Métodos numéricos y Optimización - primer cuatrimestre de 2023

Trabajo Práctico 3

Escribir un informe reportando los resultados de los siguientes experimentos numéricos. El informe debe contar con una introducción, descripción de los métodos numéricos, análisis de los resultados y conclusiones.

1 Compresión de imágenes

En el archivo *dataset01.tar.gz* se encuentran imágenes de 15 imagenes. Cada imagen es una matriz de $p \times p$ que puede representarse como un vector $\boldsymbol{x} \in \mathbb{R}^{p*p}$. A su vez, es posible armar un matriz de datos apilando los vectores de cada imagen. Se desea aprender una representación de baja dimensión de las imágenes mediante una descomposición en valores singulares.

1. Visualizar en forma matricial $p \times p$ las 10 primeras y las 10 últimas dimensiones (autovectores) de la descomposición obtenida. ¿Qué diferencias existen entre unas y otras? ¿Qué conclusiones pueden sacar?

2. Dada una imagen cualquiera del conjunto (por ejemplo la primera) encontrar d, el número mínimo de dimensiones a las que se puede reducir la dimensionalidad de su representación mediante valores singulares tal que el error entre la imagen comprimida y la original no exceda el 5% bajo la norma de Frobenius. ¿Qué error obtienen si realizan la misma compresión (con el mismo d) para otra imagen cualquiera del conjunto?

2 Clustering de datos

En el archivo *dataset02.csv* se encuentra el dataset X. Este contiene un conjunto de n muestras

$$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n\}$$

con $\boldsymbol{x}_i \in \mathbb{R}^p$ (X es por lo tanto una matriz de $n \times p$ dimensiones). Si bien el conjunto tiene, *a priori*, dimensión alta, suponemos que las muestras no se distribuyen uniformemente, por lo que podremos encontrar grupos de muestras (clusters) similares entre sí. La similaridad entre un par de muestras $\boldsymbol{x}_i, \boldsymbol{x}_j$ se puede medir utilizando una función no-lineal de su distancia euclidiana

$$K_1(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right),$$

para algún valor de σ.

1. Determinar si existen clusters o grupos de alta similaridad entre muestras en el dataset.

2. Determinar a que cluster pertenece cada muestra $\boldsymbol{x}_i$

3. Encontrar en centroide de cada cluster y a partir de estos, armar una clasificador basado en la distancia de una muestra a cada centroide.

Como la dimensionalidad inicial del dataset es muy alta y hay ruido en las muestras va a ser conveniente trabajar en un espacio de dimensión reducida d. Para hacer esto hay que realizar una descomposición de X en sus valores singulares, reducir la dimensión de esta representación, y luego trabajar con los vectores $\boldsymbol{x}$ proyectados al nuevo espacio reducido, es decir $V_d^T \boldsymbol{x}$. Realizar los puntos anteriores para $d = 2$, 4, 20, y p. ¿Para qué elección de d resulta más fácil hacer el análisis? ¿Cómo se conecta esto con los valores singulares de X? ¿Qué conclusiones puede sacar al respecto?