Ciencia de Datos
Tarea 2

Para entregar el 4 de marzo de 2020

1. Este ejercicio es sobre PCA.

   a) Realiza PCA a la matriz

   $$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

   donde $\rho > 0$. Ahora, cambia la escala de X_1, es decir, considera la covarianza de cX_1 y X_2. ¿Cómo cambian los componentes principales al realizar este escalamiento?

   b) Considera los datos del archivo `ushealth.csv`, que contiene el número reportado de muertes en los 50 estados de los Estados Unidos durante 1985, clasificado de acuerdo a 7 categorias: `acc`, `card`, `canc`, `pul`, `pneu`, `diab` y `liv`:

   ```
   'land area': a numeric vector
   'popu 1985': population 1985, a numeric vector
   'acc': accident, a numeric vector
   'card': cardiovascular, a numeric vector
   'canc': cancer, a numeric vector
   'pul': pulmonar, a numeric vector
   'pneu': pneumonia flu, a numeric vector
   'diab': diabetis, a numeric vector
   'liv': liver, a numeric vector
   'doc': doctors, a numeric vector
   'hosp': hospitals, a numeric vector
   'reg': US region, a factor with levels 'Northeast' 'Midwest'
        'South' 'West'
   'div': US division, a factor with levels 'New England' 'Mid
        Atlantic' 'E N Central' 'W N Central' 'S Atlantic' 'E S
        Central' 'W S Central' 'Mountain' 'Pacific'
   ```

   Realiza PCA, con y sin normalización e interpreta los resultados. ¿Qué puedes decir sobre la relación entre las causas y el número de muertes? Usa el resultado del inciso anterior para explicar el efecto de usar PCA normalizado y sin normalizar. ¿Cuál prefieres usar en este caso y porqué? ¿Qué recomendación darías al respecto al usar PCA en general?

2. Considera nuevamente las imágenes de rostros que vimos en clase. Usaremos las imágenes pequeñas `olivetti_faces` del `sklearn`, que tienen 10 fotos por cada uno de los 40 sujetos.

   a) Separa un conjunto de entrenamiento (90 %) y prueba (toma por ejemplo, la última foto de cada sujeto como prueba). Obtén las eigenfaces del conjunto de entrenamiento. Visualiza los primeros dos componentes principales ¿Encuentras patrones interesantes?

   b) Proyecta los datos de prueba en los componentes principales. Verifica si se "ubican" en su "individuo" correspondiente al graficarlos en los primeros dos componentes principales.

   c) Usa el método del vecino más cercano para identificar a un "sujeto" de prueba en las imágenes de entrenamiento. Usa la distancia euclideana en el espacio de los p componentes principales. Decide qué valor de p usar. El obejtivo es obtener algo como lo que se muestra en la Figura 1: ¿Puedes identificar correctamente a

Figura 1: Identificación de un individuo de prueba usando el vecino más cercano en el espacio de los primeros p PC.

   los sujetos usando éste criterio? ¿Qué tanto influye el valor de p?

   d) Considera una(s) imágen(es) que no están la base de datos ¿Qué se te ocurre para detectar casos como los que muestran en la Figura 2?

3. Supón que eres asesor técnico de la Secretaría de Desarrollo Social de Nuevo León. Para establecer estrategias de desarrollo, la Secretaría desea primero, hacer un análisis del estado actual de la entidad, por lo que ha revisado el índice de marginación elaborado por el Consejo Nacional de Población (CONAPO) y ha subrayado dos cosas: 1) no entiende cómo lo calcularon y 2) le gustaría explorar otra forma de

Figura 2: Datos fuera de muestra.

hacerlo. Para esto, recurre a ti para que ayudes a analizar la información y a resolver las dudas que surgieron.

a) Reproduce los resultados del índice de marginación a nivel localidad para el estado de NL, el cual se muestra en la Figura 3, y puedes encontrar con mayor detalle en el archivo `conapo_marginacion_nl.xls`. [1].

