Universidad Autónoma De Nuevo León

Facultad De Ciencias Físico Matemáticas

Introducción al Aprendizaje Automático

Lic. Ángel Adrián Domínguez Lozano

Tarea 2

Resumen

Equipo 1

Nombre del alumno:	Matrícula:
Humberto Gerardo Peña Páez	1862464
Alberto Natanael Sanchez Robles	1861608

San Nicolás de los Garza, Nuevo León 17 de septiembre de 2021

1. Regresión

La **regresión lineal** es un modelo matemático que se usa para aproximar relaciones de dependencia entre una variable dependiente, n variables independientes y un termino aleatorio. Los supuestos en este modelo son:

- **Independencia**: Los residuos son independientes entre sí, es decir, los residuos constituyen una variable aleatoria
- **Linealidad**: la ecuación de regresión tiene forma de recta, por lo que la variable dependiente constituye la suma de un conjunto de elementos que son; el origen de la recta, una combinación lineal de variables independientes o predictoras y los residuos
- **Homocedasticidad**: Para cada valor de la variable independiente, la varianza de los residuos es constante
- **Normalidad**: Para cada valor de la variable independiente, los residuos se distribuyen normalmente con media cero
- **No colinealidad**: no existe relación lineal exacta entre ninguna de las variables independientes

También existe la **regresión logística**, la cual es el conjunto de modelo estadísticos utilizados cuando se desea conocer la relación entre: una variable dependiente cualitativa ya sea binaria o multinomial, una o mas variables explicativas independientes, llamadas covariables, ya sean cualitativas o cuantitativa. Los supuestos de este modelo son:

- **Linealidad**: en este caso se trata de que existe una relación lineal entre cada variable predictora continua y el logaritmo de la variable respuesta
- **Independencia de los errores**: los distintos casos de los datos no deben estar relacionados.
- **Multicolinealidad**: Las variables predictoras no deben estar altamente correlacionadas

Cuando se trata de una regresión lineal múltiple, pueden surgir diversos problemas, como la incorporación de predictores correlacionados, no realizar selección de predictores o simplemente no poder ajustarse cuando el numero de predictores es superior al numero de observaciones, por lo que algunas estrategias que se pueden aplicar son la regresión **Lasso** y la **Ridge**.

Regresión Lasso (least absolute shrinkage and selection operator): es un modelo de análisis de regresión que realiza selección de variables y regularización para mejorar la exactitud e interpretabilidad del modelo estadístico. Al igual que la regresión lineal, los coeficientes estimados no necesariamente son únicos si las variables independientes son colineales

El modelo tiende a ignorar algunas de las características predictivas, por lo que puede ser considerado un tipo de selección automática de características. Penaliza

la suma del valor absoluto delos coeficientes de regresión, la cual se conoce como L1 y tiene el efecto de forzar a que los coeficientes de los predictores tiendan a cero.

Regresión Ridge: Para este modelo se penaliza la suma de los coeficientes elevados al cuadrado, lo cual se conoce como L2 tiene el efecto de reducir de forma proporcional el valor de todos los coeficientes del modelo, pero sin que estos lleguen a cero. A diferencia de la regresión lasso, a Ridge se incluyen todos los predictores puesto que la penalización fuerza a que los coeficientes tiendan a cero, nunca llegan a ser exactamente cero. Este método consigue minimizar la influencia sobre el modelo de los predictores menos relacionados con la variable respuesta, pero, en el modelo final, van a seguir apareciendo.

2. KNN

Es un algoritmo basado en instancia de tipo supervisado que proviene del inglés K-Nearest-Nieghbor, en español, K-Vecinos-Cercanos donde la "K" hace referencia a los "puntos vecinos". Este puede usarse para clasificar nuevas muestras o para predecir. Esencialmente sirve para clasificar valores buscando los puntos de datos "más similares".

Se aplica comúnmente en la resolución de multitud de problemas, como en sistemas de recomendación, búsqueda semántica y detección de anomalías.

Metricas utilizadas frecuentemente en este algoritmo:

- **Minkowski**.

Es una métrica más compleja que la mayoría utilizada ya que utiliza los espacios vectoriales normados, es decir, espacios reales de n-dimensiones, por lo que puede utilizar espacios donde las distancias se pueden representar como un vector con longitud. Se deben tomar en cuenta los requisitos de un vector normal, es decir, que pueda contener un vector cero, que al multiplicarse un vector con un número positivo su longitud mantenga la dirección y que la distancia más corta entre dos puntos siempre es una línea recta.

Lo interesante de esta métrica es que utiliza el parámetro p, que se puede utilizar para manipular la métrica para que se logren parecer a otras.

Casos de uso.

