BOURBAKI

COLEGIO DE MATEMÁTICAS

BOURBAKI

COLEGIO DE MATEMÁTICAS

Índice

ÍNDICE

01 Introducción

Bienvenidos a nuestro curso de Matemáticas Avanzadas para la Ciencia de Datos, nuestro curso tiene cuatro módulos dedicados a estudiar las ideas matemáticas más útiles para comprender los algoritmos y modelos matemáticos más comúnes en Ciencia de Datos. Los cuatro módulos son los siguientes

- Fundamentos de probabilidad
- Álgebra Lineal
- Estadística e inferencia bayesiana
- Optimización y cálculo diferencial

Todos los módulos tienen una duración de 6 semanas. El curso está acompañado de ejercicios y tareas en Python para practicar y reforzar los conocimientos aprendidos así como las implementaciones en bases de datos de los algoritmos estudiados. Pueden consultar el repositorio de esta semana en este link.

La estructura de cada una de las semanas es la siguiente:

1. Veinte minutos dedicados a estudiar un artículo de referencia que motivará los conceptos matemáticos de esta semana.
2. Dos horas cuarenta dedicadas a estudiar el tema de la semana y algunos ejercicios.

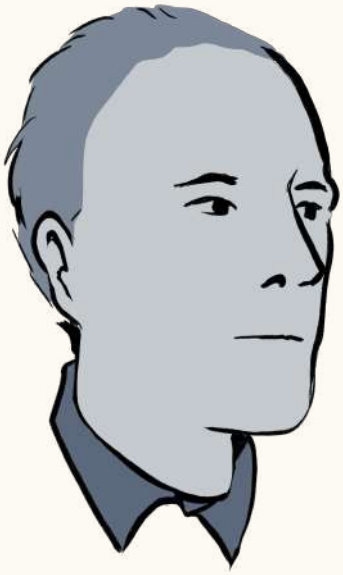
3. Una hora y media dedicada a practicar lo aprendido utilizando Python.

El primer módulo de probabilidad consta de los siguientes temas:

1. Aleatoriedad, independencia y condicionamiento.
2. Variables aleatorias, sus momentos y las regresiones.
3. Ley de los grandes números y el teorema límite central.
4. Tests estadísticos.
5. Cadenas de Markov y Page Rank.
6. Metropolis-Hastings y mensajes codificados.

🔗 El repositorio de Github para esta semana se puede encontrar en éste link.

02 Karl Pearson



En esta semana hablaremos, entre otras cosas, sobre el cálculo de la correlación entre dos variables aleatorias. La fórmula de la correlación se le atribuye al matemático inglés Karl Pearson, aunque también los matemáticos Auguste Bravais y Francis Galton participaron en el desarrollo de ella.

Karl Pearson (1857 - 1936) fue uno de los matemáticos más importantes en el desarrollo de la estadística moderna. Abogado, filósofo, escritor e historiador del arte. La mayoría de sus trabajos los desarrolló en el Kings College of London. Su influencia en la academia inglesa fue muy poderosa y por muchos es considerado el padre de la estadística.

Entre sus contribuciones está la idea de estudiar la relación de datos a través de los patrones geométricos que generan. Sus trabajos junto a Walter Eldon

acerca de evolución y características hereditarias en plantas son pioneros en lo que hoy se conoce como la bioestadística. Algunas técnicas estadísticas que requieren de grandes bases de datos, se le atribuyen a Pearson. Más adelante en el curso hablaremos sobre sus resultados teóricos sobre los tests estadísticos.

03 Lectura de referencia: las distribuciones de Laplace

En la semana 4 del curso hablaremos sobre un teorema que es llamado el Teorema Límite Central y que justifica su importancia. Por el momento solo haremos referencia a la idea intuitiva que todos tenemos cuando escuchamos o vemos una"campana de gauss".

La distribución gaussiana es ampliamente conocida y utilizada para modelar distintos fenómenos tanto en Ciencia de Datos como en otras áreas del conocimiento.

Además del matemático alemán Carl FriedrichGauss a quien le debe su nombre la distribución gaussiana, en su descubrimiento participó el matemático francés Pierre SimónLaplace quien descubrió otra distribución de probabilidad hoy conocida como distribución de Laplace. Esta distribución tiene la importante diferencia con la gaussiana que la cantidad de valores atípicos

(outliers)es mucho mayor.

Les compartimos este texto de referencia [1] sobre la distribución que siguen las mediciones del crecimiento de algunas compañías japonesas. Por medio de observaciones empíricas y algunos resultados teóricos el autor concluye que las distribuciones de Laplace son una mejor hipótesis para este fenómeno.

04 Las variables aleatorias y sus momentos

La semana pasada hablamos sobre los espacios de probabilidad y en esta lo haremos sobre las variables aleatorias, estas últimas son la base de la teoría moderna de probabilidades. A continuación explicaremos a relación entre ambos conceptos.

