Air Pollution in Seoul - Logistic Regression

Sergio Alejandro Paucara Saca
Universidad Mayor de San Andres
sergiopaucara@gmail.com

***Resumen*—En base a un conjutno de datos obtenida de la comunidad en linea de cientificos de datos y profesionales del aprendizaje automático Kaggle, se aplico un algoritmo supervisado de Regresion Logistica para poder predecir el nivel de contaminacion.**

***Palabras Clave*—Regresion Logistica, Kaggle, Conjunto de Datos, Contaminacion**

***Abstract*—Based on data collected from Kaggle's online community of data scientists and machine learning professionals, a supervised Regresion Logistics algorithm was applied to predict the level of contamination.**

***Index Terms*—Logistic Regression, Kaggle, Dataset, Contamination**

I. Introducción

Es de conocimiento común que la contaminación del aire puede causar varios problemas en el medio ambiente y en nuestra salud. La foto de arriba fue tomada el 11 de diciembre de 2019 y muestra cómo puede impactar severamente los paisajes de Seúl. En esta ocasión, un smog de polvo ultrafino, proveniente desde China, duró dos días e hizo que el gobierno local dictara medidas de emergencia para la reducción de emisiones. Según The Korea Times, el Centro de Pronóstico de la Calidad del Aire, con el Ministerio de Medio Ambiente, informó que el 11 de diciembre a las 10 p.m la concentración de partículas PM2.5 era de aproximadamente 118 microgramos por metro cúbico en Seúl.
Para empezar, suponga que hay una unica variable explicativa X, que es cuantitativa. Para una variable de respuesta binaria Y, recuerde que $\pi(x)$ denota la probabilidad de ëxitoën valor x. Esta probabilidad es el parámetro de la distribución binomial. La logistica El modelo de regresión tiene forma lineal para el logit de esta probabilidad

$$logit[\pi(x)] = \log\left(\frac{\pi(x)}{1-(x)}\right) = \alpha + \beta x \tag{1}$$

Para x cuantitativo, la fórmula implica que $\pi(x)$ cambia como una función en forma de S de x. La regresión logística tiene una fórmula correspondiente para $\pi(x)$, utilizando el función exponencial $\exp(\alpha+\beta x) = \exp^{\alpha+\beta x}$

$$\pi(x) = \frac{\exp^{\alpha+\beta x}}{1+\exp^{\alpha+\beta x}} \tag{2}$$

El parámetro de efecto β determina la tasa de aumento o disminución de la curva en forma de S para $\pi(x)$. El signo de β indica si la curva asciende ($\beta > 0$) o desciende ($\beta < 0$). La tasa de cambio aumenta cuando $|\beta|$ aumenta. Cuando $\beta = 0$, la curva se aplana a una horizontal línea recta. La variable de respuesta binaria es entonces independiente de la variable explicativa.

I-A. Interpretaciones de la razón de probabilidades y la aproximación lineal

La fórmula de regresión logística (1) indica que el logit aumenta en β por cada 1 unidad aumento en x. La mayoría de nosotros no pensamos de forma natural en una escala logit, por lo que a continuación sugerimos alternativas interpretaciones. Al exponenciar ambos lados de la ecuación de regresión logística (1), obtenemos una interpretación que utiliza las probabilidades y la razón de probabilidades. Las probabilidades de éxito son

$$\frac{\pi(x}{1-\pi(x)} = \exp(\alpha+\beta x) = e^{\alpha}(e^{\beta})^{x}$$

Por lo tanto, las probabilidades se multiplican por e^{β} por cada aumento de 1 unidad en x. Es decir, las probabilidades a nivel $x+1$ es igual a las probabilidades en x multiplicadas por $e^{b}eta$. Cuando $\beta = 0$, $e^{\beta} = 1$ y las probabilidades no cambiana medida que x cambia.

Una interpretación más simple se refiere a la probabilidad $\pi(x)$ en sí. La figura 1. muestra la apariencia en forma de S del modelo para $\pi(x)$, como se ajusta para el siguiente ejemplo. Ya que es curvo en lugar de una línea recta, la tasa de cambio en $\pi(x)$ por 1 unidad de aumento en x depende de la valor de x.
Una línea recta trazada tangente a la curva en un valor de x particular, como se muestra en la Figura 1, describe la tasa de cambio en ese punto. Para el parámetro de regresión logística β, esa línea tiene una pendiente igual a $\beta\pi(x)[1-\pi(x)]$. Por ejemplo, la recta tangente a la curva en x para lo cual $\pi(x) = 0{,}50$ tiene pendiente $\beta(0{,}50)(0{,}50) = 0{,}25\beta$; por el contrario, cuando $\pi(x) = 0{,}90$ o $0,10$, tiene pendiente $0,09\beta$. La pendiente se acerca a 0 cuando $\pi(x)$ se acerca a 1.0 o 0. La pendiente más pronunciada la pendiente ocurre cuando $\pi(x) = 0{,}50$. Ese valor x se relaciona con los parámetros de regresión logística por $x = \alpha/\beta$. Este valor de x a veces se denomina nivel medio efectivo. Representa el punto en el que cada resultado tiene un $50\,\%$ de probabilidad.

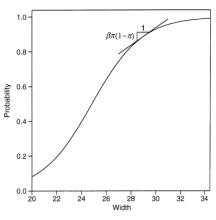

Figura 1. Linear approximation to the logistic regression curve

I-B. Regresion Logistica Multiple

A continuación, consideraremos el modelo de regresión logística general con múltiples explicaciones. variables. Denote los k predictores para una respuesta binaria Y por $x_1, x_2, ..., x_k$. Elmodelo para las probabilidades de registro es

$$logit[P(Y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

El parámetro β_i se refiere al efecto de x_i en las probabilidades de registro de que $Y = 1$, controlando los otros xs. Por ejemplo, $\exp(\beta_i)$ es el efecto multiplicativo sobre las probabilidades de una unidad de 1 aumento en x_i, a niveles fijos de los otros xs

II. Objetivos

II-A. Objetivo General

En base ha hechos se desea predecir como sera la contaminación dentro una ciudad, en este caso el conjunto de datos nos proporciona la información de medición de la contaminación del aire en Seul, Corea del Sur.

II-B. Obejtivo Especifico

Aplicar el modelo de algoritmo de regresión logistica para predecir el nivel de contaminación en base a seis contaminantes de un conjunto de datos.

III. Materiales y Métodos

El conjunto de datos tabulados que utilizaremos lo obtuvimos de Kaggle, cuenta con las siguientes columnas:

- Measurement date = Hora y fecha de cuando se midio
- Station code = Codigo de Estacion
- Address = Direccion donde se obtuvieron los datos
- Latitude = Latitud
- Longitude = Longitud
- SO_2 = El dióxido de azufre, u óxido de azufre
- NO_2 = Dióxido de nitrógeno
- CO = Monóxido de carbono
- O3 = Ozono
- PM_10 = Pequeñas partículas sólidas o líquidas de polvo, cenizas, hollín, partículas metálicas, cemento o polen, dispersas en la atmósfera
- PM_2,5 = Pequeñas partículas que incluye sustancias químicas orgánicas, polvo, hollín y metales.
- class = Interpretacion si la contaminacion es:
  - Good : Es normal
  - Bad: Es mala

Una vez obtenidos los datos empezamos con el preprocesamiento de los mismo, se aplico cuatro tipos diferentes de preprocesamiento donde dos fueron para hacer la prueba y los otros dos los utilizamos para el desarrollo del proyecto.

III-A. Preprocesamiento

III-A1. Preprocesamiento Imputer: Nos sirve como un transformador de imputación para completar valores perdidos. En caso del data set no fue necesario.

III-A2. Preprocesamiento MinMaxScaler: Nos sirve para que transforme las características escalando cada característica a un rango determinado.
Este estimador escala y traduce cada característica individualmente de modo que esté en el rango dado en el conjunto de entrenamiento, p. Ej. entre cero y uno

III-A3. Preprocesamiento StandardScaler: Estandarice las características eliminando la media y escalando a la varianza de la unidad, este fue una de los preprocesamientos que se realizo para una mejor presicion de datos a la hora de aplicar el clasificador.

III-A4. Preprocesamiento LabelEncoder: Nos permite codificar las etiquetas de destino con un valor entre 0 y $n\_classes - 1$. Se aplico para los datos de salido o target, ya que se tenia datos categóricos.

III-B. Clasificador - Regresion Logistica

El parámetro β_j se refiere al efecto de x_j en las probabilidades de registro de que $Y = 1$, ajustando los de los otros $x's$. Por ejemplo, $\exp(\beta_1)$ es el efecto multiplicativo sobre las probabilidades de un aumento de 1 unidad en x_1, a un valor fijo para $\beta_2 x_2 + \cdots + \beta_p x_p$, como cuando podemos mantener constante $x_2, ..., x_p$.
El modelo de regresión logística general tiene múltiples variables explicativas que pueden ser cuantitativas, categóricas o ambas. Para las variables explicativas de p, el modelo para las probabilidades logarítmicas es

$$logit[P(Y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ + \beta_p x_p$$

En la primera iteracion de entrenamiento con un margen academico de datos de testeo y entrenamiento, se obtuvo la siguiente matriz de confusion y presicion del modelo

$$ConfusionMatrix = 629132931290118628$$

$$Exactitud\_Modelo = 0{,}96461$$

Posteriormente se realizaron 100 splits sobre el conjunto de datos con $80\,\%$ de los datos para entrenamiento y $20\,\%$ de los datos para testeo y verificar la exactitud del modelo, luego de realizar los splits la mediana de la exactitud del modelo dio:

$$Mediana\_Exactitud\_Academico = 0{,}96455$$

Finalmente se realizaron 100 splits sobre el conjunto de datos, pero ahora en forma de investigacion con $50\,\%$ de los datos para entrenamiento y $50\,\%$ de los datos para testeo y verificar la exactitud del modelo, luego de realizar los splits la mediana de la exactitud del modelo dio:

$$Mediana\_Exactitud\_Investigacion = 0{,}96450$$

III-C. Principal Component Analysis

Principal Component Analysis o PCA siempre se puede utilizar para simplificar los datos con grandes dimensiones (mayores de 2) en datos bidimensionales eliminando las características menos influntiales de los datos. Sin embargo, debemos saber que la eliminación de datos hace que la variable independiente sea menos interpretable. Antes de comenzar

a tratar con el PCA, primero debemos aprender cómo el PCA utiliza los vectores propios para obtener una matriz de covarianza de diagonalización.

Aquí usamos una matriz simple (2x2) A para explicarlo.

$$A = 1432$$

En general, el vector propio v de una matriz A es el vector donde se cumple lo siguiente:

$$Av = \lambda v$$

para el cual λ representa el autovalor tal que la transformación lineal en v puede definirse mediante λ Además, podemos resolver la ecuación de la siguiente manera:

$$Av - \lambda v = 0v(A - \lambda I) = 0$$

Mientras que I es la matriz identidad de A

$$I = A^T A = AA^T$$

En este caso, si v es un vector sin cero entonces $Det(A - \lambda I) = 0$, ya que no puede ser invertible, y podemos resolver v para A depende de esta relación.

$$I = 1001$$

$$(A - \lambda I) = 1 - \lambda 432 - \lambda$$

Para resolver el λ podemos usar la función resolver en sympy o calculando. En este caso, $\lambda_1 = -2$ y $\lambda_2 = 5$, y podemos calcular los vectores propios en dos casos. Por $\lambda_1 = -2$ Con base en la matriz, podemos inferir que el vector propio puede ser

$$v_1 = -43$$

Por $\lambda = 5$ Con base en la matriz, podemos inferir que el vector propio puede ser

$$v_2 = 11$$

Con todo, la matriz de covarianza A' ahora puede ser:

$$A' = v * A$$

De tal manera que podamos obtener la matriz V

$$V = -4131$$

donde $A' = V^{-1}AV$

Luego de aplicar Principal Component Analysis sobre el conjunto de datos mas de 5 veces vemos que los resultamos no cambian mucho en cuando se toma casi todas las componenetes, pero a medida que se van eliminando o se toma menos componentes la exactitud del modelo tiene mayor variabilidad. Por ejemplo cuando se toman 2 componentes la exactitud del modelo es:

$$Exactitud\_PCA_2 = 0{,}92609$$

En cambio cuando se toma casi todas las componentes la exactitud es similar a la mediana que se mostró en resultados anteriores

$$Exactitud\_PCA_8 = 0{,}96317$$

IV. Resultados

La presicion de los datos fue bueno supero nuestras expectativas con un gran porcentaje de exactitud, Vimos como aplicando un buen preprocesamiento o un buen algoritmo supervisado o no supervisado puede hacer que los resultados varien mucho.

V. Conclusion

Despues de ver como funciona la regresion logistica y como fue bueno aplicarlo a un conjunto de datos de contaminacion de aire, es interesante ver donde mas se podria mejorar, la inteligencia artificial esta en pleno desarrollo, a este trabajo aun le falta un desarollo muy profundo de trabajar tal vez las diferentes apliacacion que no vimos, pero se pudo lograr parte de nuestro cometido.

VI. Recomendaciones

El humo es algo muy dañino, aunque muchas veces no le damos importancia al igual que muchas otras cosas juega un rol importante en nuestra vida, es por ello que tenemos que cuidarnos. Obviamente no podemos dar una demanda a todo aquel que bote humo, sin embargo, un proyecto similar o mejorado a este podria decir por camaras quienes son las personas que más contaminan y sancionarlas.

Referencias

[1] Agresti A. *An Introduction to Categorical Data Analysis*, Recuperado de: https://mregresion.files.wordpress.com/2012/08/agresti-introduction-to-categorical-data.pdf

[2] Expower *LOS GASES Y EL HUMO, PRODUCTOS DE LA COMBUSTIÓN*, Recuperado de: http://www.expower.es/humos-gases-combustion.htm

[3] GitHub *Guide to Semantic Segmentation*, Recuperado de: https://codeac29.github.io/projects/linknet/

[4] K, Scott *Air Pollution in Seoul*, Recuperado de: https://www.kaggle.com/bappekim/air-pollution-in-seoul

[5] NHDES *Smoke Pollutants*, Recuperado de: https://www.des.nh.gov/

[6] Wikipedia *Análisis de componentes principales*, Recuperado de: https://es.wikipedia.org/wiki/An%C3%A1lisis_de_componentes_principales

[7] Wikipedia *Regresión logística*, Recuperado de: https://es.wikipedia.org/wiki/Regresi%C3%B3n_log%C3%ADstica

[8] Wikipedia *Humo*, Recuperado de: https://es.wikipedia.org/wiki/Humo