Ciencia de Datos
Tarea 1

Para entregar el 15 de febrero de 2018

1. Considera los datos que se encuentran en el archivo `wine_quality.csv`, que contienen diferentes características fisico-químicas de las variantes tinto y blanco del *Vinho Verde*, un vino Portugués que cuenta con denominación de origen controlada. El archivo contiene 11 atributos (covariables) y una variable categórica (sensorial) que indica la calidad del vino en escala del 0 (muy malo) al 10 (excelente). Se agregó además una columna que identifica el tipo de vino (tinto o blanco). Para mas detalles, consulta el archivo `winequality.names.txt`.

   Realiza un análisis exploratorio de los datos con las herramientas que consideres apropiadas. Comenta tus hallazgos. ¿Es posible distinguir el tipo de vino a partir de sus características fisicoquímicas?

   Un objetivo interesante es explorar la relación entre las características medidas y la calidad del vino (tinto y blanco por separado). Verifica si es posible encontrar visualmente tal relación. Puedes simplificar o agrupar la escala de calidad en, por ejemplo, 3 categorias: malo, medio y excelente.

   El resultado debe ser un reporte corto donde describas los pasos que seguiste y las conclusiones a las que llegas, incluyendo solamente las gráficas más ilustrativas o informativas que tu consideres.

2. 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?

3. Supón que un miembro del gabinete del gobierno de Nuevo León quiere plantear una estrategia de desarrollo social en el estado. Para esto, ha visto los últimos índices de marginación de Nuevo León y ha subrayado dos cosas: 1) no entiende cómo los calcularon y 2) le gustaría explorar otra forma de hacerlo. Para esto, está buscado personas que puedan ayudarlo a analizar la información (¡seguro que pagan muy bien!).

   En este ejercicio, abordarás esta tarea usando estadísticas oficiales.

   a) Trata de reproducir los resultados del índice de marginación a nivel localidad para el estado de NL [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`. Los resultados reportados por la CONAPO se encuentran en el archivo `conapo_marginacion_nl.xls`. Realiza un reporte ejecutivo (como para que lo entienda El Bronco), explicando los resultados y la metodología usada. 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?

4. En los datos que se presentan en `oef.train` y `oef.test` se encuentran dígitos escritos a mano, digitalizados y normalizados en 16 por 16 pixeles. Se codificó cada imagen como un vector: en la primera posición se encuentra el número que representa la imagen y después, renglón por renglón, los valores de los pixeles. Todos estos vectores son puestos uno tras otro.

[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

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 `oef.train` para ajustar el modelo y `oef.test` 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)
## incluye la ruta donde esta la carpeta con archivos server y ui
runApp(appDir="~/ruta/carpeta_archivos_server_ui/")
```

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