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

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"

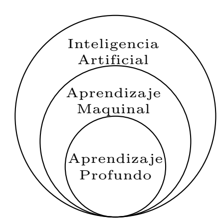


Figura 1: AP es un subcampo de AM, que es un subcampo de IA.

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

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

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 3 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 otros tipo de variables. Un modelo de regresión lineal simple consiste en generar una recta que permita explicar la relación *lineal* que existe entre dos variables. La variable dependiente Y y la variable predictora o independiente X se relacionan como:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{5}$$

donde β_0, β_1 son valores estimados y ε el error aleatorio con distribución normal. La ecuación (5) es una aproximación de la verdadera relación entre X e Y, en el cual para un valor dado de X el modelo es capaz de predecir un cierto valor para Y. Un ejemplo del modelo se muestra en la figura 2.

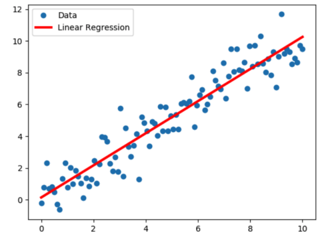


Figura 2: Regresión lineal

La eficacia de la predicción del modelo lineal depende directamente de la estimación de los parámetros β_0 y β_1 para calcular la recta

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{6}$$

que se ajusta lo mejor posible a los datos con $E(\varepsilon) = 0$. Una vez determinado las estimaciones puntuales de esos parámetros es conveniente calcular los intervalos de confianza para obtener una medida de precisión y más aún realizar contrastes de hipótesis para comprobar si un valor determinado puede ser el auténtico valor del parámetro.

Existe también otro modelo lineal donde relaciona la variable dependiente Y con K variables explícitas $X_1, X_2, \ldots, X_k$ de la forma:

$$Y = \beta_0 + \sum \beta_k X_k + \varepsilon \tag{7}$$

donde ε corresponde a los errrores de los estimadores β_k. Este modelo es llamado *modelo lineal multivariable* y el procedimiento para los estimadores $\hat{\beta}_k$ sigue un procedimiento similiar al caso simple.

Una medida que será útil al relacionar dos variables será la *dependencia lineal* que exista entre ellas, se usará la *covarianza*, la cual está definida como:

$$cov(X, Y) = E[(X - EX)(Y - EY)]. \tag{8}$$

Se sabe que si hay una relación lineal positiva, el valor de la covarianza será positiva y grande. En caso de que la relación lineal sea negativa, la covarianza será negativa y grande. Finalmente, cuando la covarianza es cercana a cero se sabe que no hay relación entre ambas variables.

La covarianza depende de las unidades de medida de las variables por lo que es conviente usar una medida que no dependa de las unidades, esta medida recibe el nombre de *coeficiente de correlación* definida como:

$$r_{(X,Y)} = \frac{cov(X,Y)}{s_1 s_2} \tag{9}$$

donde s_1^2 y s_2^2 son las varianzas muestrales de las variables X y Y respectivamente.

En el caso del modelo lineal simple se busca que el valor de la pendiente de la recta (β_1) sea distinto de cero, por lo que se pondrá un especial interés al contraste:

$$H_0: \quad \beta_1 = 0 \tag{10}$$
$$H_1: \quad \beta_1 \neq 0 \tag{11}$$

donde la región de rechazo de la hipótesis nula es:

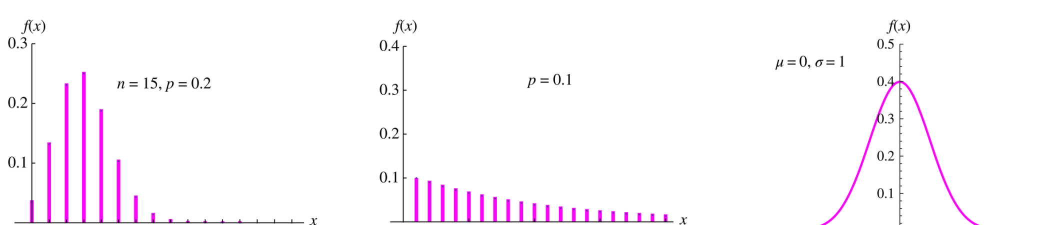
$$\left| \frac{\hat{\beta}_1}{S/\sqrt{Sxx}} \right| > t_{n-2,\alpha/2}, \tag{12}$$

donde $t_{n-2,\alpha/2}$ es la distribución de t-student con $n-2$ grados de libertad y un nivel de significancia de α.

