adecuada:

**Estructura tabular de los datos:** Los datos de expresión génica generalmente tienen una estructura tabular, donde cada fila representa una muestra (paciente, célula, etc.) y cada columna representa la expresión de un gen. Este tipo de datos es ideal para redes de perceptrones multicapa, ya que estas redes están diseñadas para trabajar con entradas estructuradas de esta manera (vectores de características).

**Capacidad de modelar interacciones complejas:** Las redes MLP son muy efectivas para capturar relaciones no lineales entre las variables, lo cual es crucial en el análisis de expresión génica. Los genes pueden interactuar de manera compleja, y un MLP tiene la capacidad de aprender esas interacciones sin la necesidad de que el usuario las especifique explícitamente.

**Flexibilidad:** Las redes MLP tienen una gran flexibilidad en términos de arquitectura (número de capas, número de neuronas por capa) y se pueden adaptar fácilmente a diferentes problemas. En el análisis de expresión génica, donde los conjuntos de

datos pueden variar mucho en tamaño y complejidad, las redes MLP pueden ajustarse para modelar de manera efectiva.

**Conclusión:**

Las redes de perceptrones multicapa son la opción más adecuada para procesar datos de expresión génica, ya que están diseñadas para trabajar con datos tabulares y tienen la capacidad de modelar interacciones complejas entre las variables (genes), lo cual es fundamental en este tipo de análisis.