Para esto, utiliza los datos del Censo de Población y Vivienda 2010 reportados en el INEGI, los cuales, para facilitarte la tarea, he concentrado y adecuado en el archivo `censo_nl.csv`. El diccionario de las variables del censo puedes verlos en `diccionariodatosscince.pdf`. Realiza un reporte ejecutivo (como para que lo entienda un político), explicando los resultados y la metodología usada para crear el índicador. Agrega apéndices técnicos a tu reporte si lo consideras necesario [2].

b) ¿Qué otra información propondrías que se incluyera dentro de la elaboración del índice (ya sea de estadísticas oficiales o de otra fuente)? ¿Estás de acuerdo con la metodología usada? ¿Tienes alguna otra propuesta para la elaboración del índice?

[1] Si no pudieras reproducirlo, explica porqué, ya que en teoría, tienes disponible toda la información para hacerlo.

[2] Ten cuidado con los datos faltantes y NA, que en este caso se muestran con valores negativos. Decide cómo tratarlos y especifícalo en el reporte.

Puedes recurrir también al documento oficial que reporta la CONAPO, que se encuentra en Capitulo01.pdf al Capitulo03.pdf, pero sobre todo en AnexoC.pdf

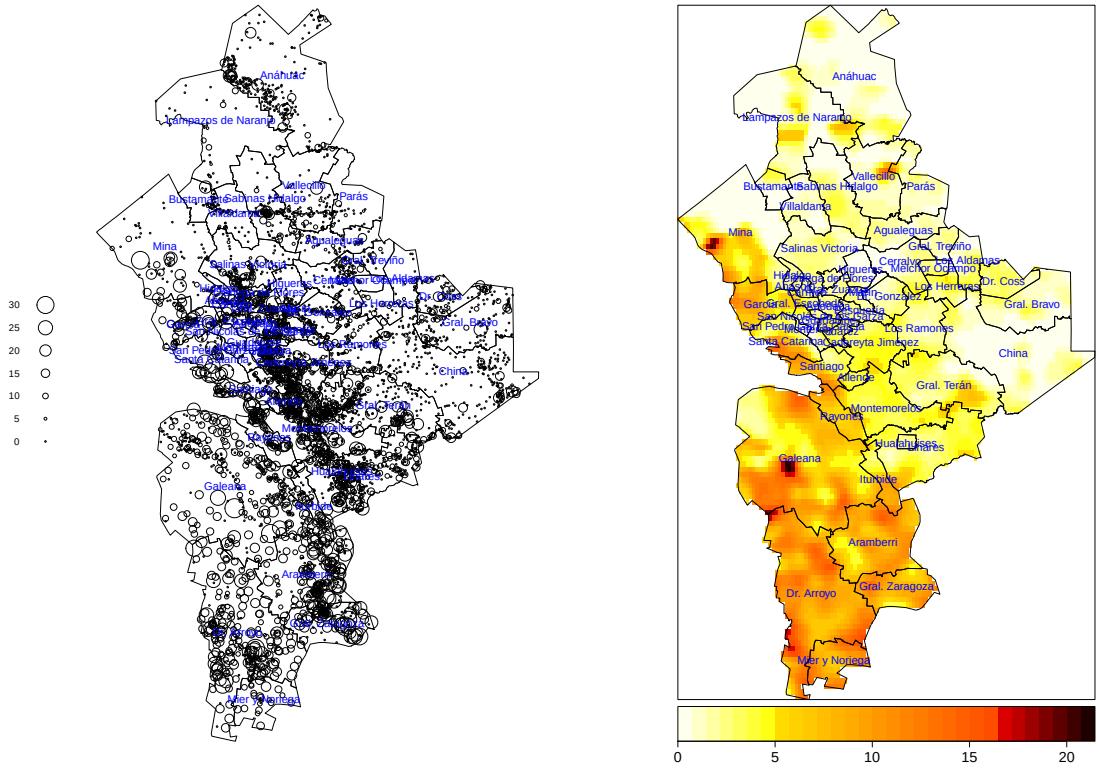

Figura 3: Índice de marginación en escala 0 a 100 del estado de Nuevo León. En la izquierda se muestra el índice para cada localidad (georeferenciada), donde el diámetro del írculo es proporcional al valor del indicador. En la derecha, se muestra una representación suavizada de los datos.