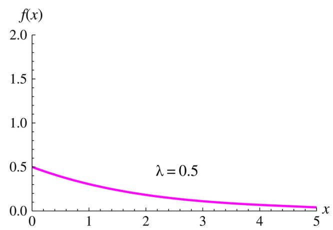
Utilizaremos el contraste de hipótesis para encontrar una mejor estimación de los parámetros β_0 y en particular β_1. Dicho procedimiento se puede consultar en

Ponerlo como referencia en la bibliografía **Devore, Jay L. Probabilidad y estadística para ingeniería y ciencias. México: Thomson, 2002. Hollander, Myles y Douglas Wolfe. Nonparametric statistical Methods. New York: Wiley, 1999.**

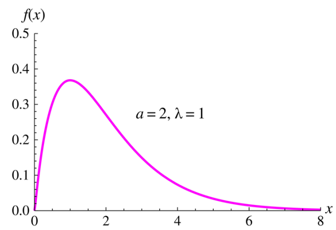
Figura 3: Diversas de distribuciones pueden modelar v.a.s, a continuación se muestran las mas importantes de acuerdo a Balakrishnan, Koutras y Politis (2020).



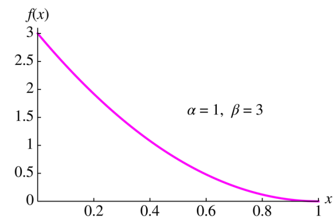
Binominal $b(n,p)$ Geométrica $G(p)$ Normal $N(\mu,\sigma^2)$



Exponencial $Expo(\lambda)$



Gama $Ga(\alpha, \beta)$



Beta $Be(\alpha, \beta)$

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 estadística, ciencias computacionales, aprendizaje autónomo (machine learning) para el 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 (distribuciones y rectas de regresión lineal).
7. 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 UIA, Ciudad de México.
Preferentemente

- Estación de trabajo con CPU de 8 núcleos físicos a 4MHz, RAM 16 GB, almacenamiento 2TB SSD.
- Servidor virtual en Google Cloud, Amazon, Azure o IBM.

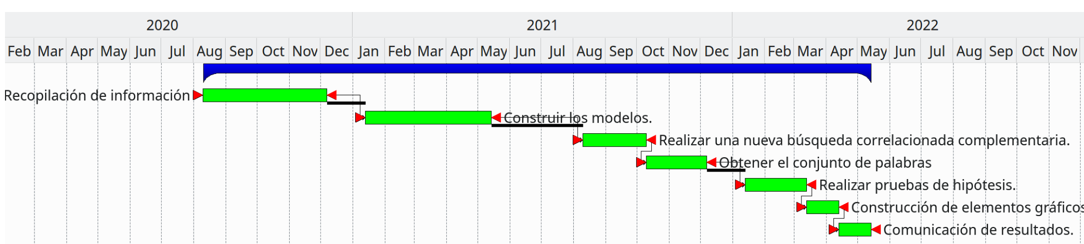
7. Cronograma de actividades

En primer semestre se recolectará información de forma aislada así como integrada con aplicaciones afines al proyecto.

Durante segundo semestre se construirán los modelos para la obtención de la base de datos.

En tercer semestre se obtendrá el conjunto de palabras para extender la búsqueda y estimar modelos de regresión, y se realizará una búsqueda correlacionada complementaria.

En cuarto semestre se obtendrán nuevos modelos de regresión lineal utilizando los datos de la búsqueda y se realizarán pruebas de hipótesis del modelo anterior para obtener mejores estimadores. Se construirán elementos gráficos usando distribuciones y rectas de regresión lineal y se comunicarán los resultados.



8. Resultados comprometidos

- Generar guía de implementación de modelos PLN.
- Crear 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.