Ciencia de Datos 2021

Lista 02

13.febrero.2021

1. Para un estudio se mide la temperatura en diferentes posiciones del cuerpo de una muestra de personas. Un investigador expresa todas las temperaturas en grados Celcius. Otro investigador convierte primero todas estas temperaturas a grados Fahrenheit.

   ¿Cómo se relacionan las matrices de covarianza de sus datos?
   Si ambos deciden proyectar en la dirección de máxima varianza, ¿obtendrán las mismas direcciones de proyección? Explica tu respuesta.

2. Dibuja un ejemplo de una distribución (2D) donde cualquier proyección es de varianza máxima.

3. Sea $X = (X_1, X_2, \ldots, X_d)$ una v.a. multidimensional con matriz de covarianza $Cov(X)$ y $\mathbb{E}(X) = 0$.
   Si $\mathbf{l} = (l_1, l_2, \ldots, l_d)^T \in \mathbb{R}^d$ es un autovector de $Cov(X)$ con autovalor λ y $Y = \langle \mathbf{l}, X \rangle$ muestra que $Cov(Y, X_i) = \lambda l_i$.

4. Sea $\mathbb{X}$ una matriz arbitraria $n \times d$ y $\mathbb{J}$ la matriz $n \times n$ para centrar $\mathbb{X}$. Verifica que

   a) $\mathbb{J}$ es una matriz de proyección.
   b) el vector $\mathbf{1} = (1, \ldots, 1)^T$ de longitud n es un autovector de $\mathbb{J}$ con autovalor 0.

5. Considera los datos `weather.csv`. Se trata de los promedios mensuales de la temperatura (en Celsius) en 35 estaciones canadienses de monitoreo. El interés es comparar las estaciones entre sí con base en sus curvas de temperatura.
   Considerando las 12 mediciones por estación como un vector X, aplica un análisis de componentes principales. Como X representa (un muestreo de) una curva, este tipo de datos se llama datos funcionales. Interpreta y dibuja (como curva) los primeros dos componentes, p_1, p_2 es decir grafica $\{(i, p_{1i})\}$ y $\{(i, p_{2i})\}$. Agrupa e interpreta las estaciones en el biplot (ten en mente un mapa de Canadá).

6. 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. Podemos medir la distancia entre dos observaciones como 1 menos su correlación (las correlaciones en la tabla son siempre positivas). Así, la distancia mínima (0) corresponde a correlación máxima (1) entre las variables correspondientes.

   Encuentra una visualización usando escalamiento multidimensional para estas observaciones y busca una interpretación del eje principal.

7. Históricamente uno de los primeros usos de PCA en el área de procesamiento de imágenes fue como método de compresión. Para ello, se divide la imagen en bloques de $c \times c$ pixeles (por ejemplo, tome c un denominador común de las dimensiones de la imagen). Con los valores de los pixeles en cada bloque se forma un vector $(x_1, x_2, \ldots, x_{c^2}) \in \mathbb{R}^{c^2}$. La matriz de datos se forma con todos estos vectores provenientes de los bloques vectorizados. La compresión consiste en proyectar los datos sobre los primeros k componentes principales. La decompresión consiste en reconstruir los datos a partir de estas proyecciones.

   Implementa lo anterior para algunas imágenes sencillas (en escala de gris) y muestra el efecto del valor de k sobre la calidad de la reconstrucción.