Definition 00.1. Una **variable aleatoria** X es un fenómeno que es aproximado por un conjunto con N observaciones numéricas generadas por el mismo fenómeno $X \approx \{x_1, x_2, \ldots, x_N\}$.

Example 00.1. *Al elegir a N personas al azar y registrar sus estaturas $\{x_1, x_2, \ldots, x_N\}$ estamos aproximando una variable aleatoria que corresponde a la población total.*

Proposition 00.2. *Asociado a una variable aleatoria X es posible construir un espacio de probabilidad sobre el conjunto Ω de valores **distintos** que toma la variable X. La fórmula de la función de probabilidad asociada a la variable X es*

$$\mathbb{P}_X(x) = \frac{n_x}{N} \tag{04.1}$$

donde n_x es el número de veces que aparece el valor x en la variable.

En términos de los registros de una base de datos X con N observaciones

$x_1, x_2, \dots x_N$, donde sólo hay k registros distintos, digamos $x_1, x_2, \dots x_k$ (con $k \leq N$), entonces su espacio de probabilidad asociado Ω_X tiene k elementos $\Omega = \{x_1, \dots x_k\}$ donde $\mathbb{P}(x_i) = \frac{n_{x_i}}{N}$. Como veremos en el siguiente ejemplo.

Example 00.3. *Supongamos que hay $N = 8$ personas, con alturas $X \approx \{1.55, 1.57, 1.57, 1.60, 1.65, 1.70, 1.70, 1.72\}$. El espacio de probabilidad asociado a la variable altura X, consta de los valores distintos $\Omega = \{1.55, 1.57, 1.60, 1.65, 1.70, 1.72\}$. La probabilidad asociada a esta variable es:*

$$\mathbb{P}_X(1.55) = \frac{1}{8}; \quad \mathbb{P}_X(1.57) = \frac{2}{8} = \frac{1}{4}; \quad \mathbb{P}_X(1.60) = \frac{1}{8}$$
$$\mathbb{P}_X(1.65) = \frac{1}{8}; \quad \mathbb{P}_X(1.70) = \frac{1}{4}; \quad \mathbb{P}_X(1.72) = \frac{1}{8}$$

La importancia del espacio de probabilidad asociado es que nos permite distinguir matemáticamente a dos variables aleatorias.

Inversamente podemos asociar una variable aleatoria a un espacio de probabilidad dado.

Proposition 00.4. *Consideremos un espacio de probabilidad (finito por el momento) $(\Omega, \mathbb{P})$, cuyos valores son númericos (números reales) es posible construir una variable aleatoria X en donde cada número $\omega \in \Omega$ aparezca en X la parte proporcional de veces correspondiente a $\mathbb{P}(\omega)$.*

Desafortunadamente existe una ambigüedad en esta asignación pues no hemos fijado el tamaño de la base de datos; sin embargo como explicaremos en las clases, esto no es un problema muy grave.

Example 00.5. *Supongamos que tenemos dado un espacio de probabilidad*

$$\left(\Omega = \{w_1, w_2\}, \mathbb{P} = \left\{\mathbb{P}(w_1) = \frac{1}{4}, \mathbb{P}(w_2) = \frac{3}{4}\right\}\right)$$

La ambigüedad que se menciona arriba se refiere a que la variable asociada a $(\Omega, \mathbb{P})$ *podría ser:* $X \approx \{x_1, x_2, x_2, x_2\}$ *o bien,* $X' \approx \{x_1, x_1, x_2, x_2, x_2, x_2, x_2, x_2\}$*, o de muchas otras maneras; por lo cual es importante fijar* N*.*

Remark 00.6. *La enorme ventaja de las variables aleatorias respecto a los espacios de probabilidad es que al constar de observaciones numéricas,podemos operar con ellas, es decir sumarlas, restarlas, etc...*

Example 00.7. *Si* $\Omega = \{2, 3, 4, 5, \ldots, 12\}$ *es el espacio de probabilidad correspondiente a la experiencia aleatoria de lanzar dos dados de manera independiente y sumarlos. Su variable aleatoria correspondiente* X_{sum} *es evidente.*

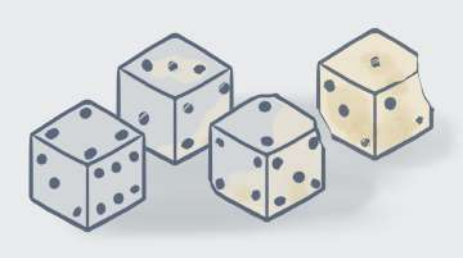
Veamos algunos ejemplos.

$$\mathbb{P}(X_{sum} = 2) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$
$$\mathbb{P}(X_{sum} = 3) = \frac{2}{36} = \frac{1}{18}$$
$$\mathbb{P}(X_{sum} = 4) = \frac{3}{36} = \frac{1}{12}$$
$$\mathbb{P}(X_{sum} = n) = \frac{\#\text{ combinaciones que suman } n}{\#\text{ de combinaciones posibles: } 6 \times 6}$$

Por ejemplo, hay dos maneras de sumar 3: Que en el primer dado salga 1 y en el segundo 2; y que en el primer dado salga 2 y en el segundo 1. Siguiendo esta misma idea, $X_{sum} = 4$ *se puede obtener de 3 maneras distintas:* $1 + 3$*,* $2 + 2$ *y*

$3+1$.

