Análisis de datasets

Alejandra Campo Archbold
Universidad del Rosario
Semillero de Machine Learning

11 de Marzo de 2019

1. Suicide Rates Overview 1985 to 2016

Es un conjunto de datos sobre los casos de suicidios de distintos paises del mundo por año, desde 1985 hasta el 2016. Dentro de sus variables, está el año, el país, los casos de suicidios por sexo y rango de edad, suicidios por cada 100.000 habitantes, país-año, Índice de Desarrollo Humano (HDI), PIB por año, PIB per cápita por año y el tipo de generación.

1.1. Regresión lineal:

De acuerdo a los datos, se puede plantear un modelo de regresión lineal donde la variable dependiente Y sea suicidios por cada 100.000 habitantes y como variable independiente X la variación del PIB per cápita, donde la pendiente B_1 de este modelo indica ¿cuánto cambia la tasa de suicidios (Y) cuando la variación del PIB per cápita (X) aumenta una unidad?

$$Y = B_0 + X_1 B_1 + u_i \tag{1}$$

Donde luego se analizaría las propiedades estadísiticas y plantearse la pregunta ¿tiene una influenza económica en las tasas de suicidio?

1.2. Regresión lineal múltiple:

El modelo de regresión múltiple extiende el modelo de regresión simple (con regresor único X_1) para incluir variables adicionales como regresores. Puede ser que no sólo el PIB per cápita sea variable explicativa de la tasa de suicidios por cada 100.000 habitantes pueden entrar otras variables como el Índice de Desarrollo Humano, una variable dummy si es mujer u hombre, otra variable dummy si la persona es de un tipo de generación. Esto con el fin de hacer los análisis más robustos. Luego, encontrar inferencias estaídisticas, intervalos de confianza, entre otros métodos.

1.3. Problemas de la regresión lineal y otros métodos:

En los datos encontramos una variable cualitativa como son el tipo de generación, la cual está la generación G.I, los Boomers, Silent, Generación X, Millenials y Generación Z, la definición de cada una de estas generaciones son disntintas ya que son de acuerdo al año o la época de nacimiento y presentan características distintas de acuerdo al contexto social, económico, político, cultural.

Una de las principales causas de suicidio, se ha pronunciado por enfermedades mentales, deperesión, rupturas amorosas, entre otros. Para el análisis de las tasas de suicidio, es posible que los casos de muerte se dan más a menudo de acuerdo al tipo de generación. Plantear una regresión lineal donde la variable dependiente contenga binarios (Y):

$$Y = \begin{cases} 1, & \text{si es Generación X;} \\ 2, & \text{si es Millenials;} \\ 3, & \text{si es Generación Z.} \end{cases}$$

Usando esta codificación, los mínimos cuadrados podrían usarse para ajustar un modelo de regresión lineal para predecir Y sobre la base de un conjunto de predictores $X_1, ..., X_p$.Esta codificación implica un ordenamiento de los resultados, colocar una Millenials entre Generación X y Generación Z, y decir que la diferencia entre Generación X y Millenials es la misma que la diferencia Millenials y Generación Z. Pueden tener cambios totalemente distintos de acuerdo a la definición de generación.Por otro lado, no existe una forma natural de convertir una variable de respuesta cualitativa con más de dos niveles en una respuesta cuantitativa que esté lista para la regresión lineal.

La pregunta para este caso: ¿tiene estas generaciones la misma probabilidad de incidir en las tasas de suicidio? Puede que las estimaciones de los parámetros de las generaciones sobrepasen el rango de probabilidad [0,1], para este caso es preferible utilizar métodos de clasificación que sea verdaderamente adecuado para los valores de respuesta cualitativos.

