Estadística Multivariante

Práctica Evaluable: Aprendiendo de los datos

Juan Antonio Villegas Recio

13 de enero de 2022

Resumen

A partir de una base de datos con información variada de una serie de países, se pretende realizar un análisis exploratorio de los datos y descubrir posibles relaciones entre ellos. Contamos con información como la densidad de población, el porcentaje de población activa, población urbana o libros publicados en cada país. Las variables con las que contamos de hecho son muy diversas en cuanto a contexto, por ello interesa buscar posibles relaciones entre variables que aparentemente no tienen nada que ver.

Para ello se han preprocesado los datos mediante la corrección de valores perdidos y valores extremos y utilizado técnicas de análisis univariante como tests de normalidad, gráficos *box-plot* o *qqplot*; multivariantes como análisis de componentes principales o análisis factorial junto con herramientas de aprendizaje automático como análisis de clusters. En conjunto, se busca con todas estas técnicas extraer la información de interés humano subyacente a este conjunto de datos, buscando con qué variables obtener la mayor cantidad de información, cuáles son más importantes, y posibles variables latentes subyacentes.

1. Introducción

Nuestro conjunto de datos cuenta con un total de 34 instancias, la cual cada una representa un país del mundo. Hay representación de países de cada uno de los continentes, con la variedad que eso implica a todos los niveles. El conjunto de datos cuenta además con 12 variables de las cuales 11 son numéricas y la restante es categórica, el nombre del país, por lo que claramente hay tantos valores como instancias y esta variable será excluida de los cálculos por tratarse principalmente de un identificador. Las variables numéricas están ya normalizadas con media 0 y varianza 1, teniendo así de antemano solucionados posibles problemas relacionados con las distintas escalas. Estas variables son:

- **Ztlibrop**: Número de libros publicados.
- **Ztejerci**: Cociente entre el número de individuos en ejército de tierra y población total del estado.
- **Ztpobact**: Cociente entre población activa y total.
- **Ztenergi**: Tasa de consumo energético.
- **Zpservi**: Población del sector servicios.
- **Zpagricu**: Población del sector agrícola.

- **Ztmedico**: Tasa de médicos por habitante.
- **Zespvida**: Esperanza de vida.
- **Ztminfan**: Tasa de mortalidad infantil.
- **Zpobdens**: Densidad de población.
- **Zpoburb**: Porcentaje de población urbana.

Como vemos, hay variables muy distintas pertenecientes a varios contextos diferentes: medicina, esperanza de vida, población urbana, población activa, consumo, etc. Esta diversidad nos sugiere estudiar posibles agrupaciones de variables inesperadas. Para ello se ha hecho uso de técnicas como el análisis factorial. También podemos buscar reducir la dimensión conservando la mayor parte de la información, y para ello aplicamos análisis de componentes principales, en ambos casos cerciorandonos previamente de los requisitos que las técnicas requieren. Por último, se ha aplicado una conocida técnica de aprendizaje automático no supervisado: el clustering, buscando agrupaciones entre países que puedan ser significativas.

Por tanto, buscamos estudiar a fondo los datos que se nos dan, apreciando la distribución de cada variable por separado, para posteriormente estudiar la distribución multivariante. Se pretende buscar además agrupaciones de variables relacionadas junto con agrupaciones de instancias con características similares, haciendo así por tanto un estudio de variables relacionadas y de instancias con rasgos comunes en lo que a las variables con las que contamos respecta.

2. Materiales y Métodos

2.1. Materiales

2.2. Métodos estadísticos

3. Resultados

4. Discusión

5. Conclusión

Referencias

Author, A.N and Another, A. N., 2010, MNRAS, 431, 28.