Exercise 00.8. *Calcule los valores faltantes para el ejemplo 00.7*



Recordemos la **ley de probabilidad binomial** $\mathbb{P}_{Bin_{(n,p)}}(i) = \left(\frac{n!}{(n-i)!i!}\right) p^i \left(1-p\right)^{n-i}$ con parámetros $0 \leq p \leq 1$ y n un número natural. Pensemos en una base de datos donde n corresponde al número de columnas $\{X_1, \dots X_n\}$. Pensemos que las columnas son independientes entre sí, y que satisfacen la ley de probabilidad de Bernouli

$$\left(\Omega = \{0,1\}, \mathbb{P}_{Bernulli} = \{1-p, p\}\right)$$

es decir, p es la proporción de unos en cada columna. En este escenario, $\mathbb{P}_{Bin_{(n,p)}}(i)$ corresponde a la probabilidad de obtener renglones con exactamente i unos.

Example 00.9. *Con las $n = 4$ columnas de la siguiente tabla donde cada una tiene probabilidad $\mathbb{P}_{Bernouli}(X_i) = \frac{3}{5}$. Entonces, para $i = 2$, $\mathbb{P}_{Bin_{(n,p)}}(2)$ es la probabilidad de obener arreglos (permutaciones de las columnas) en los que aparecen renglones con exactamente dos unos.*

X_1	X_2	X_3	X_4
1	*1*	*1*	*1*
0	*1*	*1*	*1*
0	*0*	*1*	*0*
1	*0*	*0*	*1*
1	*1*	*0*	*0*

Exercise 00.10. *Determine el espacio de probabilidad asociado a la suma de n variables aleatorias $X_1, X_2, \dots X_n$ de bernoulli independientes. A esta variable la denotamos como*

$$S_n = \sum_{k=1}^{n} X_k \tag{04.2}$$

Valor esperado y varianza

Definition 01.1. *Si X es una variable aleatoria y $(\Omega, \mathbb{P}_X)$ su espacio de probabilidad asociado, definimos el **valor esperado** de X de la siguiente forma:*

$$\mathbb{E}[X] = \sum_{x \in \Omega} x \cdot p_x \tag{04.3}$$

Donde $p_x = \mathbb{P}(x)$.

Example 01.2. *Sea X la variable aleatoria de Bernoulli $(\{0,1\},\mathbb{P})$, entonces*

$$\mathbb{E}[X] = \big(0\cdot\big(1-p\big)\big) + \big(1\cdot p\big) = p \tag{04.4}$$

Example 01.3. *Consideremos la variable X_1 de la tabla del ejemplo 00.9. Entonces $\mathbb{E}(X_1) = \frac{3}{5}$.*

Exercise 01.4. *Defina dos variables aleatorias que corresponda a las 10 calificaciones durante un semestre de dos alumnos, todas ellas entre 5 y 10 y calcule sus esperanzas.*

Exercise 01.5. *Calcular la esperanza de X_{sum} del ejemplo 00.7.*

Una propiedad muy importante de la esperanza es que es un **funcional lineal**, esto significa que respeta la suma de variables y producto por escalares en el siguiente sentido:

Proposition 01.6. *Si X, Y son dos variables aleatorias y $a, b \in \mathbb{R}$, entonces $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$.*

Proposition 01.7. *Sea $\left(\{0,1\}^n, \mathbb{P}_{Bin_{(n,p)}}\right)$ la ley de probabilidad binomial (definida en las notas de la semana pasada). Llamemos S_n la variable aleatoria asociada a dicho espacio. Entonces $\mathbb{E}[S_n] = np$.*

Example 01.8. *En el ejemplo de las olimpiadas que vimos en la notas pasadas, la esperanza es $307 \times 0.005 = 1.535$ que es lo que se esperaría en promedio de medallas olímpicas ganadas por Argentina de acuerdo al tamaño de su población.*

Definition 01.9. *Sea X una variable aleatoria. Definimos **la varianza** de X como*

$$Var(X) = \sum_{x \in \Omega} (x - \mathbb{E}(X))^2 \cdot p_x \tag{04.5}$$

Definition 01.1. Definimos la **desviación estándar** de una variable aleatoria como la raíz positiva de la varianza

$$\sigma_X = \sqrt{Var(X)} \tag{04.6}$$

Example 01.10. *Sea X la variable aleatoria de Bernoulli. Utilizando el resutlado 04.4 tenemos que*

$$Var(X) = (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p = p \cdot (1-p)$$

