Ciencia de Datos
Tarea 2

Para entregar el 26 de febrero de 2021

1. Considera una matriz de datos $\mathbf{X}_{n\times d}$. PCA puede formularse también como el problema de encontrar un subespacio (ortonormal) de baja dimensión de forma tal que se minimicen los errores de las proyecciones de los datos en tal subespacio.

   Si consideramos una base ortonormal $\{\mathbf{u}_j\}$, $j = 1, \ldots, d$, ya vimos que una observación $\mathbf{x}_i$ puede expresarse como una combinación lineal

   $$\mathbf{x}_i = \sum_{j=1}^{d} \alpha_{ij}\mathbf{u}_j.$$

   Por la ortogonalidad de $\mathbf{u}_j$, podemos expresar $\alpha_{ij} = \mathbf{x}_i'\mathbf{u}_j$. Entonces

   $$\mathbf{x}_i = \sum_{j=1}^{d} (\mathbf{x}_i'\mathbf{u}_j)\mathbf{u}_j.$$

   Ahora, considera una aproximación basada en los primeros $p < d$ vectores de la base de acuerdo al modelo lineal:

   $$\widehat{\mathbf{x}}_i = \sum_{j=1}^{p} z_{ij}\mathbf{u}_j + \sum_{j=p+1}^{d} b_j\mathbf{u}_j.$$

   Observa que los coeficientes z_{ij} *dependen* de la observación i, mientras que b_j son constantes para todas las observaciones.

   Considera la minimización de la siguiente función de costo:

   $$L = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \widehat{\mathbf{x}}_i\|^2. \tag{1}$$

   a) Muestra que en el mínimo de (1):

   $$\begin{aligned} z_{ij} &= \mathbf{x}_i'\mathbf{u}_j, \qquad j = 1, \ldots, p \\ b_j &= \bar{\mathbf{x}}'\mathbf{u}_j, \qquad j = p+1, \ldots, d \\ \mathbf{x}_i - \widehat{\mathbf{x}}_i &= \sum_{j=p+1}^{d} \left[(\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{u}_j\right]\mathbf{u}_j \end{aligned},$$

es decir, la "desviación" está en el espacio ortogonal de los componentes principales.

b) Considerando lo anterior, muestra que (1) puede escribirse como

$$L = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=p+1}^{d} (\mathbf{x}_i' \mathbf{u}_j - \bar{\mathbf{x}}' \mathbf{u}_j)^2 = \sum_{j=p+1}^{d} \mathbf{u}_j' \mathbf{S} \mathbf{u}_j,$$

es decir, la solución se obtiene resolviendo un problema de valores y vectores propios (restringida), como vimos antes.

Usando el método de Lagrange, es fácil ver (no es necesario demostrarlo) que lo anterior es equivalente a minimizar $L = \sum_{j=p+1}^{d} \lambda_j$, donde λ_j son los valores propios de $\mathbf{S}$, por lo tanto, debemos escoger los vectores propios correspondientes a los valores propios más chicos: $0 \leq \lambda_d \leq \lambda_{d-1} \leq \lambda_{d-p}$, por lo que la mejor aproximación de $\mathbf{x}$ (en los componentes principales) está dada por los eigenvectores que corresponden a los eigenvalores más grandes, tal como lo vimos en clase.

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

   a) Trata de reproducir los resultados del índice de marginación a nivel localidad para el estado de NL, el cual se muestra en la Figura 1, 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].

