Pontificia Universidad Católica de Chile
Instituto de Ingeniería Matemática y Computacional
Álgebra Lineal Avanzada y Modelamiento - IMT2230
Segundo Semestre 2023

Proyecto 2

Profesor: Cristóbal Rojas
Ayudante: Pablo Rademacher

El proyecto se puede realizar en grupo de hasta tres personas. **Importante:** Se debe entregar un informe (en formato `.pdf`) donde se reporten los resultados, gráficos y figuras obtenidos. Además, se deben entregar los códigos que se usaron para realizar las partes prácticas. Si bien pueden usar un Jupyter Notebook para los códigos, este **no** cuenta como informe: deben estar todos los resultados documentados en el pdf. La no entrega del informa conlleva un descuenta de 10 décimas.

NO está permitido usar la librería `sklearn` o `scipy`. Para realizar PCA, debe implementarlo usando la función dada en el siguiente párrafo. Además, el uso de la librería `numpy` esta permitido SOLO para realizar operaciones básicas de vectores y matrices, como sumas, productos punto, producto matriz-vector, e invertir matrices. Si desea usar alguna otra función de esta librería, **pregunte antes**.

Para varios puntos del proyecto, deberá obtener la aproximación de rango k de una matriz. Para calcularla, puede usar el siguiente código:

```
import numpy as np


def truncated_svd(A, k):
    U, S, Vh = np.linalg.svd(A, full_matrices=False)
    return U[:, :k], S[:k], Vh[:k]
```

donde U contiene los k primeros vectores singulares izquierdos (como columnas), S es un vector que contiene los primeros k valores singulares, y Vh contiene los primeros k vectores singulares derechos (como filas).

P 1. El objetivo de esta pregunta es construir un identificador de rostros. La matriz `data.npy` contiene 72 imagenes de rostros, correspondientes a 12 mujeres. Puede cargar la matriz y mostrar las imagenes usando el siguiente código:

```
# Cargar matriz
A = np.load("data.npy")

# Mostrar la i-esima imagen
i = 0 # Cambiar si se quiere mostrar otra imagen
imagen = A[i].reshape((241, 181))
plt.imshow(imagen, cmap="gray")
plt.show()
```

Las filas están ordenadas de manera que las imágenes correspondientes a la primera persona están en las filas 1 a 6, las correspondientes a la segunda persona en las filas 7 a 12, y así.

1. Realice la SVD de la matriz A, y úsela para proyectar las imagenes sobre el espacio lineal de dimensión k que mejor las aproxima. Use valores de $k = 1, 5, 10, 25, 50$ y 72. Elija un par de rostros, y para cada valor de k, muestre los "rostros aproximados"(correspondientes a las proyecciones) y compárelos con los originales.

2. Realice PCA a la matriz A (es decir, el paso anterior pero centrando los datos previamente), y úsela para proyectar las imagenes sobre el espacio afín de dimensión k que mejor las aproxima. Use valores de $k = 1, 5, 10, 25, 50$ y 72. Nuevamente, para cada k, muestre un par de "rostros aproximados" y compárelos con los originales. Recuerde descentrar las imágenes una vez proyectadas.

3. Muestre una tabla donde se compare el error cuadrático medio de las aproximaciones encontradas en los items anteriores (es decir, para los valores de k mencionados, y tanto para el caso centrado como el no centrado). Recuerde que este error esta dado por la fórmula

$$ECM(A^{(k)}, A) = \frac{\|A^{(k)} - A\|_F^2}{m}$$

   donde $A^{(k)}$ contiene las proyecciones de las imagenes como filas, y m es la cantidad de imágenes. Comente.

4. Visualize los datos como puntos, proyectándolos en sus dos primeras componentes principales. Coloree los rostros correspondientes a una misma persona de un mismo color, y use colores distintos para rostros de personas distintas. Para mayor claridad, muestre los datos correspondientes a las dos primeras personas.

5. Repita el paso anterior, proyectando ahora sobre la tercera y cuarta componentes principales. Comente. En particular, responda: ¿qué par de componentes (1 y 2 vs 3 y 4) parece ser más relevante para separar a las 2 primeras personas? ¿Cómo explica esto?

6. La matriz `test.npy` contiene seis imágenes, tres de ellas correspondientes a rostros y tres de ellas correspondientes a otros objetos cotidianos. Visualice las seis imágenes. Luego, aproxímelas proyectandolas sobre las 20 primeras componentes principales (de la misma matriz A anterior con los 72 rostros), y visualice esta aproximación. Comente sobre la calidad de las aproximaciones obtenidas. ¿Que imagenes se aproximan bien sobre las componentes principales de A? ¿Porqué?

7. Basado en el punto anterior, proponga un método que permita distinguir si una foto corresponde a un rostro o no (sólo se pide describirlo, no es necesario que lo implemente). Fundamente su propuesta.

P 2. El objetivo de esta pregunta es obtener una visualización sobre las posiciones políticas de los distintos miembros de la cámara de diputados de Chile. En el archivo `votes.csv` se encuentra la información sobre los votos de cada diputado/a en el último año sobre las distintas propuestas de leyes revisadas. Estos votos toman el valor 1 (apruebo), 0 (abstengo) y -1 (rechazo). Además, se da información sobre el partido político de cada persona (columna *party*).

1. Realice una exploración de los datos. En particular notará que hay varios valores `Nan`. ¿Cómo los interpretaría?

2. Hay varios métodos que se pueden usar para lidiar con datos faltantes. Para el caso de este problema, algunas alternativas son:

   i) Eliminar las filas o columnas que contengan algún elemento nulo.

   ii) Reemplazarlo por el valor del voto promedio en esa votación, entre todos los diputados que no presenten valores nulos.

   iii) Reemplazarlo por el valor del voto promedio en esa votación, entre los diputados de la misma bancada que no tengan valores nulos.

   La primera opción no es viable en nuestro caso, ya que en casi todas las filas y columnas de nuestros datos se encuentra algún valor nulo, por lo que enfrentaríamos una reducción muy drástica en la cantidad de datos.

   a) Reemplace los valores nulos, usando la segunda opción. Proyecte los datos (centrandolos previamente) sobre el subespacio de dimensión 2 que mejor los aproxime. Plotee los datos sobre este subespacio, coloreando cada punto de acuerdo a su partido político.

   b) Ahora, reemplace los valores nulos usando la tercera opción. Proyecte estos datos sobre el subespacio de dimensión 2 que mejor los aproxime. Plotee los datos sobre este subespacio, coloreando cada punto de acuerdo a su partido político.

   c) Comente sobre los resultados obtenidos en ambos casos. En particular, ¿observa una división entre los políticos de partidos de derecha con los partidos de izquierda? ¿Que protocolo de reemplazo de valores nulos parece más razonable?