Exercise 01.11. *Calcule la varianza de la variable aleatoria definida en el ejercicio 01.4.*

Exercise 01.12. *Calcule la varianza de la variable X_{sum} definida en el ejemplo 00.7.*

Remark 01.13. *La volatilidad o riesgo de un activo financiero se mide mediante la desviación estándar cuando se miden los rendimientos.*

La curtosis

En ocasiones el valor esperado y la varianza no son suficientes para determinar de cuál distribución estamos hablando. En esta sección proponemos utilizar el cuarto momento de una variable aleatoria,el cual está íntimamente relacionado con los valores extremos.

Definition 02.1. Sea $X = \{x_1, x_2, \ldots, x_N\}$ una variable aleatoria, definimos la **curtosis** de X

$$\kappa_X = \sum_{x \in X} \left(\frac{x - \mathbb{E}(X)}{\sigma} \right)^4 \cdot p_x \tag{04.7}$$

Notemos que la curtosis escrita de esta manera es siempre positiva, pues es una suma de números elevados a la cuarta. Más aún, notemos que si existen datos x_i muy alejados de $\mathbb{E}[X]$, entonces los factores $(X - \mathbb{E}[X])^4$ serán números muy grandes, por lo que una curtosis grande puede indicar una mayor cantidad de datos alejados de la media hacia uno u otro lado.

Example 02.1. *La curtosis de una distribución gaussiana centrada en cero y con varianza uno es igual a cero.*

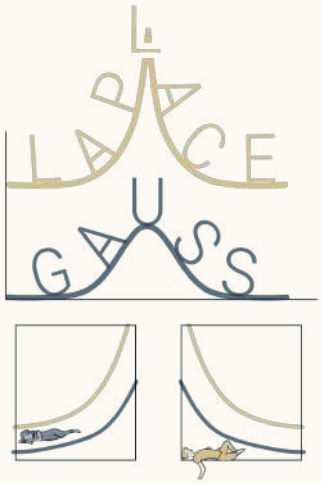
$$\kappa_X = \sum_{x \in X} \left(\frac{x - 0}{1} \right)^4 \cdot p_x = \mathbb{E}[X] = 0$$

Aunque más adelante estudiaremos detalladamente las distribuciones continuas, en esta sección vale la pena mencionar a la distribución de Laplace pues es un importante ejemplo de cómo la curtosis es útil para distinguir dos distribuciones.

Definition 02.2. Distribución de Laplace. La variable aleatoria de Laplace con media μ y varianza $2b^2$ se define el fenómeno de obtener un número x en la recta real con la siguiente probabilidad.

$$\mathbb{P}_{Laplace}(-\infty, x) = \frac{1}{2b}\int_{-\infty}^{x} e^{-\frac{|t-\mu|}{b}}\, dt \tag{04.8}$$

La curtosis de la curva normal estándar es $K_{Normal} = 3$. Se define la curtosis K_3, también conocida como **el exceso de curtosis** como $K_3(X) = K(X) - 3$ para determinar qué tan alejada está la gráfica de frecuencias de X de una distribución normal.



Example 02.2. *El exceso de curtosis de una distribución de Laplace centrada en cero y con varianza uno es igual a tres.*

Momentos para vectores aleatorios

Hasta el momento solo hemos hablado de las variables aleatorias sin embargo es común que en ciencia de datos tratemos vectores aleatorios- Comenzaremos con el caso de dos variables aleatorias.

Definition 03.1. Sean X, Y dos variables aleatorias. Definimos la **covarianza** entre X y Y como

$$Cov(X,Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X) \cdot \mathbb{E}(Y) \tag{04.9}$$

Example 03.1. *Consideremos dos variables aleatorias independientes idénticamente distribuidas, o bien, columnas de una base de datos*

X	Y
4	5
8	1
2	3
5	4
1	2

Calculemos sus medias $\mathbb{E}(X) = \mu_X = \frac{1}{5}(4+8+2+5+1) = \frac{20}{5} = 4$ *y* $\mathbb{E}(Y) = \mu(Y) = \frac{1}{5}(5+1+3+4+2) = \frac{15}{5} = 3$. *Calculemos también el producto entrada a entrada* $X \cdot Y = \{20, 8, 6, 20, 2\}$ *y su promedio* $\mathbb{E}(X \cdot Y) = \mu_{X \cdot Y} = \frac{56}{5}$. *Entonces la covarianza es* $Cov(X,Y) = E(X \cdot Y) - E(X) \cdot E(Y) = \frac{56}{5} - 12 = -0.8$

Exercise 03.2. *Verifique que* $Cov(X,X) = Var(X)$.

Exercise 03.3. *Verifique que Si* X, Y *son variables aleatorias independientes entonces* $Cov(X,Y) = 0$.

