Modelo para procesamiento de lenguage natural?

Victor Hugo Matus Maldonado

21 de marzo de 2020

Resumen ejecutivo

Resumen ejecutivo.

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 *ingeligencia artificial es*

> el esfuerzo por automatizar tareas intelectuales normalmente realizadas por humanos(Chollet, 2018)
>
> "es una de las disciplinas más sofisticadas creadas por el ser humano (...) está permitiendo obtener resultados similares a los que observamos en las capacidades de la inteligencia humana: reconocimiento del entorno y percepción espacial, predicción y anticipación, entendimiento del lenguaje y capacidades de comunicación..." recordar cómo citar entrevistas en páginas

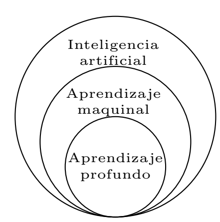

Figura 1: Inteligencia artificial

> web www.apd.es/ entrevista-gfi-estamos-asistiendo -al-auge-de -agentes-inteligentes-capaces-de-comunicarse -como-una-persona

de este campo general se desprenden el *aprendizaje maquinal* y el *aprendizaje profundo*, figura (1).

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

Aprendizaje maquinal

Burkov (2019) lo define 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

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.

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.

Procesamiento del lenguaje natural

Es 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 dos

Representamos la dependencia entre dos variables, en el que una aumenta o disminuye cuando la otra cambia con el *coeficiente de correlación* p

$$p = \frac{Cov(Y_1, Y_2)}{\sigma_1 \sigma_2} \tag{1}$$

La covarianza mide qué tanto o tan poco las dos variables aleatorias cuyos valores esperados existen y son positivos tienen dependencia lineal, denotada $cov(X, Y)$ la covarianza de X y Y es definida como

$$cov(X, Y) = E[(X - EX)(Y - EY)] \tag{2}$$

Cuando no se tiene una referencia para usar la covarianza, tiene sentido escalarla de acuerdo a la desviación estándar de las variablesMatloff 2017

$$p(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)}\sqrt{var(Y)}} \tag{3}$$

denotada $corr(X, Y)$ de esta es la *correlación* de X y Y. Un coeficiente de relación $p = 0$ indica que no hay relación.

1.4. Mate

Una *variable aleatoria* (v.a.) es una 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 $R_x = \{x_1, x_2 x_3, \ldots, x_n, \ldots\}$ tiene una función de distribución

$$\begin{aligned} f(x) &= 0 \text{ para cada } x \notin R_x; \\ f(x) &= P(X = x) \text{ para } x \in R_x \end{aligned} \tag{4}$$

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

$$P(X \in A) = \int_A f(x) dx \tag{5}$$

El *valor esperado* de una v.a.d. X con una función de probabilidad (4) es definida como

$$\mu = E(X) = \sum_{x \in R_x}^{\infty} x f(x), \tag{6}$$

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

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx \tag{7}$$

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

$$\sigma^2 = Var(X) = [(X - \mu)^2] \tag{8}$$

Figura 2: Existe diversidad de distribuciones para modelar v.a.s, a continuación se muestran las mas importantes de acuerdo a Balakrishnan, Koutras y Politis (2020). De las siguientes son aptas para v.a.d.s hasta la distribución Normal, mientras que las v.a.c.s pueden ser modeladas desde la distribución Uniforme.

`binominal.png`

Binominal $b(n, p)$. Número de éxitos en n ensayos de Bernoulli independientes con la misma probabilidad de éxito p.

$$\begin{gathered} f(x = k) = \tbinom{n}{x} p^x p^{n-x}, \\ x = 0, 1, \ldots, n; \\ E(X) = np,\ Var(X) = npq \end{gathered} \tag{9}$$

`geom.png`

Geométrica $G(p)$. Número n de ensayos de Bernoulli independientes con la misma probabilidad de éxito p, hasta obtener el primer éxito.

$$\begin{gathered} f(x) = q^{x-1}, \\ x = 1, 2, \ldots, n; \\ E(X) = \tfrac{1}{p},\ Var(X) = \tfrac{q}{p^2} \end{gathered} \tag{10}$$

negativabinominal.png

Negativa binominal $Nb(rp)$. Número n de ensayos de Bernoulli independientes con la misma probabilidad de éxito p, hasta obtener resultado número r.

