Ciencia de Datos
Tarea 2

Para entregar el 23 de febrero de 2019

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, clasificado de acuerdo a 7 categorias: accidentes `acc`, cardiovascular `card`, cáncer `canc`, pulmonar `pul`, neumonia `pneu`, diabetes `diab` y enfermedades del hígado `liv`.

   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. 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, ypuedes encontrar con mayor detalle en el archivo `conapo_marginacion_nl.xls`. [1].

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

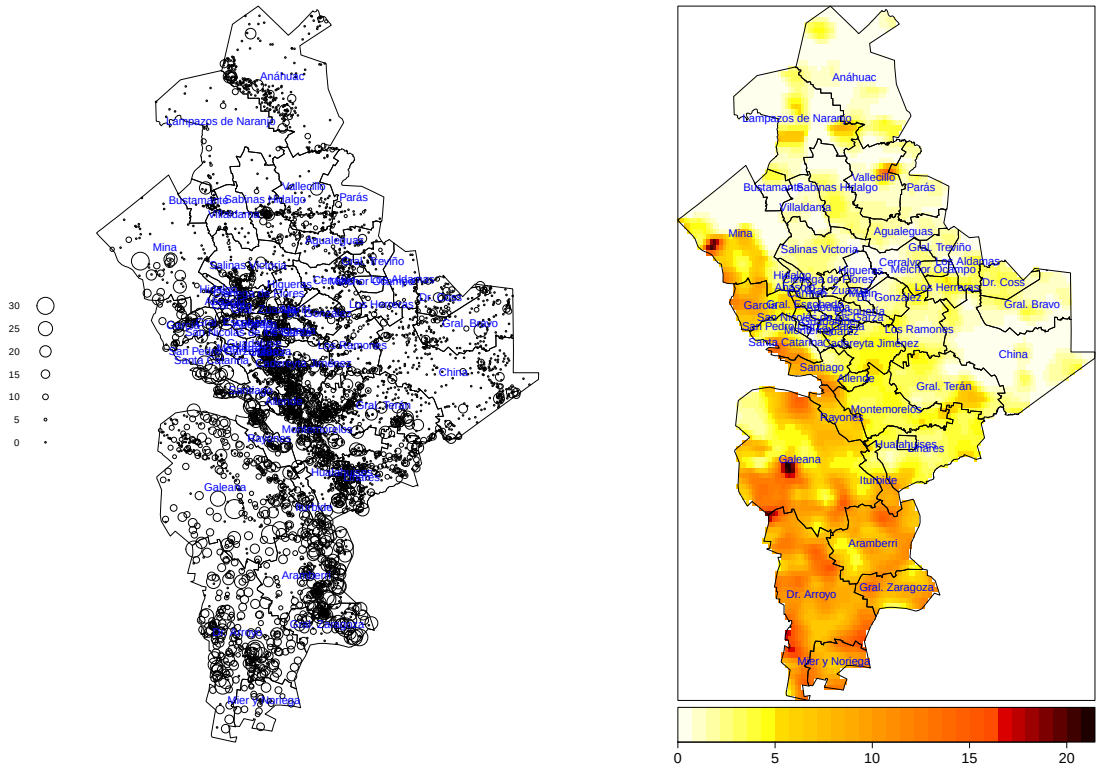

Figura 1: Índice de marginación en escala 0 a 100 del estado de Nuevo León. En la izquierda se muestra el índice para cada localidad (georeferenciada), donde el diámetro del írculo es proporcional al valor del indicador. En la derecha, se muestra una representación suavizada de los datos.

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].

[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

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. En los datos que se presentan en `mnistXtrain.dat` y `mnistXtest.dat` se encuentran dígitos escritos a mano, digitalizados y normalizados en 28×28 pixeles. Se codificó cada imagen como un vector donde cada entrada corresponde a los valores de los pixeles. Todos estos vectores son puestos uno tras otro. Los archivos `mnistYtrain.dat` y `mnistYtest.dat` contiene el dígito al que corresponde cada imagen de los archivos anteriores.

   a) Implementa un clasificador para las imágenes que pertenecen a uno de los $k \in K = \{0, 1, \ldots, 9\}$ dígitos usando regresión-PCA múltivariada:

   $$\mathbf{Y} = \mathbf{Z}_p \widehat{\mathbf{B}}_p,$$

   donde $\mathbf{Y}_{n \times |K|}$ es una matriz indicadora, donde cada renglón tiene ceros excepto en el lugar que corresponde al valor y_k, donde colocamos un 1. Por ejemplo, si alguna imagen corresponde al dígito "3", el renglón correspondiente en $\mathbf{Y}$ será $(0, 0, 0, 1, 0, 0, 0, 0, 0, 0)$.

   $\mathbf{Z}_p$ es una matriz con los primeros p componentes principales y $\widehat{\mathbf{B}}_p$ es una matriz cuyas columnas contienen los $|K|$ coeficientes $\widehat{\boldsymbol{\beta}}_p$ obtenidos como lo vimos en clase.

   Con esta formulación, asumimos un modelo lineal para cada respuesta $\mathbf{y}_k$:

   $$\widehat{\mathbf{y}}_k = \mathbf{Z}_p \widehat{\boldsymbol{\beta}}_p^k,$$

   y la clasificación para alguna observación $\mathbf{z}$ se obtiene mediante

   $$\widehat{C}(\mathbf{z}) = \arg \max_{k \in K} \hat{y}_k.$$

   Utiliza los datos de `mnistXtrain.dat` para ajustar el modelo y `mnistXtest.dat` para probarlo. Obten el error obtenido, tanto en los datos de entrenamiento como los de prueba, usando diferentes valores de p componentes principales. Realiza una gráfica de error vs p. ¿Qué valor de p recomendarías usar?

   Nota: Puedes usar la función general para modelos lineales `lm()` de `R`, la cual puede usarse también para regresión lineal multivariada. Revisa la ayuda de la función.

b) **Opcional (puntos extra)**. Programa una aplicación interactiva donde dibujes un número y te diga qué dígito es usando el clasificador del inciso anterior. Puedes usar y modificar el script `ui.r` y `server.r` que les proporciono, los cuales se ejecutan con

```
library(shiny)
## carga el objeto que contiene los PC calculados previamente
## pon la ruta donde guardaste el objeto que contiene los datos de PCA
pc <- readRDS("/home/victor/cursos/ciencia_de_datos_2019/programs/pc_mnist.rds")
## sus categorias
y.train <- as.numeric(unlist(read.table(
           "/home/victor/cursos/ciencia_de_datos_2019/data/digits_mnist/mnistYtrain.dat",
           header=TRUE)))

## incluye la ruta donde esta la carpeta con archivos server y ui
runApp(appDir="~/cursos/ciencia_de_datos_2019/programs/")

library(shiny)
## incluye la ruta donde esta la carpeta con archivos server y ui
runApp(appDir="~/cursos/ciencia_de_datos_2019/programs/")
```

Tu programa debe mostrar al menos, el área para dibujar el número y el dígito que estimó tu clasificador.