

Métodos numéricos y Optimización 2024

Trabajo Práctico 3 - SVD y reducción de la dimensionalidad

Fecha de entrega: 06 Junio 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 Reducción de la dimensionalidad y Cuadrados Mínimos

En el archivo *dataset.csv* se encuentra el dataset X . Este contiene un conjunto de n muestras que fueron medidas a través de un sensor

$$\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n\}$$

con $\mathbf{x}_i \in \mathbb{R}^p$ (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 similitud entre sí. La similitud entre un par de muestras $\mathbf{x}_i, \mathbf{x}_j$ se puede medir utilizando una función no-lineal de su distancia euclidiana

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

para algún valor de σ .

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 . 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 \mathbf{x} proyectados al nuevo espacio reducido \mathcal{Z} , es decir $\mathbf{z} = V_d^T \mathbf{x}$. Realizar los puntos anteriores para $d = 2, 6, 10$, y p . ¿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?

1. Determinar la similitud 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 similitud. Ayuda: ver de utilizar una matriz de similitud para visualizar todas las similitudes par-a-par juntas.
2. De las p dimensiones originales del dataset, cuales son las mas representativas con respecto a las dimensiones d obtenidas por SVD? Indicar que dimensiones originales del conjunto p son las mas importantes y el método utilizado para determinarlas.

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. Usando PCA, ¿qué dimensión d mejora la predicción? ¿Cuáles muestras son las de mejor predicción con el mejor modelo? Resolviendo el problema de cuadrados mínimos en el espacio original X , ¿qué peso se le asigna a cada dimensión original si observamos el vector β ?

2 Compresión de imágenes

En el archivo *dataset_imagenes1.zip* se encuentran n imágenes. Cada imagen es una matriz de $p \times p$ que puede representarse como un vector $\mathbf{x} \in \mathbb{R}^{p \times p}$. A su vez, es posible armar una matriz de datos apilando los vectores de cada imagen generando una matriz de $n \times (p \times p)$. 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 distintos valores de d dimensiones ¿Qué conclusiones pueden sacar?
3. Utilizando compresión con distintos valores de d medir la similaridad entre pares de imágenes (con alguna métrica de similaridad que decida el autor) en un espacio de baja dimensión d . Analizar cómo la similaridad entre pares de imágenes cambia a medida que se utilizan distintos valores de d . ¿Cuáles imágenes se encuentran cerca entre sí? ¿Alguna interpretación al respecto? Ayuda: ver de utilizar una matriz de similaridad para visualizar todas las similaridades par-a-par juntas.
4. Dado el dataset *dataset_imagenes2.zip* 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 de cada imagen comprimida y su original no exceda el 10% bajo la norma de Frobenius. Utilizando esta última representación aprendida con el dataset 2 ¿Qué error de reconstrucción obtienen si utilizan la misma compresión (con la misma base de d dimensiones obtenida del dataset 2) para las imágenes *dataset_imagenes1.zip*?