Poder operar el parámetro p nos da la posibilidad de iterar sobre él y encontrar la medida que funcione mejor. Tiene gran flexibilidad, lo que puede ser tomado como beneficio si se ha trabajado con p muchas medidas de distancia.

$$D(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

- **Hamming**. Compara dos strings de datos binarios.

Esta métrica compara el número de valores que son diferentes entre dos vectores que por lo general contienen datos binarios y son de igual longitud. También se puede utilizar para comparar que tan similares son entre sí calculando la cantidad de caracteres diferentes.

Es difícil de usar cuando los vectores no son de la misma longitud, además no tiene en cuenta el valor real siempre que sean diferentes o iguales.

Casos de uso.

Incluyen corrección/detección de errores. Se puede usar para determinar el número de bits distorsionados en una palabra binaria como forma para estimar el error.

- **Cosine**.

También conocido como la similitud del coseno. Sirve utilizado para orientaciones ya que es simplemente el coseno del ángulo entre dos vectores. Basándose en que, dos vectores con exactamente la misma orientación tienen similitud de coseno de 1, mientras que si son opuestos entre sí tienen una similitud de -1. Entre menor sea el ángulo, más similares serán.

Casos de uso

Se utiliza cuando tenemos datos de alta dimensión y cuando la magnitud de los vectores no es importante.

$$D(x,y) = cos(\theta) = \frac{x \cdot y}{||x||||y||}$$

- **Haversine**.

Esta métrica es la distancia entre dos puntos en una esfera dadas sus longitudes y latitudes. Similar a la distancia euclidiana, ya que las dos basan en calcular la línea más corta entre dos puntos, siendo la principal diferencia que en Haversine no es una línea reta, ya que los puntos estan en una esfera. En la practica no es muy utilizada ya que es necesario que los puntos estén situados en una esfera y la tierra no es completamente redonda.

Casos de uso

Se utiliza a menudo en la navegación, para calcular la distancia entre dos países cuando se vuela entre ellos.

$$hav(\theta) = sin^2\left(\frac{\theta}{2}\right)$$

$$d = 2r\ arcsin\left(\sqrt{sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + cos(\varphi_1)cos(\varphi_2)sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

Donde: $\varphi \Rightarrow Latitud(rad).$
$\lambda \Rightarrow Longitud(rad).$
$r = radio\ de\ la\ esfera.$

3. DT

- **Impureza de Gini**

Es uno de los métodos utilizados para decidir la división óptima de un nodo raíz y las divisiones posteriores. Es la forma más popular y fácil de dividir un árbol de decisión y solo funciona con objetivos categóricos, ya que solo realiza divisiones binarias.

Cuanto menor sea, mayor es la homogeneidad del nodo. La impureza de Gini de un nodo puro (misma clase) es cero.

$$I_G = 1 - \sum_{i=1}^{C}(p_i)^2$$

- **Ganancia de información**

Es la información que puede aumentar el nivel de certeza después de la división. Podemos pensar en esto y la entropía como opuestos. Se basa en el concepto de entropía de la teoría de la información.

$$I_E(f) = -\sum_{i=1}^{m} f_i\ log_2 f_i$$

- **Reducción de la varianza**

Es un algoritmo usado para variables objetivo-continuas (problemas de regresión). Este algoritmo usa la fórmula estándar de la varianza para escoger el criterio de división. La división con la varianza más baja se escoge para dividir la población.

$$I_V = \frac{\sum(X - X)^2}{n}$$

Referencias

Alvear, J. O. (2018, Noviembre). *Arboles de Decisión - Parte I.* Retrieved from bookdown.org

Grootendorst, M. (n.d.). *9 medidas de distancia en ciencia de datos.* Retrieved from www.ichi.pro

Hojas, I. M. (n.d.). *Construyendo árboles de decisión.* Retrieved from www.statdeveloper.com

Interactive Chaos. (n.d.). *Árbol de decisión.* Retrieved from interactivechaos.com

Li, B. (2014, Agosto). *Machine Learning Algorithms: DT, FR & SVM.* Retrieved from amazonaws.com

Na8. (2018, Julio). *Clasificar con K-Nearest Neighbor ejemplo en Python.* Retrieved from www.aprendemachinlearning.com

Sarang Anil Gokte, P. B. (2020). Retrieved from Most Popular Distance Metrics Used in KNN and When to Use Them: www.kdnuggets.com

Sitio Big Data. (2019, Diciembre). *Árbol de decisión en Machine Learning (Parte 1).* Retrieved from sitiobigdata.com

Tipos de métricas de distancia y uso de métricas de distancia definidas por el usuario en el algortimo KNN de Scikit. (n.d.). Retrieved from www.ichi.pro

Wikipedia. (2020, Diciembre). *Aprendizaje basado en árboles de decisión.* Retrieved from wikipedia.org