Las mil secciones son para no perderme durante la escritura

Índice

Las secciones son para no perderme durante la escritura

0.0.1. Distribución de Bernoulli y binominal

Una variable aleatoria tiene la *distribución de Bernoulli* con un parámetro p si $P(X = 1) = p$ y $P(X = 0 = 1 - p)$, cuando $0 < p < 1$. Se escribe como $X \sim Bern(p)$, el símbolo $\sim$ significa "distribuido como" y la probabilidad p es el *parámetro*, que determina qué distribución de Bernoulli específica tenemos.

Supóngase que se realizan n ensayos Bernoulli independientes, cada uno con probabilidad p de éxito. X sea el número de éxitos, la distribución X se llama *distribución binominal* con parámetros n y p; se escribe $X \sim Bin(p, n)$. $Bern(p)$ es la misma distribución que $Bin(1, p)$. Bernoulli es un caso especial de binominal, si $x \sim Bin(1, p)$, entonces la función de probabilidad de X es

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \tag{1}$$

para $k = 0, 1, \ldots, n$ (y por otra parte $P(X = k) = 0$).

0.0.2. Distribución de hipergeométrica

Si $X \sim HGeom(w, b, n)$, entonces la función de probabilidad de X es

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}, \tag{2}$$

para enteros k satisfaciendo $0 \leq k \leq w$ y $0 \leq n-k \leq b$, y $P(X = k) = 0$. La estructura esencial de la distribución hipergeométrica se basa en que objetos en su población están clasificados usando dos tipos de etiquetas, al menos una de estas siendo asignada al azar. Las distribuciones $HGeom(w, b, n)$ y $HGeom(n, w+b-n, 1)$ son idénticas si X y Y tienen la misma distribución, podemos demostrarlo algebraicamente:

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}} = \frac{w!b!n!(w + b - n)!}{k!(w + b)!(w - k)!(n - k)!(b - n + k)!} \tag{3}$$

$$P(X = k) = \frac{\binom{n}{k}\binom{w+b-n}{w-k}}{\binom{w+b}{w}} = \frac{w!b!n!(w + b - n)!}{k!(w + b)!(w - k)!(n - k)!(b - n + k)!}. \tag{4}$$

0.0.3. Distribución uniforme discreta

Teniendo C, un conjunto finito no vacío de números, se elige un número uniformemente al azar (o sea que todos los números tienen la misma posibilidad de ser elegidos), llámese X. Entonces se dice que X una *distribución uniforme discreta* con el parámetro C. Se dice entonces que la función de probabilidad de $X \sim DUNif(C)$ (la distribución uniforme discreta de X) es

$$P(X = x) = \frac{1}{|C|} \tag{5}$$

para $x \in C$ (de lo contrario 0) ya que la función de probabilidad debe sumar 1.

0.1. distribuciones

0.1.1. Binominal geométrica y negativa

Distribución geométrica: Se tiene una secuencia de ensayos independientes Bernoulli, cada uno con la misma probabilidad de éxito $p \in (0, 1)$, con ensayos realizados hasta que se alcanza el éxito. X es el número de *fallas* antes de la primera prueba exitosa por lo que X tiene una *distribución geométrica* con un parámetro p; denotado $X \sim Geom(p)$. Con esto podemos llegar a los teoremas de *distribución geométrica de la función de probabilidad*, cuando $X \sim Geom(p)$, entonces la función de probabilidad de X será

$$P(X = k) = q^k p \tag{6}$$

para $k = 1, 2, \ldots$, cuando $q = 1 - p$; y el teoremas de *distribución geométrica de la función de distribución acumulativa*, cuando $X \sim Geom(p)$, entonces la función de distribución acumulativa de X será

$$F(x) = \begin{cases} 1 - q^{\lfloor x \rfloor + 1}, & \text{si } x \geq 0; \\ 0, & \text{si } x < 0, \end{cases} \tag{7}$$

cuando $q = 1 - q$ y $\lfloor x \rfloor$ es el mayor entero y menor o igual a x.

