Modelo para procesamiento de lenguage natural?

Victor Hugo Matus Maldonado

23 de marzo de 2020

Resumen ejecutivo

Resumen ejecutivo.

Índice

1. Marco teórico y estado del arte

1.1. Bases de datos y álgebra relacional

El *modelo relacional de base de datos*, consiste en cinco componentes:

1. Una colección de tipos escalares, pueden ser definidos por el sistema o por el usuario.

2. Un generador de tipos de relaciones y un intérprete para las relaciones mismas.

3. Estructuras para definir variables relacionales de los tipos generados.

4. Un operador para asignar valores de relación a dichas variables.

5. Una colección relacionalmente completa para obtener valores relacionales de otros valores relacionales mediante operadores.

Las operaciones del modelo relacional están cimientadas en el *álgebra relacional*. Utilizando operaciones primitivas del álgebra se producen nuevas relaciones que pueden manipularse también por medio de operaciones del álgebra mismo. Una secuencia de operaciones de álgebra relacional forma una expresión cuyo resultado es una relación que representa el resultado de una consulta de base de datos. Estas operaciones se pueden clasificar en dos grupos, operaciones de la teoría de conjuntos: *UNIÓN*, *INTERSECCIÓN*, *DIFERENCIA* y *PRODUCTO CARTESIANO* (*PRODUCTO CRUZADO*), y el otro grupo consiste en operaciones específicas para bases de datos relacionales: *JUNTAR*, *SELECCIONAR* y *PROYECTAR*.

1.2. Inteligencia artificial

> "La *inteligencia artificial* es un campo antiguo y amplio que generalmente se puede definir como todos los intentos de automatizar el proceso cognitivo (...) la automatización del pensamiento. Esto puede ir desde lo más básico, como una hoja de cálculo de Excel, hasta lo más avanzado, como un androide que puede hablar y caminar." (Chollet, 2018)

Dentro de los múltiples tipos de inteligencia artificial, los que son de nuestro interés para este proyecto se describen a continuación.

Burkov (2019) define el *aprendizaje maquinal* como

> "preocupado con construir algoritmos que, para ser útiles dependen de una colección de ejemplos de algún fenómeno (...) el proceso de resolver problemas prácticos por 1) reunir un conjunto de datos y, 2) construir algorítmicamente un modelo estadístico basado en ese conjunto de datos"

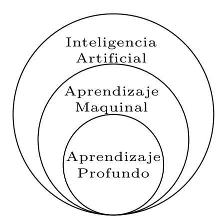

Figura 1: AP es un subcampo de AM, que es un subcampo de IA.

A diferencia del paradigma clásico de programación, donde los humanos introducen órdenes y datos para ser procesados de acuerdo con dichas reglas, en el aprendizaje maquinal el humano introduce datos y respuestas esperadas de estos datos como ejemplos, el resultado es la generalización de ciertas respuestas a partir de dichos datos sin estructurar. Con ello, se induce al conocimiento por parte de la computadora.

La *lingüística computacional* es un campo multidisciplinario de la lingüística aplicada en la informática. Se sirve de los sistemas informáticos para el estudio y el tratamiento del lenguaje. Para ello, se intenta modelar de manera lógica el lenguaje natural desde un punto de vista programable.

El *procesamiento del lenguaje natural* una disciplina de la rama de la ingeniería para la lingüística computacional. Se utiliza para la formulación e investigación de mecanismos de eficacia informática para servicios de comunicación entre las personas o entre ellas y las máquinas usando lenguajes naturales. Dos de los módulos básicos de procesamiento natural del lenguaje son búsqueda y aprendizaje con los que se pueden resolver muchos problemas con técnicas de optimización enfocadas en los diferentes parámetros involucrados.

1.3. Mate

Representamos la dependencia entre dos variables, en el que una aumenta o disminuye cuando la otra cambia con la *covarianza*

$$cov(X,Y) = E[(X - EX)(Y - EY)], \tag{1}$$

al referencia para la covarianza es conveniente escalarla de acuerdo a su desviación estándar, esto recibe el nombre de *coeficiente de correlación*

$$p = \frac{Cov(X,Y)}{\sigma_1 \sigma_2}. \tag{2}$$

Un modelo que relaciona $E(Y)$ como una función lineal únicamente de β_0 y β_1, es llamado modelo de regresión lineal *simple*

$$E(Y) = \beta_0 + \beta_1 x, \tag{3}$$

cuando más de una variable independiente es de interés, por ejemplo $x_1, x_2, \ldots, x_n$, se utiliza una generalización de (3), denominada modelo de regresión lineal *múltiple*

$$E(Y) = \beta_0 + \beta_1 x + \ldots + \beta_n x_n. \tag{4}$$

Demostrar que los conjuntos de hipótesis que contienen β_1, por ejemplo, $H_a: \beta_1 = 0$ contra $H_a: \beta_1 > 0$ $H_a: \beta_1 < 0$, así como $H_a: \beta_1 > 0$ contra $H_a: \beta_1 \neq 0$ pueden estar basadas en el estadístico

$$t = \frac{\hat{\beta}_1 - 0}{S/\sqrt{Sxx}}, \tag{5}$$

cuando se tienen muestras moderadamente grandes puede probarse la hipótesis $H_0: \rho_1 = \rho_0$ con una prueba Z

$$Z = \frac{(\frac{1}{2})\ln(\frac{1+r}{1-r}) - (\frac{1}{2})\ln(\frac{1+\rho}{1-\rho})}{\frac{1}{\sqrt{n-3}}}. \tag{6}$$

Variable aleatoria (v.a.) es la función real $X : \Omega \mapsto \mathbb{R}$ tal que el conjunto $\{\omega \in \Omega : X(\omega) \in I\}$ es un evento de Ω para cada $I \subset \mathbb{R}$, en un espacio Ω hipotético. Se le considera *variable aleatoria discreta* (v.a.d.) cuando su rango de valores R_x es finito o contablemente infinito, mientras que una *variable aleatoria continua* (v.a.c.) puede tomar cualquier valor real en un intervalo.