1.3.1. Clasificación:

Se puede utilizar los siguientes métodos para lograr los análisis que queremos del datasets teniendo en cuenta los problemas de las regresiones lineales:

- Regresión logística: es adecuada cuando la variable de respuesta Y es politómica (admite varias categorías de respuesta), ajusta las probabilidades de estas variables para una estimación más verosímil.
- Análisis discriminante lineal: Lo usamos cuando las estimaciones de los parámetros para el modelo de regresión logística son inestables. El análisis discriminante lineal no sufre de este problema.
- Usando Teorema de Bayes para la Clasificación: Para este caso queremos clasificar una observación en una de las clases K, donde $K \geq 2$, la variable dependiente cualitativa Y puede tomar K posibles valores distintos y desordenados. Luego, con el análisis de la fórmula del Teorema de Bayes podemos encontrar respuestas para el ajuste de las variables.

1.4. Referencias

Suicide Rates Overview 1985 to 2016. 2018.
Recuperado de: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

Students Performance in Exams

Para este datasets se busca comprender los resultados marcados en los exámenes de los estudiantes de Estados Unidos en las áreas de matemática, lectura y escritura de acuerdo al grupo de raza perteneciente, el almuerzo y el nivel de educación de los padres. Se pretende analizar cuál es el desempeño de un estudiante de acuerdo a su género, raza y nivel de educación de los padres de acuerdo a las áreas de matemáticas, lectura y escritura.

2.1. Análisis:

De acuerdo a las definiciones en la sección anterior, para este dataset podemos analizar los siguientes métodos estadísticos y de clasificación:

- Regresión lineal múltiple: como variable dependiente Y podemos analizar el desempeño del estudiante y como variables explicativas $(X_1, ..., X_p)$ la raza, el género y el resultado en matemáticas como ejemplo.
- En este datasets también podemos analizar métodos de clasificación donde la variable Y es politómica con la raza. De acuerdo a la base de datos encontramos grupo A, B, C y D, y analizar de acuerdo a desempeño en matemáticas por ejemplo.

2.2. Referencias

Students Performance in Exams. 2018. Recuperado de:
https://www.kaggle.com/spscientist/students-performance-in-exams

Garbage classification

En este datasets, encontramos conjunto de imágenes de un tipo de basura para hacer clasificaciones de éste en el procesamiento de reciclado. Dentro del conjunto de imágenes encontramos cartón, papel, vidrio, metal, plástico y basura sin identificación como tal.

Para este caso el método de análisis es la clasificación en teledetección:

Es un caso particular del problema general de clasificar N individuos en un conjunto de K clases en función de una serie de variables cuantitativas $(X_1, ..., X_p)$. Para resolver este problema se necesita una medida de la semejanza o diferencia entre los diferentes individuos y entre los individuos y las clases. Dos individuos muy parecidos pertenecerán probablemente a la misma clase, mientras que dos individuos distintos pertenecerán a diferentes clases.

3.1. Clasificación no supervisada

En la clasificación no supervisada no se establece ninguna clase a priori, aunque es necesario determinar el número de clases que queremos establecer,

y se utilizan algoritmos matemáticos de clasificación automática. Los más comunes son los algoritmos de clustering que divide el espacio de las variables en una serie de regiones de manera que se minimice la variabilidad interna de los pixeles incluidos en cada región.

El objetivo del método de clasificación consiste en minimizar las desviaciones entre las observaciones que pertenecen al mismo grupo y maximizar las distancias entre los centros de los grupos

3.2. Clasificación supervisada

La clasificación supervisada se basa en la disponibilidad de áreas de entrenamiento. Se trata de áreas de las que se conoce a priori la clase a la que pertenecen y que servirán para generar una signatura espectral característica de cada una de las clases.

3.3. Referencias

https://www.um.es/geograf/sigmur/temariohtml/node74.html

Garbage Classification. 2018. Recuperado de:
https://www.kaggle.com/asdasdasasdas/garbage-classification

Bibliografía

James, G. Witten, D. Hastie, T. and Tibshirani, R. An Introduction to Statistical Learning. 2014. Recuperado de: http://www-bcf.usc.edu/ gareth/ISL/