$$\begin{array}{c} f(x) = \binom{x-1}{r-1} p^r q^{x-r}, \\ r = r, r+1, r+2, \ldots, n; \\ E(X) = \frac{r}{p}, \ Var(X) = \frac{rq}{p^2} \end{array} \tag{11}$$

hyperg.png

Hipergeométrica $h(n; a, b)$. Muestra aleatoria tamaño n de elementos tipo a no reemplazada de un universo con elementos tipo a y b.

$$\begin{array}{c} f(x) = P(X = x) = \frac{\binom{a}{x}\binom{b}{n-x}}{\binom{a+b}{n}}; \\ E(X) = n \cdot \frac{a}{a+b}, \\ Var(X) = n \cdot \frac{a}{a+b} \cdot \frac{b}{a+b} \cdot (1 - \frac{n-1}{a+b-1}) \end{array} \tag{12}$$

poisson.png

Poisson $\mathcal{P}(\lambda)$. Puede utilizarse con algún parámetro λ cuando la probabilidad de éxito tiende a cero ($p \to 0$) de forma que la media $E(X) = np$ converge en algún $\lambda > 0$.

$$\begin{array}{c} f(x) = e^{-\lambda} \frac{\lambda^x}{x!}, \\ x = 0, 1, 2, \ldots; \\ E(X) = \lambda, Var(X) = \lambda \end{array} \tag{13}$$

uniforme.png

Uniforme $U[a, b]$. Útil para modelar situaciones en que todos los intervalos tengan la misma amplitud y probabilidad.

$$\begin{array}{c} f(x) \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ \text{de otro modo}, & 0; \end{cases} \\ F(t) \begin{cases} 0, & t < 0, \\ \frac{t-a}{b-a}, & a \leq t \leq b, \\ 1, & t > b; \end{cases} \\ E(X) = \frac{a+b}{2}, \quad Var(X) = \frac{(b-a)^2}{12} \end{array} \tag{14}$$

normal.png

Normal $N(\mu, \sigma^2)$. La suma y el promedio de un gran número de observaciones para una variable X puede ser aproximada por la distribución normal, independientemente de su distribución original.

$$\begin{gathered} f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}, \\ -\infty < z < \infty; \\ E(X) = \mu, \ Var(X) = \sigma^2 \end{gathered} \tag{15}$$

epsilon_lambda.png

Exponencial $\epsilon(\lambda)$. Es considerada la análoga continua de la distribución geométrica y puede ser utilizada para modelar parte de la vida de un sujeto X.

$$\begin{gathered} f(x) = \begin{cases} \lambda^{-\lambda x}, \ x \leq 0, \\ 0, \ x < 0; \end{cases} \\ F(t) = \begin{cases} 0, \ t < 0, \\ 1 - e^{-\lambda t}, \ x \geq 0; \end{cases} \\ E(X) = \frac{1}{\lambda}, \ Var(X) = \frac{1}{\lambda^2} \end{gathered} \tag{16}$$

gama.png

Gama. Generalización de la distribución *Erlang* cuando el parámetro n no es necesariamente un entero.

$$\begin{gathered} f(x) = \begin{cases} \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}, \ x \geq 0, \\ 0, \ x < 0, \end{cases} \\ \text{donde } \Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt \\ \text{es la función gama.} \\ E(X) = \frac{a}{\lambda}, \ Var(X) = \frac{a}{\lambda^2} \end{gathered} \tag{17}$$

beta.png

Beta. Ofrece modelos satisfactorios para v.a.c.s que toman valores entre dos puntos conocidos.

$$\begin{gathered} f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \ 0 < x < 1, \\ \text{de otro modo, } 0, \end{cases} \\ \text{donde } B(\alpha, \beta) = \\ \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = \\ \big(\Gamma(\alpha)\Gamma(\beta)\big) / \big(\Gamma(\alpha+\beta)\big) \\ \text{es la función beta.} \\ E(X) = \frac{a}{\lambda}, \ Var(X) = \frac{a}{\lambda^2} \end{gathered} \tag{18}$$

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

Metas científicas

3. Metodología científica

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. servidor desktop

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.

Matloff, Norman S (2017). *Statistical regression and classification from linear models to machine learning [Regresión estadística y clasificación de modelos lineales a aprendizaje maquinal]*. FLorida: CRC Press. ISBN: 9781138066565.