[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

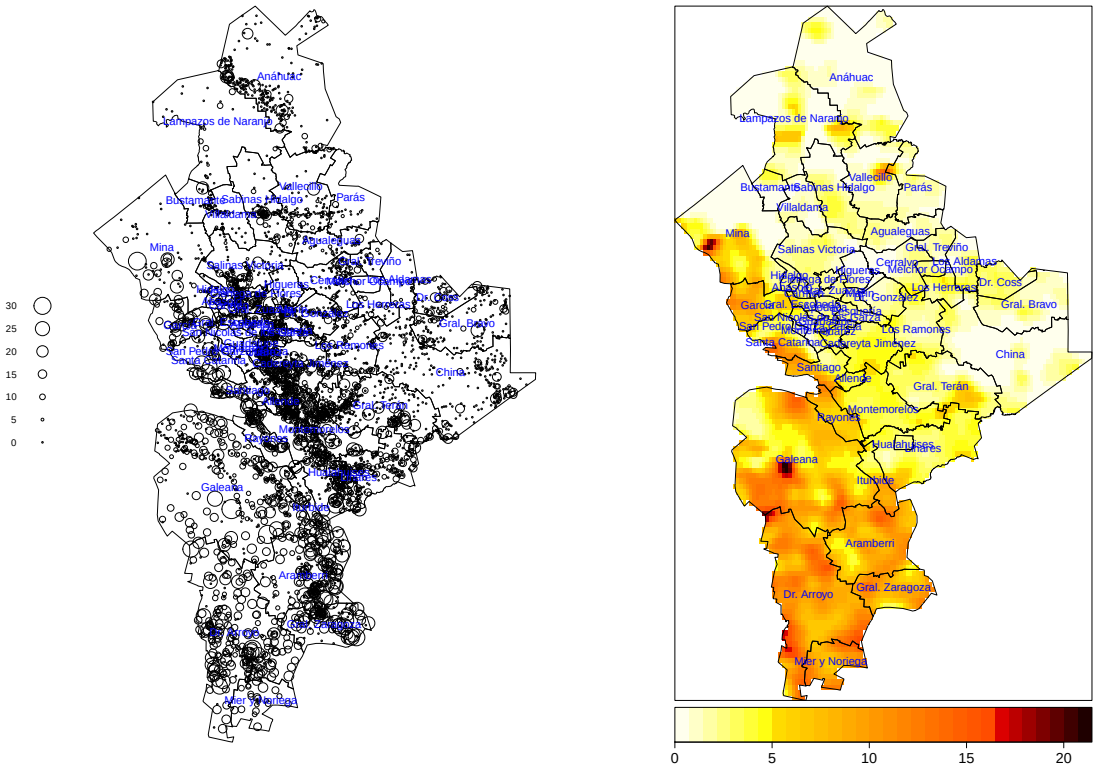

Figura 1: Índice de marginación en escala 0 a 100 del estado de Nuevo León. En la izquierda se muestra el índice para cada localidad (georeferenciada), donde el diámetro del írculo es proporcional al valor del indicador. En la derecha, se muestra una representación suavizada de los datos.

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?

3. Considera el conjunto de datos Labelled Faces in the Wild (LFW), que consiste en fotografías de rostros recolectados de internet y contenido en `sklearn`. Algunos rostros identificados, tienen varias fotos incluídas en el dataset. Vamos a considerar solo aquellas personas que tienen al menos 70 fotografías de su rostro, tambien, vamos a considerar el tamaño original de la imágen (125×94)

```
from sklearn.datasets import fetch_lfw_people
lfw_people = fetch_lfw_people(min_faces_per_person=70, resize=1)
```

Esto resulta en 1288 imágenes que pertenecen a alguna de las etiquetas:

```
>>> for name in lfw_people.target_names:
>>>     print(name)
Ariel Sharon
Colin Powell
Donald Rumsfeld
George W Bush
Gerhard Schroeder
Hugo Chavez
Tony Blair
```

Una muestra de las imágenes puede verse en la Figura 2.

Figura 2: Ejemplo de los rostros del dataset LFW.

a) Separa un conjunto de entrenamiento (80 %) y prueba (puedes usar la función `train_test_split` de `sklearn.model_selection`, por ejemplo:

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test, names_train,
names_test = train_test_split(X, y, target_names[y],
test_size=0.2, random_state=42)
```

donde antes, tuviste que declarar `X,y,target_names` (ve la documentación de `fetch_lfw_people`). 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 objetivo es obtener algo como lo que se muestra en la Figura 3:

Figura 3: Identificación de un individuo de prueba usando el vecino más cercano en el espacio de los primeros p PC.

¿Puedes identificar correctamente a 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 prevenir casos como los que muestran en la Figura 4?

4. Dado un conjunto de datos centrados $\mathbf{X}_{n\times d}$, vimos que hacer PCA, es realizar la descomposición espectral de la matriz de covarianzas muestral, que puede estimarse como $\mathbf{S} = \mathbf{X}'\mathbf{X}$ (omitimos el coeficiente $n-1$).

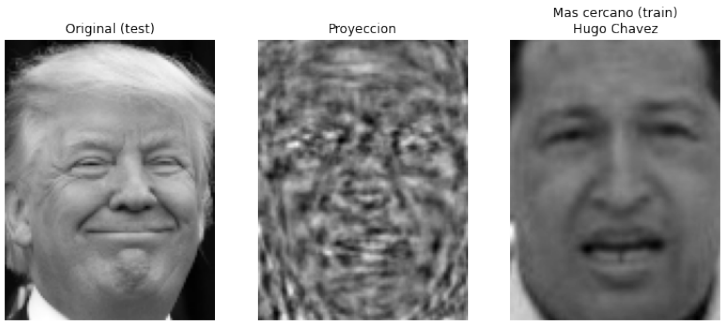

Figura 4: Identificación de un individuo de prueba usando el vecino más cercano en el espacio de los primeros p PC.

Ahora, considera la matriz $K_{n\times n} = \mathbf{X}\mathbf{X}'$.

a) Muestra que es equivalente realizar PCA en $\mathbf{S}$ o en $\mathbf{K}$, es decir, que $(\lambda^{-1/2}\mathbf{X}\mathbf{u}, \lambda)$ es un par eigenvector-eigenvalor normalizado de $\mathbf{K}$, y a su vez, $(\lambda^{-1/2}\mathbf{X}^T\mathbf{v}, \lambda)$ es un par eigenvector-eigenvalor normalizado de $\mathbf{S}$, donde $\mathbf{u}$ y $\mathbf{v}$ son vectores propios de $\mathbf{S}$ y $\mathbf{K}$, respectívamente.

b) Verifica experimentalmente el resultado del inciso previo en el conjunto de imágenes LFW que usaste en el ejercicio anterior. ¿En qué casos es recomendable usar $\mathbf{K}$?