> "La forma más natural de expresar la distribución de v.a.d.s es la *función de probabilidad*" (Blitzstein y Hwang, 2019)

Una v.a.d. X con $R_x = \{x_1, x_2 x_3, \dots, x_n, \dots\}$ tiene una función de distribución

$$\begin{array}{c} f(x) = 0 \text{ para cada } x \notin R_x; \\ f(x) = P(X = x) \text{ para } x \in R_x \end{array} \tag{7}$$

para una v.a.c. X será una función no negativa real $f : \mathbb{R} \mapsto [0, \infty)$, es decir

$$P(X \in A) = \int_A f(x)dx \tag{8}$$

El *valor esperado* de una v.a.d. X con una función de probabilidad (7) es definida como

$$\mu = E(X) = \sum_{x \in R_x}^{\infty} x f(x), \tag{9}$$

siempre y cuando la serie converja absolutamente y es también llamado *media* de X, utilizada, similar a la media aritmética en estadísticas, para obtener el valor promedio entre observaciones.

Para una v.a.c. X se define como

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)dx \tag{10}$$

Para conocer la variabilidad de la distribución de cualquier v.a se utiliza la *varianza*, para X se define

$$\sigma^2 = Var(X) = [(X - \mu)^2] \tag{11}$$

Figura 2: Diversas de distribuciones pueden para modelar v.a.s, a continuación se muestran las mas importantes de acuerdo a Balakrishnan, Koutras y Politis (2020).

binominal.png

Binominal $b(n, p)$

geom.png

Geométrica $G(p)$.

negativabinominal.png

Negativa binominal $Nb(rp)$

hyperg.png

Hipergeométrica $h(n; a, b)$

poisson.png

Poisson $\mathcal{P}(\lambda)$

uniforme.png

Uniforme $U[a, b]$

normal.png

Normal $N(\mu, \sigma^2)$

epsilon_lambda.png

Exponencial $Expo(\lambda)$

gama.png

Gama $Ga(\alpha, \beta)$

beta.png

Beta $Be(\alpha, \beta)$

2. Objetivos

2.1. General

Haciendo uso de las ciencias de la computación, las herramientas matemáticas de estadística y métodos de aprendizaje autónomo, se busca obtener información cuantitativa de textos provenientes de redes sociales, cadenas noticiosas y audio de programas de capacitación, para inferir posiciones, tendencias, comportamientos o razones de grupos sociales, considerando un ciclo de clasificación, estimación, detección y comprobación.

2.2. Particulares

1. Desarrollar un modelo de base de datos que permita la captura de categorías para un determinado problema, los elementos de identificación de cada categoría, el origen de la información y su correlación.

2. Construir una estructura de datos que capte la estimación o valores esperados para el procesamiento de textos.

3. Elaborar un sistema de objetos para el soporte de los elementos de aprendizaje autónomo.

4. Generar los elementos de captura de textos para su almacenamiento y procesamiento.

5. Elaborar un modelo estadístico que permita comprobar las estimaciones a partir de los datos y en consecuencia realizar un ajuste en los parámetros usados para el aprendizaje autónomo.

6. Producir los reportes con un análisis estadístico que faciliten la interpretación de resultados y den pauta para la obtención del conocimiento de interés.

2.3. Metas científicas

Esperando respuesta mail.

3. Metodología científica

# #

En gdrive, esperando retroalimentación.

4. Grupo de trabajo

- Dr. José Emilio Quiroz Ibarra
  Universidad Iberoamericana, Dirección.
- Dra. Alma Rocío Sagaceta Mejía
  Universidad Autónoma Metropolitana, Codirección.
- Mtra. Paloma Alejandra Vilchis León
  Universidad Tecnológica de México*, Tutoría. ↑ preguntar

5. Infraestructura disponible para el proyecto

Laboratorios y equipos disponibles en la Universidad Iberoamericana Ciudad de México.

Desktop CPU Intel i7 (ó AMD Ryzen 7) RAM 32 GB Almacenamiento 2TB SSD (ó 256 SSD y 2TB HDD)

Servidor Renta de servicio virtual en Google Cloud, Amazon AWS o Microsoft Azure, entre otros. (adquirir uno, al menos $50k$).

6. Cronograma de actividades

cronograma

7. Resultados comprometidos

Publicación

8. Visto bueno

Vo.Bo.

9. Referencias

Balakrishnan, N., Markos V. Koutras y Konstantinos G. Politis (2020). *Introduction to probability: models and applications [Introducción a la probabilidad: modelos y aplicaciones]*. Hoboken: John Wiley & Sons, Inc. ISBN: 9781118123348.

Blitzstein, Joseph K. y Jessica Hwang (2019). *Introduction to Probability [Introducción a la probabilidad]*. 2.ª ed. Texts in Statistical Science [Textos en ciencia estadistica]. Florida: CRC Press. ISBN: 9781138369917.

Burkov, Andriy (2019). *The hundred-page machine learning book*. Quebec: Andriy Burkov. ISBN: 9781999579500.

Chollet, François (2018). *Machine Learning With Python [Machine Learning]*. New York: Manning Publications Co. ISBN: 9781617294433.

Wackerly, Dennis D, william Mendenhall III y Richard L Scheaffer (2009). *Estadística matemática con aplicaciones*. Trad. por Jorge Humberto Romo Mufioz. 7.ª ed. Ciudad de México: Cenage Learning. ISBN: 9780495110811.