Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Cuadrados Mínimos Lineales

Introducción

La expectativa de vida es un indicador de la calidad de vida de los habitantes de un país mundialmente utilizado. Políticas públicas a largo plazo podrían impactar negativa o positivamente en su valor, lo que vuelve fundamental entender su relación con otras variables estadísticas. Enfermedades, indicadores económicos, educación podrían relacionarse con la expectativa de vida. El objetivo de este trabajo práctico es la realización de un análisis de datos que permita entender mejor que características de los países se relacionan con la expectativa de vida de sus habitantes. Se debe utilizar, como método central, el de cuadrados mínimos lineales (regresión lineal en la literatura estadística).

Se requiere la realización de un análisis exploratorio de datos, buscando describir cualitativa y cuantitativamente las características del dataset proporcionado. Este primer paso será la base sobre la cual, deberán luego experimentar con diversas variantes de regresión, a los efectos de intentar explicar la expectativa de vida por país.

Dataset

Se proporcionarán dos archivos: expectativa_de_vida.csv y expectativa_de_vida_descripcion.txt que son el resultado de un preprocesamiento realizado por la cátedra, de datos públicos de la Organización Mundial de la Salud (WHO). El primero contendrá los datos a utilizar mientras que el segundo contendrá un descripción de los mismos.

Análisis exploratorio de datos

En todo trabajo relacionado con datos, lo primero que hay que hacer es lograr algún entendimiento de la estructura de los mismos. Piensen en un lector que va a trabajar con estos datos por primera vez: esta sección del trabajo práctico debería ser una ayuda para esa persona. Como mínimo, deberán reportar los tamaños del dataset, los datos faltantes, la distribución de cada feature, recalcar desbalances que pudiera haber, la correlación de a pares entre las columnas. ¿Qué tipo de dato tiene cada columna? ¿son variables categóricas o numéricas, están acotadas, son porcentajes? Tienen outliers (pueden usar la técnica de la distancia intercuantil). Además, piensen en mapas, categorías de países. ¿Qué características tienen aquellos con mayor expectativa de vida? ¿Y los de menor? Por supuesto, utilicen gráficos, así como también funciones ya existentes en el ecosistema de las librerías de Python utilizadas en la materia.

Relacionando expectativa de vida con regresores

Se debe describir el proceso para llegar a uno o varios modelos de regresión donde el target Y sea la expectativa de vida, y los predictores $X_1, ..., X_j$ sean features del dataset o nuevos incorporados por el grupo. El objetivo del trabajo práctico es entender como distintas variantes

de cuadrados mínimos se comportan sobre un dataset. Deberán ir iterando el modelo, agregando o quitando variables. Para ello, pueden nombrar las versiones (V1, V2, etc). Deberán justificar, en cada iteración, el porqué de dicho movimiento. Deberán experimentar teniendo en cuenta las técnicas de diagnóstico vistas en clase, sumando puntos extra el agregado de otras técnicas de la literatura.

Agregando información

La WHO provee estadísticas anuales de muchísimos indicadores que podrían ayudar a explicar la expectativa de vida[1]. Deberán incorporar información de alguno de esos datatasets, posiblemente completando valores faltantes en los mismos. Utilicen datos promedios de 2015 a la fecha. Recomendamos no incorporar datos por año, dado que podría complejizar el problema debido a la correlación entre los años.

Entrega

En todos los casos es **obligatorio** fundamentar los experimentos planteados, proveer los archivos e información necesaria para replicarlos, presentar los resultados de forma conveniente y clara, y analizar los mismos con el nivel de detalle apropiado. El código para cuadrados mínimos debe ser implementado en C++.

- Formato Electrónico: Domingo 14 de Noviembre de 2021, hasta las 23:59 hs, enviando el trabajo (informe + código) a la dirección `metnum.lab@gmail.com`. El subject del email debe comenzar con el texto [TP3]-Número de grupo- seguido de la lista de apellidos de los integrantes del grupo. Además, el contenido debe entregarse en un archivo .zip con el número de grupo como nombre, por ejemplo, grupo_1.zip.

Nota: el presente trabajo práctico es intencionalmente más abierto en su consigna que los anteriores. Se pretende que el grupo investigue y complemente las técnicas vistas en clase con la literatura. Recomendamos también complementar la interpretación que puedan hacer de los datos, con información de dominio que puedan encontrar.

Importante: El horario es estricto. Los correos recibidos después de la hora indicada no serán considerados.

[1] https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates