Practica 1
Predicción de rendimiento de gasolina en millas por galón de un auto mediante técnicas de regresión

1st Erick Franco Gaona
Departamento de Estudios Multidisciplinarios
Universidad de Guanajuato
Yuriria, México
e.francogaona@ugto.mx

***Resumen*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

I. Introducción

El análisis de regresión se usa ampliamente para hacer predicciones y estimaciones de expectativas condicionales de variables dependientes e independientes, y sus aplicaciones se superponen con el campo del aprendizaje automático. Ejecutar regresiones permite determinar los factores más importantes, los factores que a menudo se pasan por alto y cómo se influyen entre sí. Estos factores se denominan variables y se clasifican de la siguiente manera:

- Variable dependiente: Es el factor el cual se está tratando de entender o predecir (Y).
- Variable(s) independiente(s): Es el factor que se cree que puede impactar en la variable dependiente (X).

Debido a eso la regresión suele usarse en las organizaciones ya que puede interpretar fenómenos y hacer predicciones sobre el futuro, así como obtener información empresarial valiosa y procesable. Este método proporciona información sobre cómo la estructura de costos y las características variables afectan el producto. Realizar análisis de regresión permite tomar decisiones comerciales más informadas y eficientes y desarrollar estrategias para mejorar la calidad de sus productos y servicios, lo que beneficia a su organización.

La Figura 1 muestra cómo se relaciona el salario con los años de experiencia, prediciendo el salario de un trabajador dado. La regresión muestra una relación lineal positiva entre el salario (eje y) y los años de experiencia (eje x). Puede utilizar estos datos históricos para realizar previsiones salariales. En este trabajo se presentan dos metodos para realizar regresiones y se compara la efectividad de ambos aplicados para la predicción del rendimiento de la gasolina en millas por galón tomando un conjunto de datos público (mpg-auto) de diversos autos.

Figura 1. Ejemplo de regresión aplicado en el salario vs los años de experiencia.

II. Teoría

II-A. Regresión lineal simple

El análisis de regresión lineal simple se utiliza para predecir el valor de una variable en función del valor de otra variable. La regresión lineal se ajusta a una línea recta o a una superficie que minimiza las diferencias entre los valores de salida y los reales. Una forma razonable de relación entre la respuesta Y y el regresor x es la relación lineal

$$Y = \beta_0 + \beta_1 x \tag{1}$$

en la que, β_0 es la intersección de la recta y β_1 es la pendiente.

Si la relación es exacta y no contiene componentes aleatorios o probabilísticos, se trata de una relación determinista entre dos variables. Sin embargo, en muchos casos reales, la relación no es determinista, es decir, una x

dada no siempre produce el mismo valor de Y. Debido a esa situación, los problemas son de naturaleza probabilística, toda vez que la relación anterior no puede considerarse exacta. El concepto de análisis de regresión se refiere a encontrar la mejor relación entre Y y x cuantificando la fuerza de esa relación, y empleando métodos que permitan predecir los valores de la respuesta dados los valores del regresor x. En resumen la regresión lineal simple, trata el caso de una sola variable regresora, en el que la relación entre x y y es lineal.

El método de los mínimos cuadrados se utiliza para calcular la recta de regresión lineal que minimiza los residuos, es decir, las diferencias entre los valores reales y los estimados por la recta. Con este método se debe calcular b_0 y b_1, los estimados de β_0 y β_1, de manera que la suma de los cuadrados de los residuales sea mínima. Los estimados b_0 y b_1 de los mínimos cuadrados de los coeficientes de regresión β_0 y β_1 se calculan mediante las fórmulas

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} \tag{3}$$

II-B. Regresión lineal multiple

En muchas aplicaciones habrá más de un regresor, es decir, más de una variable independiente que ayude a explicar a Y. Por ejemplo, si se tratara de un problema con dos regresores la estructura múltiple de la regresión se podría escribir como

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{4}$$

el modelo de regresión lineal múltiple general se expresa de la siguiente manera

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{5}$$

donde cada coeficiente de regresión β_i se estima por medio de b_i, a partir de los datos muestrales, usando el método de mínimos cuadrados. Si se acomoda en forma matricial se tiene que:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots x_{k1} \\ 1 & x_{12} & \cdots x_{k2} \\ 1 & n & \cdots x_{kn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{6}$$

$$Y = X\beta + \epsilon \tag{7}$$

Para encontrar las β se tiene de forma matricial que

$$\beta = (X^T X)^{-1} X^T Y \tag{8}$$

II-C. Árboles de decisión

Un árbol de decisión es un método de aprendizaje supervisado que predice valores de respuesta aprendiendo reglas de decisión derivadas de características. Se pueden utilizar tanto en contextos de regresión como de clasificación. A diferencia de los modelos lineales, los árboles de decisión pueden capturar interacciones no lineales entre características y objetivos.Los árboles de decisión funcionan dividiendo el espacio de características en múltiples regiones rectangulares simples, divididas por ejes de división paralelos. Para obtener una predicción para una observación en particular, se utiliza la media de las observaciones de entrenamiento en la partición a la que pertenece la nueva observación.

De forma matemática la función para un árbol de regresión se expresa de la siguiente manera:

$$f(x) = \sum_{m=1}^{M} w_m \phi(x; v_m) \tag{9}$$

donde w_m es la respuesta media en una región particular (R_m), v_m representa cómo se divide cada variable en un valor de umbral particular.

De forma visual un árbol de decisión con dos variables de características (X_1 y X_2) y una respuesta numérica "y" se puede observar en la Figura 2. Por otro lado, en la Figura 3 se muestra un subconjunto que contiene el espacio de caracteristicas. El dominio se divide mediante divisiones paralelas de eje, es decir, cada división del dominio se alinea con uno de los ejes de características.

II-D. Bosques aleatorios de decisión

El principal problema con los árboles de decisión es que tienden a sobreajustarse. Puedes crear un árbol que realice regresiones perfectamente con los datos de entrenamiento, pero no en el conjunto de datos de prueba. Para ese problema se crearon los bosques aleatorios de desición. La aplicación repetida del algoritmo de generación de árboles de decisión a los mismos datos con diferentes parámetros produce lo que se denomina un bosque de decisión aleatorio. Este algoritmo es uno de los métodos de pronóstico más eficientes y ampliamente utilizados para big data en la actualidad, promediando múltiples modelos sin ruido ni sesgo para reducir la variación final general.

En la práctica, se construyen diferentes conjuntos de entrenamiento y prueba sobre los mismos datos, la unión de estos árboles de diferentes complejidades y con datos de origen distinto aunque del mismo conjunto resulta un bosque aleatorio. Su principal característica es que produce modelos más robustos que los que se obtienen generando un único árbol de decisión complejo para los mismos datos. El ensamblaje de diferentes modelos (árboles de decisión) produce predicciones más sólidas. Los grupos de árboles de clasificación se combinan y se infiere una predicción a partir de la población de árboles. Mientras haya suficientes árboles en el bosque, hay poco o ningún riesgo de sobreadaptación. Los árboles de decisión también se pueden sobreajustar. Los bosques

aleatorios obtienen esto construyendo árboles de diferentes tamaños a partir de subconjuntos y combinando los resultados.

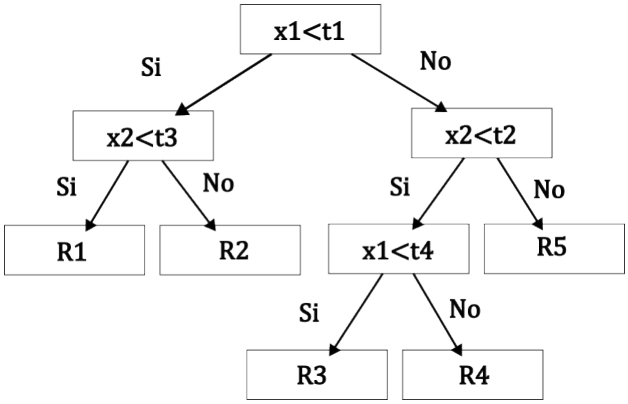

Figura 2. Ejemplo visual de un árbol de decisión.

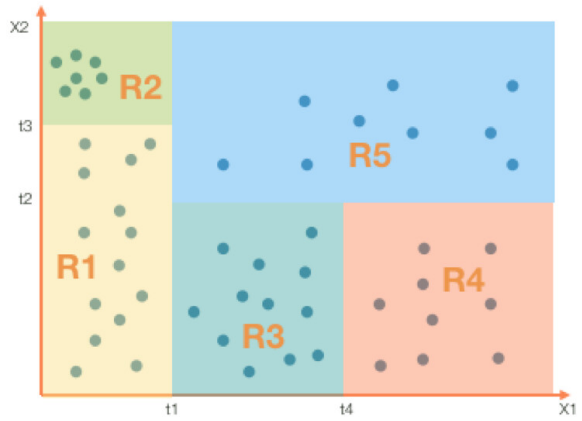

Figura 3. Divisiones de los subconjuntos que realiza un árbol de decisión.

III. Resultados

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections **??–??** below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

Conclusiones

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

Referencias

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

Referencias

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[4] K. Elissa, "Title of paper if known," unpublished.
[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove the template text from your paper may result in your paper not being published.