Métodos numéricos y Optimización 2024

Trabajo Práctico 3 - SVD y reducción de la dimensionalidad

Fecha de entrega: 14 Noviembre 2024 23h59

Escribir un informe de máximo 18 carillas reportando los resultados de los siguientes experimentos numéricos. Los códigos desarrollados se deben entregar junto al informe. El informe debe contar con una introducción, descripción de los métodos numéricos, descripción de los experimentos, análisis de los resultados y conclusiones.

1 Compresión de imágenes

En el archivo *dataset_imagenes.zip* se encuentran n imágenes. 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 generando una matriz de $n \times (p*p)$. Se desea aprender una representación de baja dimensión de las imágenes mediante una descomposición en valores singulares.

1. Aprender una representación basada en Descomposición de Valores Singulares utilizando las n imágenes.
2. Visualizar en forma matricial $p \times p$ las imágenes reconstruidas luego de compresión con d dimensiones utilizando distintos valores de d ¿Qué conclusiones pueden sacar?
3. Analizar como evoluciona el error entre las imagenes comprimidas y las originales bajo la norma de Frobenius. Que dimension d asegura que el error a lo largo de cada foto comprimida no excede el 10%?

2 Reducción de la dimensionalidad y Cuadrados Mínimos

En el archivo *dataset.csv* se encuentra el dataset X. Este contiene un conjunto de n muestras (mediciones) que fueron realizadas a través de p sensores

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

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, es de interés entender visualmente como se distribuyen las muestras. Suponemos que las muestras no se distribuyen uniformemente en el espacio $\mathbb{R}^p$, por lo que podremos encontrar grupos de muestras (clusters) con alta similaridad 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

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

¡para algún valor de σ.
Como la dimensionalidad inicial del dataset es muy alta y se supone que algunas dimensiones son mas ruidosas que otras en las muestras, va a ser conveniente trabajar en un espacio de dimensión reducida d.

1. 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 $\mathcal{Z}$, es decir $z = V_d^T \boldsymbol{x}$. Realizar los puntos anteriores para $d = 2$, 6, 10, y p.

2. Analizar la similaridad par-a-par entre muestras en el espacio de dimension X y en el espacio de dimensión reducida d para distintos valores de d utilizando PCA. Comparar estas medidas de similaridad. Ayuda: ver de utilizar una matriz de similaridad para visualizar todas las similaridades par-a-par juntas. ¿Para qué elección de d resulta más conveniente hacer el análisis? ¿Cómo se conecta esto con los valores singulares de X? ¿Qué conclusiones puede sacar al respecto?

3. Los datos X vienen acompañados de una variable dependiente *respuesta o etiquetas* llamada Y (archivo *y.txt*) estructurada como un vector $n \times 1$ para cada muestra. Queremos encontrar el vector β y modelar linealmente el problema que minimice la norma

$$\|X\hat{\beta} - \mathbf{y}\|_2$$

   de manera tal de poder predecir con $X\hat{\beta} = \hat{y}$ lo mejor posible a las etiquetas y, es decir, minimizar el error de predicción utilizando todas las variables iniciales. Resolviendo el problema de cuadrados mínimos en el espacio original X, que peso se le asigna a cada dimensión original si observamos el vector β?

4. Usando la representacion aprendida con PCA y $d = 2$: mejora la predicción $\|Z\hat{\beta} - \mathbf{y}\|_2$ en comparacion a no realizar reduccion de dimensionalidad? Cuales muestras son las de mejor predicción con el mejor modelo?