Definition 03.2. Si X, Y son dos variables aleatorias, definimos su **correlación** como

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} \tag{04.10}$$

Example 03.4. *Calculemos la varianza de cada una de las variables del ejemplo 03.1. Si eliminamos la media de cada columna obtenemos* $X - \mu_X = \{0,4,-2,1,-3\}$ *y* $Y - \mu_Y = \{2,-2,0,1,-1\}$ *y luego elevemos todo al cuadrado* $(X-\mu_X)^2 = \{0,16,4,1,9\}$, $(Y-\mu_Y)^2 = \{4,4,0,1,1\}$*; entonces la varianza es el promedio* $Var(X) = \frac{0+16+4+1+9}{5} = 6$ *y* $Var(Y) = \frac{4+4+0+1+1}{5} = 2$. *Con lo que las desviaciones estándar son* $\sigma_X = \sqrt{6}$ *y* $\sigma_Y = \sqrt{2}$. *Finalmente, su coeficiente de correlación es* $Corr(X,Y) = \frac{-0.8}{\sqrt{6}\cdot\sqrt{2}} = -0.23$.

Proposition 03.5. *Entre más cercana a* -1 *o* $+1$ *sea la correlación entre dos variables* $X = \{x_1,\ldots,x_N\}$, $Y = \{y_1,\ldots,y_N\}$ *más cercana de* ***estar en regresión lineal simple*** *estará la base de datos* $\left\{(x_1,y_1),\ldots,(x_N,y_N)\right\}$. *De hecho, los coeficientes de la regresión son:*

$$\beta_0 = \frac{Cov(X,Y)}{Var(X)} y \beta_1 = \mu_Y - \beta_0 \mu_X$$

donde μ_X *y* μ_Y *representan las medias de X y Y respectivamente.*

Correlación de Spearman

La correlación de Spearman, al igual que la correlación de Pearson, también mide la relación entre dos variables aleatorias X y Y. A diferencia de la correlación de Pearson, la de Spearman trabaja con los **estadísticos de orden** de los datos $x - y$.

El primer estadístico de orden de una muestra de tamaño n es el siguiente mínimo. Del mismo modo, el n-ésimo estadístico de orden n es el máximo:

$$X_{(1)} = \text{mín}\{X_1, X_2, \dots X_n\}$$

$$X_{(n)} = \text{máx}\{X_1, X_2, \dots X_n\}$$

El rango de la muestra es la diferencia:

$$Rango\{X_1, \dots, X_n\} = X_{(n)} - X_{(1)}$$

El coeficiente de correlación de Spearman ρ se calcula como

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} \tag{04.11}$$

Donde d_i son las diferencias entre los rangos de ambas variables en cada observación $d_i = x_{(i)} - y_{(i)}$.

El coeficiente ρ mide las relaciones monótonas entre X y Y, es decir, detecta si una de las variables aumenta o disminuye cuando la otra lo hace; sin asumir que dicha relación es lineal (como en la correlación de Pearson).

Example 04.1. *Calculemos la correlación de Spearman para las variables del ejemplo 03.1, para lo que será útil observar la siguiente tabla.*

X	Y	Rango X	Rango Y	d (diferencia)	d^2
4	5	3	5	-2	4
8	1	5	1	4	16
2	3	2	3	-1	1
5	4	4	4	0	0
1	2	1	2	-1	1

Entonces el coeficiente de correlación de Spearman se calcula como

$$\rho = 1 - \frac{6 \cdot \sum d^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot (4 + 16 + 1 + 0 + 1)}{5 \cdot (5^2 - 1)}$$
$$= 1 - \frac{6 \cdot 22}{5 \cdot 24} = 1 - \frac{22}{20} = 1 - \frac{11}{10} = -\frac{1}{10} = -0.1$$

Remark 04.2. *El coeficiente de correlación de Spearman varía en el intervalo [-1, 1]. Cuando $\rho = -1$, hay una correlación monótona perfecta negativa. Cuando $\rho = 1$ se presenta una correlación monótona perfecta positiva, y cuando $\rho = 0$ no hay relación monótona entre las variables.*

Skewness

El skewness o asimetría mide qué tan simétrica es la distribución de una variable aleatoria con valores reales. Hay distintas maneras de abordar este problema, pero entre las más usuales están el coeficiente de asimetría de Fisher y el coeficiente de asimetría de Pearson.

