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 horas del dia 03/01/2025 (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. Predicción meteorológica

En la televisión pública de vuestro país quieren relevar al meteorólogo de cabecera para calcular y presentar la predicción meteorológica en *prime time*. Después de un duro proceso de selección os acaban seleccionando y hoy es el primer día de trabajo.

Como especialistas en ciencia de datos, lo primero que queréis hacer es analizar períodos históricos temporales para observar los diferentes patrones y ver si se replican en el tiempo. Estáis interesados en conocer cuáles han sido las variables más relevantes para la predicción a lo largo del tiempo. Para realizar esta tarea, os proporcionan las observaciones diarias (a las 12 del mediodía) de diferentes variables relacionadas con la meteorología durante un período de 3 años (2006-2008).

- *weather_label*: el tiempo del día (nublado: 0, lluvioso: 1, soleado: 2).
- *temperature*: temperatura (en grados centígrados).
- *temp_app*: sensación térmica (en grados centígrados).
- *humidity*: humedad relativa (en tanto por uno [0-1]).
- *vind_vel*: velocidad del viento (en kilómetros por hora).
- *vind_dir*: dirección del viento (en grados).
- *visibility*: visibilidad (en kilómetros).
- *atm_pres*: presión atmosférica (en milibares).

Una librería que os puede ser útil para realizar la práctica es `fields`, como veréis más adelante. Recordar que debéis instalarla una sola vez y luego importarla en el código.

Antes de empezar, **debéis abrir la "Tabla resumen de la Práctica 1"** del Moodle. Allí, encontraréis el valor de los parámetros (T, E) para poder realizar la práctica. Recordar, también, que debéis indicar los valores utilizados al inicio de la memoria, así como el intento de la Tabla correspondiente (primero o segundo).

1. [10 %] Primeramente, leer el fichero de datos correspondiente al período 2006-2008.

```
> data_df <- read.csv('/home/data_0608.csv')
```

   De la tabla resultante, guardar la primer columna (*weather_label*) al vector `y` y las otras columnas a la matriz de características `X`. **Responder:** ¿qué dimensión tiene la matriz `X`?

2. [10 %] Antes de realizar cualquier tipo de análisis, es importante hacer una exploración estadística (cuantitativa y cualitativa) de los datos. Para este propósito, **observar** el número de días con tiempo T y **calcular** su temperatura media (sólo de los días correspondientes a T)

3. [10 %] Para poder aplicar la descomposición en componentes principales, debéis normalizar la matriz de datos `X` siguiendo los criterios de la Sección 2.1 de los apuntes del módulo. Para hacerlo, debéis calcular la media y la desviación típica de los datos; guardar ambas en las variables `m_X` y `s_X`, respectivamente, ya que las necesitaréis más adelante. Nombrar a la nueva matriz de datos normalizada `Xs`. Una vez hecho, **indicar** la temperatura media de todo el período 2006-2008 de los datos normalizados.

4. [15 %] Para ver la relación cruzada entre las distintas variables, observar la matriz de covarianza `CXs`. **Dibujarla** mediante la instrucción `image.plot()` de la librería `fields` y **contestar**:

   - ¿Qué variable está más asociada a la visibilidad en valor absoluto (y sin que sea ella misma)?
   - ¿Qué variable está menos asociada a la visibilidad en valor absoluto?

5. [15 %] Seguidamente, calcular la descomposición en componentes principales de la matriz de covarianza `CXs`. **Utilizar la instrucción** `eigen` y consultar la documentación si lo necesitáis[1]. Esta función proporciona los valores y vectores propios (componentes principales) de forma ordenada de mayor a menor varianza explicada de los datos originales. Así, la primera componente corresponde a la dirección de máxima varianza mientras que la última componente corresponde a la dirección de mínima varianza. **Dibujar** la distribución de la **varianza acumulada** (eje de ordenadas) para cada componente principal (eje de abscisas) respecto a la varianza total de los datos. Seguidamente, **indicar** el número mínimo de componentes necesarios P para explicar un E % de la varianza inicial de los datos.

6. [10 %] Con el tiempo, se os encarga un nuevo estudio, ahora durante el período 2009-2011. Como suposición inicial, considerar una distribución estacionaria de los datos, eso es, que sus propiedades estadísticas son constantes en el tiempo. Esto os permite utilizar la media y desviación típica anteriormente calculadas así como las componentes principales del período 2006-2008.

   Empezar leyendo los datos del nuevo período y normalizarlos usando la media y desviación típica previamente calculadas (apartado 3); guardar los datos normalizados a la matriz `Xs_test`. Una vez hecho, **contestar:** ¿cuántos días de T habéis observado en este segundo período?

[1] https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/eigen

7. [15 %] Utilizando las P primeras componentes principales calculadas anteriormente, proyectar los datos del período 2009-2011 (normalizados) al nuevo subespacio. Guardar dicha proyección a la variable `Xproj_test`. Recordar que este subespacio tiene dimensión P. **Responder:** ¿qué proporción (en porcentaje) de la varianza inicial de los datos explican los datos del subespacio, `Xproj_test`? ¿Es mayor o menor que la obtenida en el período 2006-2008?

8. [15 %] Finalmente, a partir de la proyección `Xproj_test` queréis recuperar los datos observados tal y como se indica a la Sección 2.5 y 2.5.1 de los apuntes del módulo. **Calcular** el error de reconstrucción y **responder:** ¿cuál es la desviación típica del error de reconstrucción de la temperatura? ¿Creéis que la suposición de una distribución estacionaria de los datos es correcta?