75.06/95.58 Organización de Datos
Finger 4: Reducción de dimensiones.

Puntos: Se otorgarán los siguientes puntos en orden de entrega 5,3,2,1,1,1.
Tiempo Estimado de Resolución: 3 hs.

Utilizando las herramientas que proveen distintas bibliotecas de Python (Numpy, Sklearn, etc.), se quiere realizar un análisis sobre reducción de dimensiones y visualización de datos. Se utilizará un set de datos sobre el cual se puede analizar la relación entre distintas muestras de tejido en términos de expresión genética. En el siguiente link podrán descargar el set del cual se utilizarán los siguientes dos archivos:

- GTExdata.csv
- SampleLabels.csv

El primer archivo contiene 1641 observaciones, y el segundo archivo contiene para cada una de esas observaciones su label correspondiente respetando el orden de los registros. Es decir, por ejemplo, el label del primer registro de GTExdata.csv es el primer registro del archivo SampleLabels.csv y así sucesivamente. El label que vamos a usar es *Tissue type*.

Se plantean los siguientes puntos a realizar:

1. Realizar una descomposición en valores singulares sobre el set y luego:
    a. Graficar la energía acumulada en función de la cantidad de autovalores, y graficar los autovalores del set.
    b. Con los dos gráficos anteriores, indicar qué valor de ***k*** se podría utilizar para realizar una aproximación.
    c. Realizar una reducción a dos dimensiones[1] del set de datos y graficar los puntos en un scatter-plot utilizando colores para indicar el label correspondiente a cada punto.
    Atención: Prestar demasiada atención a las dimensiones de las matrices U, Sigma y V. Según la biblioteca o función utilizada, las dimensiones pueden variar dependiendo si la SVD se calcula usando su forma reducida o no.
2. Utilizando el algoritmo t-SNE[2], se pide:
    a. Realizar un scatter-plot en el plano 2D, utilizando colores para indicar el label.
    b. Probar varios valores de perplexity (e.g. 3, 30, 1000). ¿Qué efectos tiene el perplexity en el gráfico?

Recomendamos utilizar Jupyter Notebook para organizar claramente el output y comentar los hallazgos, de tal forma que puedan compartirse por ejemplo en GitHub.

---

[1] Para poder realizar la reducción a 2D con SVD, se puede hacer el producto interno entre las primeras dos filas de V^t por A^t siendo A la matriz de datos. Esto deberá arrojará como resultado una matriz de dimensión (2,1641). Otra alternativa es realizar el producto interno entre U y Sigma y sobre el resultado quedarse con las dos primeras columnas

[2] Es posible que t-SNE consuma bastante memoria; en ese caso, se puede partir de un sample de los datos tomado de manera aleatoria.