Definition 05.1. El coeficiente de asimetría (poblacional) de Fisher para una

variable aleatoria X se calcula como:

$$\gamma_{Fisher} = \frac{E[(X-\mu)^3]}{\sigma^3} \tag{04.12}$$

Donde μ es la media, σ es la desviación estándar

Definition 05.2. El coeficiente de asimetría de Pearson se calcula como

$$\gamma_{Pearson} = \frac{\mu - Mo}{\sigma} \tag{04.13}$$

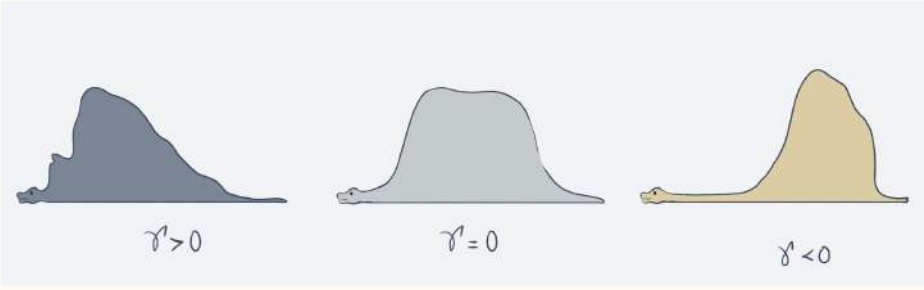
donde μ es la media y Mo es la moda de X.

Hablemos de la interpretación del signo del índice de asimetría

$\gamma = 0$ Estamos en presencia de una distribución simétrica.

$\gamma > 0$ La distribución tiene una cola más larga hacia la derecha.

$\gamma < 0$ La distribución de X muestra una cola más larga a la izquierda.



Example 05.1. *El coeficiente de asimetría de Fisher de la variable X_{sum} defi-*

nida en el ejemplo 00.7 se obtiene fácilmente:

$$E[X_{sum}] = \frac{252}{36} = 7 \quad \text{y} \quad E[X_{sum}^2] = \frac{1974}{36}$$
$$Var(X_{sum}) = \frac{1974}{36} - 7^2 = \frac{1974 - 1764}{36} = \frac{35}{6} \quad \text{y} \quad \sigma = \sqrt{\frac{35}{6}},$$
$$E[(X-\mu)^3] = \frac{-125-128-81-32+5+0+5+32+81+128+125}{36} = 0,$$
$$\gamma_{Fisher} = \frac{E[(X-\mu)^3]}{\sigma^3} = \frac{0}{\sigma^3} = 0$$

Lo que indica que la distribución X_{sum} es simétrica.

Ley multinomial

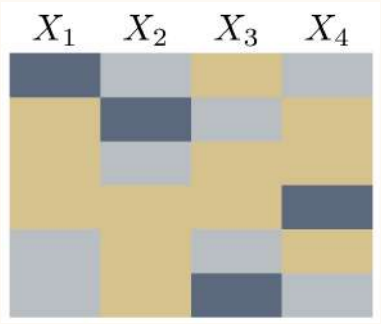
Example 06.1. ***Ley multinomial****. Fijemos K y n dos números naturales y Ω un conjunto de tamaño n^K, digamos $\Omega = \{x_1, x_2, \ldots, x_n\}^K$.Además fijemos números $0 \leq p_1, p_2, \ldots, p_K \leq 1$ cuya suma es igual a uno. Definimos la ley de probabilidad multinomial de la siguiente forma:*

$$\mathbb{P}_{Mult_{(n,p_1,\ldots,p_K)}}(x_1, x_2, \ldots, x_K) = \left(\frac{n!}{x_1!x_2!\ldots x_K!}\right) p_1^{x_1} \ldots p_K^{x_K} \tag{04.14}$$

Pensemos en una base de datos X con n columnas $X_1, X_2, \ldots, X_n$ y N renglones. En cada columna podemos tener solamente K observaciones. Éstas observaciones podrían ser incluso variables categóricas, digamos $c_1, c_2 \ldots, c_k$. Llamemos p_i a la probabilidad de que aparezca la observación c_i en cada columna. En este contexto $\mathbb{P}(x_1, x_2, x_3)$ es la probabilidad de encontrar **renglo-**

nes con exactamente x_1 observaciones c_1, x_2 observaciones c_2 y x_3 observaciones c_3.

Por ejemplo, si $K = 3$, $n = 4$, $N = 5$, c_1 es oscuro y $p_1 = \frac{1}{5}$, c_2 es dorado y $p_2 = \frac{3}{5}$ y c_3 es gris claro y $p_3 = \frac{2}{5}$. La siguiente tabla muestra que hay probabilidad de obtener arreglos con renglones para (x_1, x_2, x_3) de la forma $(1,1,2)$ (de hecho, el primero y el último de los renglones tienen esta configuración), también tenemos posibilidad para $(1,2,1)$, $(0,3,1)$,$(1,3,0)$ y $(0,2,2)$.



Proposition 06.2. *Si X_i, X_j son algunas de las K columnas distintas donde viven los valores de la distribución multinomial, entonces $Cov(X_i, X_j) = -np_ip_j$.*

05 Distribuciones infinitas

Siméon Denis Poisson

Nació en Francia solo unos años antes de la revolución francesa, estudió en una de las escuelas más emblemáticas de toda la historia de Francia, École Polytechnique. Sus trabajos científicos están relacionados con la Teoría del Potencial, con la Óptica, con la Mecánica y por supuesto con la Teoría de la Probabilidad.

