Aplicación de técnicas de aprendizaje no supervisado sobre datos biológicos

Objetivos

El objetivo de esta actividad es determinar de forma justificada aquellos algoritmos de aprendizaje no supervisado más eficientes a la hora de detectar patrones en el conjunto de datos propuesto.

Pautas de elaboración

Para esta actividad necesitarás un ordenador con [R instalado](https://sissa.crc-sas.org/wp-content/uploads/2020/10/Instalacion_R.html#instalar-r-en-windows) y el conjunto de datos de RNA-seq de cáncer de mama.

Actividad

Buscar patrones en los datos mediante técnicas de aprendizaje no supervisadotales como los métodos de reducción de dimensionalidad. Esta actividad consta de las siguientes partes:

* Preparación del entorno de trabajo (instalación y carga de paquetes adecuados).
* Implementación de cuatro técnicas de aprendizaje no supervisado a escoger por el estudiante.
* Selección razonada del algoritmo y sus parámetros más eficientes.

Extensión y formato

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

Realiza los comentarios pertinentes para explicar el código.

No hay una extensión máxima del *script.*

Preguntas sobre las actividades

1. Ambiente y librerías de trabajo (0,5 puntos):

¿Qué librerías han sido necesarias para llevar a cabo la implementación de cada uno de los métodos de aprendizaje no supervisado? Explica el rol de los argumentos de las funciones que has utilizado en cada método de reducción de dimensionalidad.

Los argumentos de las funciones quedan explicados en el Rmarkdown, siendo este documento una síntesis del original.

En cuanto a las librerías utilizadas tenemos:

* **Paquete dimRed**: paquete que ofrece **varias técnicas de reducción de dimensión** (incluyendo MDS, Isomap, PCA) de forma unificada y flexible.
* **Paquete tSNE**: nos da acceso a la librería ocn el mismo nombre.
* **Librería ggfortify** amplía ggplot2 (cargada con "library ( ggplot2)") para visualizar los resultados de los modelos de forma automática.
* **Librería stats** crea funciones estadísticas de R (cmdscale () o prcomp ()...).
* **RDRToolbox** **paquete** y **librería**: son algoritmos necesarios de isomap para reducción de dimensiones. Este a su vez se instala desde **Bioconductor**.

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

¿Qué problemas has detectado en el conjunto de datos que te han sido proporcionados? ¿Cómo los has solucionado?

La base de datos era de una dimensión muy elevada, por lo que en numerosas ocasiones he tenido que filtrar los datos de manera aleatoria o aplicar reducciones para poder proceder con los estudios.

1. Métodos no supervisados (2,5 puntos):

¿Cuál es el motivo por el cual has seleccionado estas técnicas de reducción de dimensionalidad? (1 punto).

Son las más sencillas o básicas y con las que más cómoda me sentía para indagar en ellas y poder compararlas. Además, entre ellas se representa bastante bien la diversidad de tipos de datos (lineales, no lineales...) por lo que da una buena visión, al menos preliminar de lo que podrían ser los resultados. No quería arriesgarme a que por la complejidad que me suponía el lenguaje, los resultados no fueran concluyentes.

¿Qué aspectos positivos y negativos tienen cada una de las técnicas escogidas? (0,5 puntos).

**MODELO ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)**

El modelo PCA trata los datos para **maximizar la varianza** en las primeras dimensiones.

En el gráfico, la separación es bastante buena, parecida a MDS, pero no tan nítida como en Isomap o t-SNE. Hay mezclas entre algunas clases, como BRCA y LUAD que podrían ser causa biológica, o ruido del modelo. La limitación del PCA es que asume que las relaciones son **lineales**, por eso puede perder información si la estructura real es más compleja.

**MODELO MULTIDIMENSIONAL SCALING**

El modelo MDS busca preservar las **distancias** entre los datos.

Gráficamente **l**ogra una separación razonable de los tipos de cáncer, aunque se observa solapamiento, especialmente entre BRCA (rojo) y LUAD (naranja). Sin embargo, no maneja bien estructuras no lineales.

**MODELO ISOMAP**

El Isomap supera a MDS porque preserva distancias más directas y la estructura geodésica del espacio de datos, que se traduce en mayor representación de las relaciones no únicamente lineales.

Gráficamente la separación entre clases es más clara, aunque algunos tipos como PRAD (morado) muestran mayor dispersión interna y algún solapamiento. La ventaja de este modelo es que funciona bien cuando los datos tienen formas curvas o no lineales.

**MODELO DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING (t-SNE)**

El modelo t-SNE se centra en preservar relaciones locales, es decir, que muestras cercanas en alta dimensión sigan estando cerca en el mapa reducido. Visualmente ofrece una separación muy clara entre los tipos de cáncer, sin apenas solapamiento. Entre sus limitaciones, el t-SNE no respeta bien las distancias globales (entre grupos lejanos).

Comenta para cada caso el resultado de representar gráficamente las primeras dos variables de cada uno de los métodos. ¿Qué se observa? ¿Tiene sentido biológico? Razona tu respuesta (1 punto).

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.En el gráfico vemos que los tipos de cáncer BCRA, COAD y LUAD se entremezclan. Esto quiere decir que, aunque son tres tipos de cancer distintos, compartirán mucha expresion génica del mismo tipo de genes. No son iguales porque se diferencian en algunos genes, pero tienen muchos que compraten.

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.Vemos que, BRCA y LUAD pueden compartir ciertas **vías celulares**, **mutaciones** o **patrones de expresión**. Por ello, aparecen más próximos y entremezclados que los grupos de KRIC y PRAD que distan más.

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.Este **gráfico del isomap** muestra bastante dispersión en los grupos KIRC y COAD junto con LUAD. Sin embaro, también están entremezclados PRAD con BRCA. Esto nos llevaría a pensar que estos dos últimos tipos de cáncer sean más parecidos molecularmente (compartan más genes) mientras que los otros se diferencian más. Por otro lado, hay algunos puntos esporádicos de COAD y LUAD que podrían haberse mezclado por errores de procesamiento o toma de daros, así como por algún caso aislado biológicamente hablando.

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.Vemos qué dimensiones representan o reflejan la mayor varianza residual. Aplicando la **regla del codo** cogeríamos desde dos dimensiones o tres, que es donde se estabiliza la variabilidad explicada. En cambio, si cogemos menos de dos dimensiones, hay mucha variabilidad que queda sin explicar. En el segundo gráfico, dos dimensiones explican más información que en el primero, ya que la caída de varianza residual es más rápida.

Ilustración 1. Primer gráfico de varianza explicada.

Ilustración 2 Segundo gráfico de varianza explicada

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.En el **modelo t-SNE** podríamos pensar que cada tipo de cáncer tiene una base molecular distinta puesto que los grupos no están tan entremezclados ni pegados como en los otros análisis.

Rúbrica

|  |  |  |  |
| --- | --- | --- | --- |
| Aplicación de técnicas de aprendizaje no supervisado sobre datos biológicos | 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 | 3 | 30 % |
| Criterio 3 | Preguntas sobre las actividades | 4,5 | 45 % |
| Criterio 4 | Orden y claridad del código | 0,5 | 5 % |
| Criterio 5 | Creatividad: se muestra un enfoque innovador o una solución creativa para resolver el problema planteado | 0,5 | 5 % |
|  |  | **10** | **100 %** |