Modelo de procesamiento de lenguaje natural a partir de una estimación estadística

Victor Hugo Matus Maldonado

25 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"

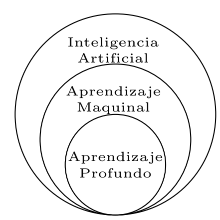


Figura 1: AP es un subcampo de AM, que es un subcampo de IA.

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. Fundamentos matemáticos✓

Para el desarrollo del presente proyecto se usará fuertemente la estadística, se observará la cantidad de veces que aparece la palabra clave en la base de datos recolectada para comprender el comportamiento o distribución que tiene. Además se buscará la relación que existe de la palabra clave con algún otro conjunto de palabras.

Para fundamentar todo el trabajo es importante definir algunos conceptos primordiales de la estadística. Una *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 del espacio muestral Ω para cada intervalo $I \subset \mathbb{R}$. Existen dos tipos de variable, la primera de ellas es la *variable aleatoria discreta* (v.a.d.), la cual está definida cuando su rango de valores R_x es finito o contablemente infinito y la segunda es la *variable aleatoria continua* (v.a.c.), es aquella que puede tomar cualquier valor real en un intervalo.

La *distribución de probabilidad* es un modelo teórica que describe la forma en que varían o cambian los resultados de un experimento aleatorio, en otras palabras, a través de una función obtenemos las probabilidad de todos los posibles resultados que podrían obtenerse cuando se realiza un experimento aleatorio. Se hace una diferencia en las funciones de distribución para las variables aleatorias discretas y las variables aleatorias continuas.

Para una v.a.d. X con $\Omega = \{x_1, x_2, x_3, \ldots, x_n, \ldots\}$ donde $f(x)$ es la función de distribución, la probabilidad se encuentra dada por:

$$P(X = x) = \begin{cases} 0, & x \notin \Omega \\ f(x), & x \in \Omega. \end{cases} \tag{1}$$

Para una v.a.c. X, la función de densidad probabilística es una función no negativa real $f : \mathbb{R} \mapsto [0, \infty)$, la probabilidad se encuentra dada por:

$$P(X \in A) = \int_A f(x)dx. \tag{2}$$

El *valor esperado, la esperanza* de una v.a. X discreta o continua, con una función de probabilidad $f(x)$ discreta o continua se define como

$$\mu = E(X) = \sum_{x \in \Omega} xf(x), \quad \mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx, \tag{3}$$

respectivamente, la definición anterior es también llamada la *media* de X, cuyo uso es similar a la media aritmética en estadísticas.

Otro valor numérico importante que se usa para conocer la variabilidad de la distribución de cualquier v.a se es la *varianza*, se define como

$$\sigma^2 = Var(X) = E(X^2) - E(X)^2. \tag{4}$$

En la figura 4 se muestran las distribuciones más comunes para variables aleatorias discretas y continuas [Balakrishnan, Koutras y Politis 2020].

Una vez identificado el comportamiento de la variable aleatoria es útil pensar en la relación que tiene con otro tipo de variables. Se hace uso de un modelo que relaciona $E(X)$ como una función lineal únicamente dependiente de los parámetros β_0 y β_1, este modelo es llamado modelo de regresión lineal *simple*

Hasta aqui Alma

$$E(X) = \beta_0 + \beta_1 x, \tag{5}$$

cuando más de una variable independiente es de interés, por ejemplo $x_1, x_2, \ldots, x_n$, se utiliza una generalización de (18), denominada modelo de regresión lineal *múltiple*

$$E(Y) = \beta_0 + \beta_1 x + \ldots + \beta_n x_n. \tag{6}$$

La *dependencia lineal* entre dos variables se puede medir usando la *covarianza* definido como:

$$cov(X, Y) = E[(X - EX)(Y - EY)], \tag{7}$$

dicho valor es conveniente escalarlo de acuerdo a su desviación estándar, esto recibe el nombre de *coeficiente de correlación*

$$p = \frac{cov(X,Y)}{\sigma_1 \sigma_2}, \qquad -1 \leq p \leq 1,$$

$$\text{correlación} \begin{cases} p > 0, \text{ positiva}; & p < 0, \text{ negativa}; \\ p \approx 0, \text{ débil}; & p = 0, \text{ no hay } \therefore \text{ variables independientes}; \\ p = 1, \text{ perfecta}; & p = -1, \text{ perfecta inversa } (\textit{anticorrelación}). \end{cases} \tag{8}$$