Estudiando los juicios criminales en materia civil desde el punto de vista de la probabilidad dedujo lo que actualmente se conoce como ley de Poisson la cual estudiaremos más adelante en este texto.

Sus interacciones con otros matemáticos de la época son naturales debido al gran prestigio que cosechó durante su carrera académica, sustituyendo en

algún momento al mismísimo Joseph Fourier. Una de estas interacciones es bastante notable, con el joven y genial Évariste Galois a quien invitó a publicar su trabajo completo a pesar de describirlo como incomprensible.

Lectura de referencia: La maldición de la Dimensión y MLE

Uno de los problemas más complicados dentro de Machine Learning es el de la maldición de la dimensión pues cuando el espacio de parámetros es demasiado grande la búsqueda de los valores óptimos puede ser tan complicada como imposible gracias a la compleja geometría de estos espacios.

La técnica de aproximación que veremos en este capítulo se le conoce como máxima verosimilitud y es ampliamente conocida en estadística como un método para aproximar parámetros. Esta técnica al igual que otros métodos de machine learning también se ven afectados por la maldición de la dimensión.

En esta lectura de referencia recomendamos el siguiente artículo [2] en el que los autores introducen algunas mejoras a la clásica regresión logística para aproximar los parámetros. Aunque en esta semana no estudiaremos la regresión logística sino la regresión de Poisson, ambos métodos son muy parecidos.

Distribuciones infinitas

Hasta ahora hemos hecho énfasis de espacios de probabilidad finitos sin embargo las variables aleatorias son mucho más expresivas para espacios de probabilidad infinitos. En esta sección nos dedicaremos a estudiar variables aleatorias un poco más complicadas.

Example 03.1. *Si E es la experiencia aleatoria de lanzar un dado justo hasta obtener el número seis, definimos la siguiente variable aleatoria: $X(\omega_1, \omega_2, \ldots) = \{w_1, w_2, \ldots w_k\}$ donde k es el menor número de tiradas del lado en las que obtuvimos un 6, es decir, $k = \min_{j \geq 1} \left\{ j : \omega_j = 6 \right\}$.*

Definiremos la siguiente experiencia aleatoria numerable (esto es, cuando Ω tiene tantos elementos como números naturales $\Omega = \{1, 2, 3, \ldots\}$), llamada **Ley de Poisson.**

Definition 03.2. *Sea $\lambda > 0$, definimos*

$$\mathbb{P}_{Poisson,\lambda}(i) = \frac{e^{-\lambda} \cdot \lambda^i}{i!} \tag{05.1}$$

Proposition 03.3. *Para una variable aleatoria de Poisson X con parámetro λ,la media y la varianza son $\mathbb{E}[X] = Var[X] = \lambda$*

La ley anterior corresponde a la probabilidad de que un evento raro ocurra después de muchas repeticiones. La justificación matemática de esta intuición es la siguiente proposición:

Proposition 03.4. *Consideremos una distribución binomial $Bin_{(n,p_n)}$ en la que* $\lim_{n\to\infty} p_n \cdot n = \lambda$. *Entonces* $\lim_{n\to\infty} \mathbb{P}_{Bin(p_n,n)}(i) = \mathbb{P}_{Poisson,\lambda}(i)$.

La condición de la proposición anterior dice que cuando n crece, las probabilidades p_n deben ser cada vez más pequeñas, pues de no ser así, el límite $\lim_{n\to\infty} P_n \cdot n$ sería infinito.

Example 03.5. *Pensemos en el evento de tirar n monedas donde ganar significa que salga águila. La condición correspondería a que entre más monedas lancemos, menos probabilidades tenemos de ganar.*

Example 03.6. *Comparemos la distribución de Poisson con la ley binomial que calculamos en el ejemplo de las olimpiadas en la primera semana, en este caso nuestro parámetro λ será igual a la esperanza de la variable aleatoria de Bernouilli, lo cuál corresponde con* 1.535 *gracias al cálculo que hicimos en aquellas notas, tenemos:*

$$\mathbb{P}_{Poisson,1.535}(0) = 0.215, \qquad \mathbb{P}_{Poisson,1.535}(1) = 0.33,$$

$$\mathbb{P}_{Poisson,1.535}(2) = 0.253, \quad \mathbb{P}_{Poisson,1.535}(3) = 0.129$$

Exercise 03.7. *Calcule la curtosis de una distribución de Poisson con parámetro λ.*

Notemos que si en lugar de considerar la variable aleatoria S_n consideramos la variable aleatoria $n - S_n$ (o equivalentemente la variable aleatoria de Poisson con $\lambda = n(1-p)$ es posible aproximar de la misma manera eventos alta-

mente probables. Una pregunta inmediata es ¿qué pasa si deseamos calcular eventos cuya probabilidad no es muy pequeña ni muy alta? Para ello será necesario utilizar el célebre Teorema Límite Central de Lévy.

