Elements of Machine Learning 2024

Tarea 02

02.febrero.2024

En esta tarea vamos a explorar dos conjuntos de datos a través de PCA: (1) un conjunto de datos de estaciones climáticas en Canadá; y (2) un conjunto de datos muy simple sobre crímenes.
Además, veremos una aplicación simple a imágenes: el PCA como método de compresión de información.

1. El conjunto de datos **weather.csv**, trata de los promedios mensuales de la temperatura (en Celsius) en 35 estaciones canadienses de monitoreo (meses 1 a 12). El interés del análisis es comparar las estaciones entre sí con base en sus curvas de temperatura.

   Considerando las 12 mediciones por estación como un vector $\mathbf{x} = (x_1, x_2, \ldots, x_{12})$, aplica un análisis de componentes principales PCA a los datos. Como $\mathbf{x}$ representa (un muestreo de) una curva, este tipo de datos se llaman datos funcionales. Realizar lo siguiente:

   - Hacer un análisis exploratorio muy breve de los datos.
   - Proyectar los datos a sus primeras dos componentes principales. Grafica e interpreta como curva (de longitud 12) a la primera y segunda componentes principal $\mathbf{p}_1$ y $\mathbf{p}_2$, esto es, grafica $(i, \mathbf{p}_{1i})$ e $(i, \mathbf{p}_{2i})$, para $i = 1, 2, \ldots, 12$.
   - Representar los datos e interpretar los primeros dos componentes en un biplot. Agrupa (de manera intuitiva) e interpreta las estaciones en el biplot. (Aquí se sugiere mostrar las etiquetas con los nombres de las estaciones, y tener a la mano un mapa de Canadá).

2. A partir de una base de datos con actos delictivos en EE.UU (1970), se construyó la tabla con las correlaciones entre la ocurrencia de 7 clases de delitos, como aparece en la tabla **crimes.dat**.

   Consideramos cada clase de delito como una observación. Con la matriz de correlación, podemos inferir una distancia entre dos observaciones como 1 menos su correlación $d(\mathbf{x}_i, \mathbf{x}_j) = 1 - \rho_{ij}$. Observe que las correlaciones en la tabla son siempre positivas. Así, la distancia mínima 0 corresponde a correlación máxima 1 entre las variables correspondientes.

   Encontrar una visualización usando escalamiento multidimensional para estas observaciones y buscar una interpretación sencilla del primer eje principal.

3. A partir de los datos de distancias en el archivo **distances.csv**, construir un mapa visual de las ciudades con coordenadas sintéticas, usando las técnicas de escalamiento multidimensional. Asegurarse que su visualización sea acorde a la posición usual en la lectura de un mapa (norte: arriba, sur: abajo, oeste: izquierda, este: derecha).

   Calcular las distancias (euclideanas) entre las ciudades a partir de las coordenadas obtenidas, y comparar con la matriz original de distancias. ¿Se parecen o no?

4. Históricamente uno de los primeros usos de PCA en el área de procesamiento de imágenes fue como método de compresión. Para ello, si tenemos una imagen de tamaño $H \times W$ pixeles, ésta se subdivide en bloques de $C \times C$ pixeles (por ejemplo, tomar C un factor común de las dimensiones H y W de la imagen). Con los valores de los pixeles en cada bloque se forma un vector

$$\mathbf{b}_i = (x_1, x_2, \ldots, x_{c^2}) \in \mathbb{R}^{c^2}$$

La matriz de datos $\mathbb{X}$ se forma con todos estos vectores provenientes de los bloques $\mathbf{b}_i$ vectorizados. La compresión consiste en proyectar los datos sobre los primeros k componentes principales, mientras que la decompresión consiste en reconstruir la imagen a su tamaño original $H \times W$ a partir de estas proyecciones.

Implementar lo anterior para 2 imágenes sencillas (en escala de gris o a color) y mostrar el efecto del valor de k sobre la calidad de la reconstrucción.
Analizar y comentar cómo cambia el error de reconstrucción y la calidad visual a medida que se incrementa o disminuye k.

No se olviden de redactar su análisis de datos, destacando sus conclusiones e *insights* más importantes.