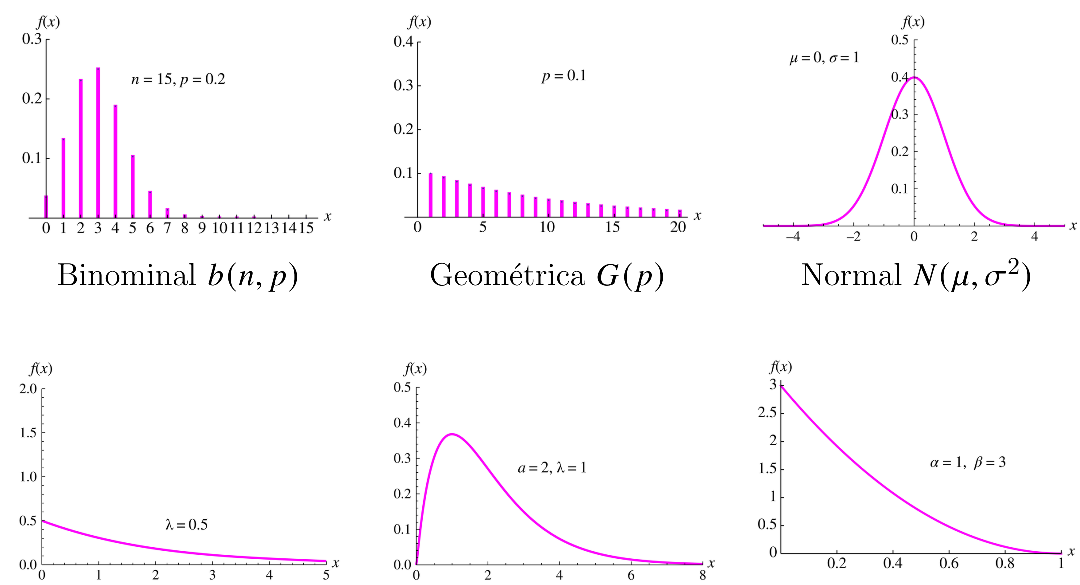
Demostrar que los conjuntos de hipótesis que contienen β_1, por ejemplo, $H_a\colon \beta_1 = 0$ contra $H_a\colon \beta_1 > 0$ $H_a\colon \beta_1 < 0$, así como $H_a\colon \beta_1 > 0$ contra $H_a\colon \beta_1 \neq 0$ pueden estar basadas en el estadístico

$$t = \frac{\hat{\beta_1} - 0}{S/\sqrt{Sxx}}, \tag{9}$$

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

$$Z = \frac{(\frac{1}{2})\ln(\frac{1+r}{1-r}) - (\frac{1}{2})\ln(\frac{1+\rho}{1-\rho})}{\frac{1}{\sqrt{n-3}}}. \tag{10}$$

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



# #

1.4. Fundamentos matemáticos×

Variable aleatoria (v.a.) es la función real $X : \mathbf{\Omega} \mapsto \mathbb{R}$ tal que el conjunto $\{\omega \in \mathbf{\Omega} : X(\omega) \in I\}$ es un evento de $\mathbf{\Omega}$ para cada $I \subset \mathbb{R}$, en un espacio $\mathbf{\Omega}$ 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 $R_x = \{x_1, x_2 x_3, \ldots, x_n, \ldots\}$ tiene una función de distribución

$$\begin{aligned} f(x) = 0 \text{ para cada } x \notin R_x; \\ f(x) = P(X = x) \text{ para } x \in R_x \end{aligned} \tag{11}$$

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

$$P(X \in A) = \int_A f(x)dx \tag{12}$$

El *valor esperado* de una v.a.d. X con una función de probabilidad (11) es definida como

$$\mu = E(X) = \sum_{x \in R_x}^{\infty} x f(x), \tag{13}$$

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.