El valor esperado geométrico de $X \sim Geom(p)$ es

$$E(X) = \sum_{k=0}^{\infty} k q^k p, \tag{8}$$

cuando $q = 1 - p$. Aunque esta no es una serie geométrica, podemos llegar a ello

$$\begin{gathered} \textstyle\sum_{k=0}^{\infty} q^k = \frac{1}{1-q} \\ \textstyle\sum_{k=0}^{\infty} k q^{k-1} = \frac{1}{1-q^2}, \end{gathered} \tag{9}$$

finalmente multiplicamos ambos lados por pq, recuperando la suma original que queríamos encontrar

$$E(X) = \sum_{k=0}^{\infty} kq^k p = pq \sum_{k=0}^{\infty} kq^{k-1} = pq \frac{1}{(1-q)^2} = \frac{q}{p}. \tag{10}$$

Primer valor esperado de éxito *FS*, podemos definir a $Y \sim FS(p)$ como $Y = X + 1$ donde $X \sim Geom(p)$, por lo que tenemos

$$E(Y) = E(X+1) = \frac{q}{p} + 1 = \frac{1}{p}. \tag{11}$$

Las *distribuciones binominales negativas* generalizan la distribución geométrica en lugar de esperar por un éxito, podemos esperar por cualquier número predeterminado r de éxitos. En una secuencia de ensayos independientes Bernoulli con probabilidad de éxito p, si X es el número de *fallas* antes del éxito número r, entonces se dice que Xtiene una distribución binominal negativa con parámetros r y p, denotado $X \sim NBin(r, p)$.

La distribución binominal cuenta el número de éxitos en un número fijo de ensayos, mientras que la binominal negativa cuenta el número de fallas hasta alcanzar cierto número de éxitos. Si $X \sim NBin(r, p)$, entonces la función de probabilidad de X es

$$P(X = n) = \binom{n+r-1}{r-1} p^r q^n \tag{12}$$

para $n = 0, 1, 2 \ldots$, donde $q = 1 = p$.

0.2. Inteligencia artificial

Inteligencia artificial es definida como "el esfuerzo por automatizar tareas intelectuales normalmente realizadas por humanos"[Cho18], de este campo general se desprenden el *aprendizaje maquinal* y el *aprendizaje profundo*.

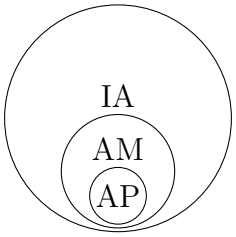

Figura 1: Aprendizaje profundo (AP), es un subcampo del aprendizaje maquinal (AM), que a su vez es un subcampo de la inteligencia artificial (IA)[Cho18].

0.2.1. Aprendizaje maquinal

La definición de aprendizaje maquinal de Tom Mitchell[Mit97] dice que "un programa de computadora aprende de experiencia E con respecto a una tarea T y una medición de rendimiento P, si su rendimiento en T, medido por P, mejora con experiencia E."

Esto dignifica que 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 "y el producto son las reglas"*..

Si no es programado explícitamente, entonces un sistema de aprendizaje maquinal es entrenado: se le presentan muchos ejemplos relevantes a una tarea, y si encuentra una estructura estadística en ellos, genera reglas para automatizar la tarea.

0.2.2. Procesamiento de lenguaje natural

El *procesamiento de lenguaje natural*, es el conjunto de métodos para hacer accesible el lenguaje humano a las computadoras[Eis19].* Existen dos enfoques en lo que debe ser su tarea central:

- Entrenar sistemas de extremo a extremo* que transmuten texto sin procesar a cualquier estructura deseada.
- Transformar texto en una pila de estructuras lingüísticas de uso general que en teoría deben poder soportar cualquier aplicación.

Dos de los módulos básicos de NLP son *búsqueda* y *aprendizaje* con los que se puede resolver muchos problemas que podemos describir en la siguiente forma matemática

$$\hat{y} = \underset{y \in Y(x)}{argmax} \Psi(x, y; 0), \tag{13}$$

donde,

- x es la entrada, un elemento de un conjunto X.
- y es el resultado, un elemento de un conjunto Y.
- Ψ es una función de puntuación (también conocida como *modelo*), que va desde el conjunto $X \times Y$ hasta los números reales.
- $\emptyset$ es el vector de parámetros para Ψ.

↑ "and out come the rules"

↑ end-to-end, de principio a fin?

↑ como que falta algo aquí?

- $\hat{y}$ es el resultado previsto, que es elegido para maximizar la función de puntuación.

El módulo de búsqueda se encarga de computar el $argmax$ de la función $\mathbf{\Psi}$, es decir, encuentra el resultado $\hat{y}$ con la mejor puntuación con respecto a la entrada x. El módulo de aprendizaje encuentra los parámetros θ por medio del procesamiento de grandes conjuntos de datos de ejemplos etiquetados $\{(x^i, y^i)\}_{i=1}^{N}$.