Presentación

En esta práctica resolveréis un caso de uso propuesto mediante el análisis de componentes principales. Este caso de uso os permitirá poner en práctica los conceptos trabajados en este reto, entender y coger destreza en su aplicación a un caso de uso concreto utilizando datos reales o realistas. Veréis también la necesidad de utilizar un lenguaje de programación como, por ejemplo, `R` para su resolución y cogeréis destreza en su utilización.

Competencias

En esta PEC se trabajan las siguientes competencias del Grado en Ciencia de Datos Aplicada:

- Que los estudiantes hayan demostrado poseer y comprender conocimientos en un área de estudio que parte de la base de la educación secundaria general, y se suele encontrar a un nivel que, si bien se apoya en libros de texto avanzados, incluye también algunos aspectos que implican conocimientos procedentes de la vanguardia de su campo de estudio.
- Utilizar de forma combinada los fundamentos matemáticos, estadísticos y de programación para desarrollar soluciones a problemas en el ámbito de la ciencia de datos.
- Uso y aplicación de las TIC en el ámbito académico y profesional.

Objetivos

Los objetivos concretos de esta Práctica son:

- Comprender la utilidad de los conceptos de álgebra lineal que se han trabajado en los retos 1-3 en la aplicación en el ámbito de la ciencia de datos mediante el análisis de componentes principales y la descomposición en valores singulares.
- Ser capaz de resolver un problema utilizando la descomposición en valores singulares en un caso de uso utilizando datos reales o realistas.
- Entender la utilidad de utilizar un lenguaje de programación para el tratamiento de grandes volúmenes de datos.

- Coger destreza en la utilización del lenguaje R para la resolución de problemas con un gran volumen de datos.

Descripción de la Práctica a realizar

Ser capaces de reducir la dimensionalidad de datos es muy importante en el ámbito de la ciencia de datos donde normalmente trabajamos con altos volúmenes de información. En este reto veremos dos técnicas muy extendidas que nos permitirán reducir la dimensionalidad de nuestros datos: la descomposición en valores singulares y el análisis de componentes principales, que están muy relacionadas. Ambas técnicas, basadas en los conceptos del álgebra lineal analizados en los retos 1, 2 y 3, permiten considerar un conjunto de datos inicial y transformarlo de manera que, o bien la dimensión resultante sea inferior o bien la nueva representación de los datos permita desvelar información relevante.

Por un lado, os pedimos que respondáis un **cuestionario** (se puede encontrar en el aula Moodle entrando en el enlace "Cuestionarios" en la parte derecha del aula) en el que vamos a trabajar la parte más instrumental de este reto en una serie de preguntas genéricas.

Os pedimos también que resolváis la práctica descrita en este documento. Estos ejercicios os plantearán escenarios propios de la ciencia de datos y veréis como los conceptos trabajados en este reto tienen relevancia en estos contextos.

Recursos

Recursos Básicos

- Documento introductorio a la descomposición en valores singulares para la ciencia de datos
- Módulo 4
- Documento de problemas sobre la descomposición en valores singulares enfocados a la ciencia de datos

Recursos Complementarios

- Caso de uso y guía de resolución en R.

Criterios de valoración

- La práctica se ha de resolver de manera individual.
- Es necesario justificar todos los pasos realizados en la resolución de la Práctica.

Tened en cuenta que las dos actividades que se plantean en este reto (la resolución de la práctica que se plantea en este documento y la tabla resumen) serán parte de la nota de prácticas (Pr = (Pr1 + Pr2) / 2). La nota de estas actividades corresponde a la Pr1 (con un peso del 20 % para el cuestionario semanal y un 80 % para la práctica). Para más información sobre el modelo de evaluación de la asignatura, consultad el plan docente.

Formato y fecha de entrega

Para realizar la práctica correctamente, se debe consultar y contestar la tabla resumen asociada a la práctica con los resultados obtenidos. La podéis encontrar en el Moodle "RETO 4 - Tabla resumen de la Práctica 1". Hay dos intentos para responder el cuestionario, el primero de los cuales tendrá feedback en las respuestas. En el enunciado de la tabla hay los parámetros necesarios para realizar la práctica; fijaros que varian entre intentos distintos de responder el cuestionario.

Como entregable, debéis subir al registro de evaluación (REC) un único documento PDF que contenga:

- La resolución de la práctica (memória técnica detallada). **Importante: debéis especificar a qué intento de la tabla corresponde**.
- El código en `R`.
- Las imágenes y/o figuras que se os pidan.

La fecha límite de entrega es a las 23:59 hores del dia 16/06/2023 (CEST).

Recordad que la práctica es **individual**. La detección de falta de originalidad será penalizada de acuerdo con la normativa vigente de la UOC. Además, comprovad que el archivo subido es el correcto, ya que es responsabilidad del alumnado hacer la entrega correctamente.

No se aceptarán entregas fuera de plazo ni en formatos que no sean los especificados.

1. Estudio de las desigualdades en tu localidad.

Cuando se acerca un año electoral, los partidos políticos buscan añadir analistas de datos a su equipo para que los ayuden a estudiar la situación actual del pueblo o ciudad donde se presentan. Con su ayuda, podrán diseñar un programa electoral ceñido a la realidad que viven día a día los vecinos y vecinas de su localidad.