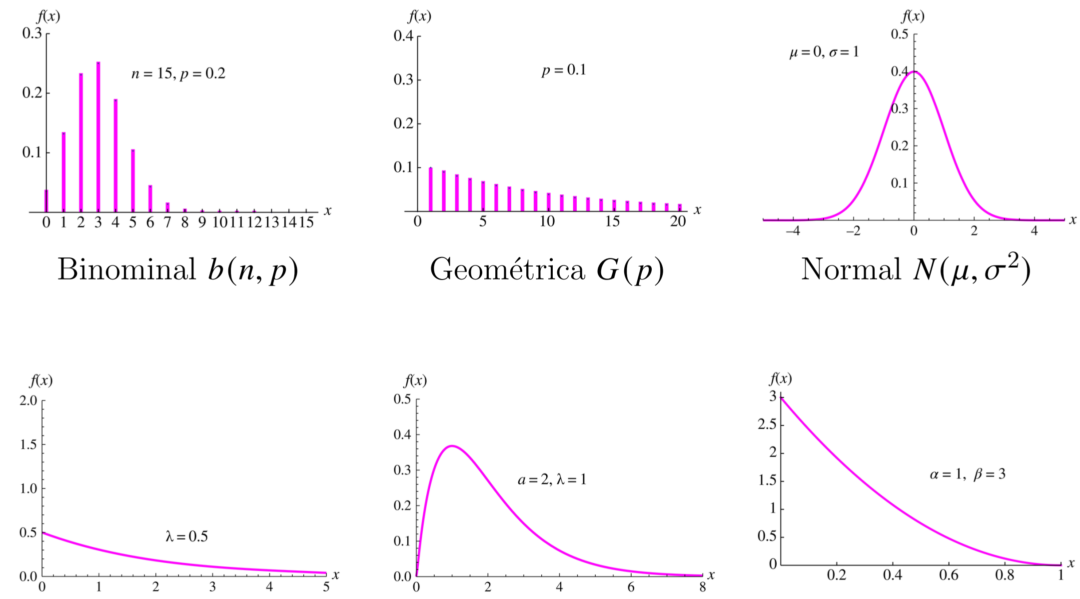
Para una v.a.c. X se define como

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)dx \tag{14}$$

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

$$\sigma^2 = Var(X) = [(X - \mu)^2] \tag{15}$$

Figura 4: Diversas de distribuciones pueden modelar v.a.s, a continuación se muestran las mas importantes de acuerdo a Balakrishnan, Koutras y Politis (2020).



Binominal $b(n, p)$ Geométrica $G(p)$ Normal $N(\mu, \sigma^2)$

Exponencial $Expo(\lambda)$ Gama $Ga(\alpha, \beta)$ Beta $Be(\alpha, \beta)$

Se representa la dependencia entre dos variables, cuando una aumenta o disminuye al presentarse cambios en la otra con la *covarianza*

$$cov(X, Y) = E[(X - EX)(Y - EY)], \tag{16}$$

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

$$p = \frac{cov(X,Y)}{\sigma_1 \sigma_2}, \qquad -1 \leq p \geq 1,$$

$$\text{correlación} \begin{cases} p > 0, \text{ positiva;} & p < 0, \text{ negativa;} \\ p \approx 0, \text{ débil;} & p = 0, \text{ no hay} \therefore \text{variables independientes;} \\ p = 1, \text{ perfecta;} & p = -1, \text{ perfecta inversa } (\textit{anticorrelación}). \end{cases} \tag{17}$$

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

$$E(Y) = \beta_0 + \beta_1 x, \tag{18}$$

cuando más de una variable independiente es de interés, por ejemplo $x_1, x_2, \ldots, x_n$, se utiliza una generalización de (18), denominada modelo de regresión lineal *múltiple*

$$E(Y) = \beta_0 + \beta_1 x + \ldots + \beta_n x_n. \tag{19}$$

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

$$t = \frac{\hat{\beta_1} - 0}{S/\sqrt{Sxx}}, \tag{20}$$

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

$$Z = \frac{(\frac{1}{2})\ln(\frac{1+r}{1-r}) - (\frac{1}{2})\ln(\frac{1+\rho}{1-\rho})}{\frac{1}{\sqrt{n-3}}}. \tag{21}$$

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. La fuente de información será redes sociales, cadenas noticiosas y audio de programas de capacitación, para inferir posiciones, tendencias, comportamientos o razones de grupos sociales. El de trabajo es 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.

3. Metas científicas

La meta de este proyecto es -la integración -de elementos de estad, ciencia comp, apr autónomo(ML) para procesamiento de lenguage natural (PLN).

4. Metodología científica

1 Recopilación de información en forma aislada e integrada con aplicaciones relacionadas al proyecto.
2 Construir los modelos para la obtención de la base de datos.
3 Obtener el conjunto de palabras correlacionadas con la palabra clave para realizar una búsqueda más extensa en los textos y estimar los modelos de regresión lineal y/o regresión múltiple con las palabras seleccionadas.
4 Realizar una búsqueda correlacionada complementaria.
5 A partir de los nuevos datos se obtienen otros modelos de regresión lineal y se realizan pruebas de hipótesis para comparar los parámetros β_0 y β_1 del modelo anterior. El objetivo de este proceso es obtener los mejores estimadores (con medidas de dispersión optimizadas).
6 Construcción de elementos gráficos con elementos de regresión
7 Construcción de elementos gráficos (distribuciones y rectas de regresión lineal).
8 Comunicación de resultados vía publicación y/o congreso en eventos especializados.

5. Grupo de trabajo

- Dr. José Emilio Quiroz Ibarra, Director.
  Universidad Iberoamericana.

- Dra. Alma Rocío Sagaceta Mejía, Codirectora
  Universidad Autónoma Metropolitana.

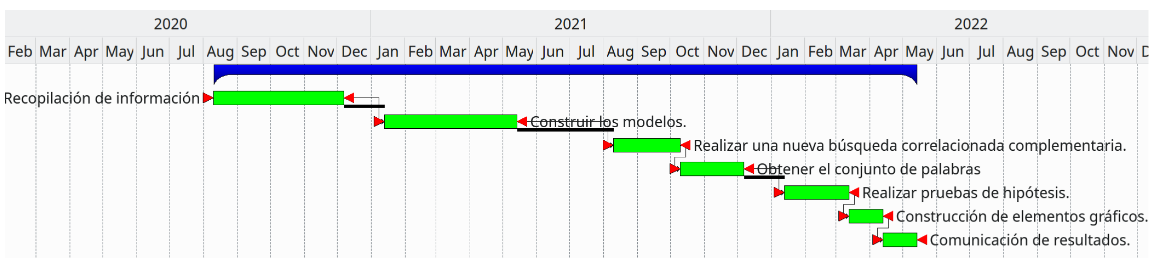
6. Infraestructura disponible para el proyecto

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

(preferentemente
estación de trabajo
CPU 8 núcleos físicos a 4MHz
RAM 16 GB
Almacenamiento 2TB SSD.

Servidor
Renta de servicio virtual en Google Cloud, Amazo, Microsoft Azure.)

7. Cronograma de actividades



8. Resultados comprometidos

Generar guía de implementación de modelos PLN Generar una interfaz de fácil acceso para aplicaciones de PLN

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 [Introduccién 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.