En esta práctica, nos imaginaremos que vosotros sois los candidatos o candidatas a la alcaldía de vuestra localidad. Debido a que las diferentes crisis que ha habido en este siglo (económica, social, sanitaria) han afectado de forma diferente a las familias, queréis estudiar las desigualdades presentes en vuestra localidad. Por eso, se os proporcionan las siguientes variables agregadas para cada sección censal (división de la localidad por lugar de votación).

- **id**: identificador de la sección censal.
- **rent**: renta bruta por persona.
- **inc_sal**: ingresos provenientes del salario.
- **inc_ret**: ingresos provenientes de pensiones de jubilación.
- **inc_emp**: ingresos provenientes de prestaciones del paro.
- **inc_non**: ingresos provenientes de otros tipos de prestaciones.
- **inc_oth**: otros ingresos.
- **gini**: coeficiente de Gini que mide la desigualdad.
- **dist8020**: relación de renta entre el percentil 80 (P80) y el percentil 20 (P20) – P80/P20.
- **mean_age**: edad media de la población.
- **perc_chil**: porcentaje de población menor de 18 años.
- **per_ret**: porcentaje de población mayor de 65 años.
- **home_size**: tamaño medio del hogar (m^2).

Estos indicadores se encuentran en el fichero *variables.csv*, que podéis leer de la siguiente manera:

```
> var_df <- read.csv('/home/variables.csv')
```

Como tenemos múltiples indicadores de diferente índole (económicos, sociales, demográficos), utilizaremos el análisis de componentes principales para poder agrupar y mezclar estos indicadores en un menor número de variables que relacionen distintas características de cada sección censal.

Antes de empezar, **tenéis que abrir la "Tabla resumen de la Práctica 1"** del Moodle. Allí, encontraréis el valor de los parámetros C y V necesarios para poder realizar la práctica. Recordar, también, que deberéis indicar los valores utilizados en el inicio de la memoria, así como el intento al que corresponden las respuestas entregadas.

1. [10 %] Para empezar, generar la matriz `X` a partir del *dataframe* `var_df` utilizando, por ejemplo, la instrucción `as.matrix`. Aseguraros de eliminar el valor de la primera columna ('id') ya que no proporciona ninguna información relevante. Comprobar, también, que `X` es una matriz y no un *dataframe* utilizando la siguiente instrucción:

```
> class(X)
[1] "matrix" "array"
```

   Responder: ¿cuántas secciones censales tiene la ciudad?

2. [10 %] Como alcaldables, os interesa tener una primera impresión general de las variables medidas y explorar los datos en crudo. Una de las características interesantes a estudiar es la razón (M/m) entre el valor máximo (M) y el mínimo (m) de una variable. **Calcular** la razón de la variable V.

3. [15 %] Para poder realizar el análisis de componentes principales debéis, inicialmente, normalizar los datos y guardarlos a la variable `Xs`, como se muestra a la Sección 2.1 de los apuntes del módulo. Una vez normalizados, calcular la matriz de covarianzas de los datos; guardarla en la variable `CXs` y **adjuntarla** [1] como una imagen. Finalmente, **indicar** cuáles son el par de variables (distintas) con mayor covarianza (en valor absoluto) y el par con menor covarianza (en valor absoluto).

4. [5 %] Finalmente, calcular la descomposición en componentes principales de la matriz de covarianzas `CXs`. **Dibujar** la distribución de la varianza explicada en porcentaje (eje de ordenadas) para cada componente principal (eje de abscisas) respecto la variancia total de los datos.

5. [20 %] Como habéis visto, la mayor parte de la varianza queda concentrada en unas pocas componentes principales. Por esto, podemos reducir la dimensión del subespacio, proyectar nuestros datos allí y utilizar estas representaciones para análisis posteriores. Un buen criterio para el diseño del nuevo subespacio es restringir el porcentaje total de varianza explicada por el subespacio a un cierto umbral. En esta práctica, os quedaréis con las L primeras

[1] Para dibujarla, podéis utilizar la instrucción *image.plot()* de la librería *fields*

componentes principales que expliquen, al menos, un 75 % de la varianza inicial. **Calcular** el valor mínimo de L, es decir, el mínimo número de componentes principales necesarias para explicar un 75 % de la varianza de nuestros datos.

6. [10 %] Considerar la componente principal C e **indicar** qué variables contribuyen en mayor y menor peso (en valor absoluto).

7. [10 %] Calcular las nuevas variables proyectadas a las componentes principales. Para la componente principal C, **anotar** las secciones censales (relacionarlo con la variable id) con el valor máximo y mínimo.

8. [20 %] Cuando reducimos la dimensión del subespacio generado por los datos iniciales a L, se produce una pérdida de información. Una manera de medir el error cometido en esta aproximación es calculando el error residual, tal y como se indica en la Sección 2.5.1 de los apuntes del módulo. Considerando el valor de L calculado en el apartado 5, **calcular** la desviación típica del error residual cuando se consideran solo las L primeras componentes principales.

Una vez hecha la descomposición por componentes principales, podéis realizar múltiples tipos de análisis de interés: clasificación, regresión o predicción, entre otros.