03.1 Leyes de probabilidad continuas

Las leyes de probabilidad continuas (es decir definidas sobre el conjunto total de los números reales) son más complicadas de definir porque en ese caso las funciones de probabilidad no actúan sobre la familia total de subconjuntos, si lo hicieran esto generaría algunos problemas matemáticos los cuales trascienden el objetivo de este curso. En esta sección hablaremos de las llamadas leyes continuas con densidad.

Definition 03.1. La ley de probabilidad uniforme sobre el intervalo de números reales $[a,b]$ es la ley de probabilidad tal que

$$\mathbb{P}_{unif_{((a,b))}}(x) = \begin{cases} 0 & \text{si } x \leq a \\ \frac{x-a}{b-a} & \text{si } a < x < b \\ 1 & \text{si } x \geq b \end{cases} \tag{05.2}$$

Example 03.8. *La variable aleatoria uniforme del tiempo medio (en minutos) de entrega de un pedido en una cafetería está distribuida en el intervalo* $[5,15]$. *La probabilidad de que su café se entregue en menos de 7 minutos es de* $\mathbb{P}_{unif_{[5,15]}}(7) = \frac{7-5}{15-5} = \frac{1}{5}$

Definition 03.2. La ley de probabilidad Gaussiana o normal con parámetros $\left(\mu, \sigma^2\right)$ se define para los intervalos $(-\infty, x]$ de la siguiente manera:

$$\mathbb{P}_{Gauss_{(\mu,\sigma^2)}}(-\infty, x] = \frac{1}{\sigma \cdot \sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\cdot\sigma^2}} dt \tag{05.3}$$

En estas notas no hemos definido la esperanza ni la covarianza para leyes de probabilidad no numerables, sin embargo es posible hacerlo:

Proposition 03.9. *Si X es una variable aleatoria con probabilidad continua uniforme $\mathbb{P}_{unif_{((a,b))}}(X)$, entonces $\mathbb{E}(X) = \frac{a+b}{2}$ y $Var(X) = \frac{(b-a)^2}{12}$.*

Proposition 03.10. *Si X es una variable aleatoria tal que $\mathbb{P}_X = \mathbb{P}_{Gauss_{(\mu,\sigma)}}$ entonces $\mathbb{E}(X) = \mu$ y $Var(X) = \sigma^2$.*

Example 03.11. ***Distribución gaussiana multivariada***

Sea $X = (X_1, X_2, \ldots, X_d)$ una variable aleatoria con $E[X_i] = \mu_i$, denotamos por $\mu = (\mu_1, \mu_2, \ldots, \mu_d) \in \mathbb{R}^d$ al vector de medias. Sea $C \in \mathbb{R}^{d\times d}$ la matriz de covarianza de X, es decir, $C_{i.j} = Cov(X_i, X_j)$ (de hecho, C es simétrica $C = C^T$ y satisface que $xCx^T > 0$ para cualquier $x \in \mathbb{R}^d \backslash \{\bar{0}\}$). Definimos la ley de probabilidad gaussiana d dimensional con parámetros μ, C de la siguiente forma:

$$\mathbb{P}_{Gauss}((-\infty, x_1] \times \ldots \times (-\infty, x_d]) = \frac{1}{(2\pi)^{d/2}|C|^{d/2}} \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_d} e^{-\frac{1}{2}(t-\mu)C^{-1}(t-\mu)^T} dt$$

Regresiones de Poisson y el aprendizaje supervisado

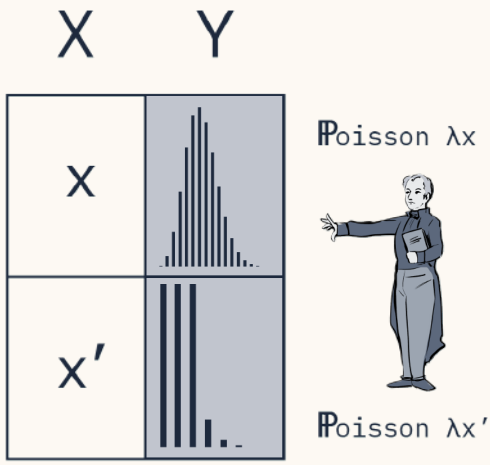
En esta semana por primera vez hablaremos sobre un ejemplo de aprendizaje supervisado en el que tendremos acceso a un muestreo S de lo que se conoce como una **probabilidad conjunta** de dos variables aleatorias (X, Y). Utilizando a S deseamos averiguar una relación funcional que exista entre X e Y, a saber una función f de la variable X tal que en la medida de lo posible expliquen a Y, es decir $f(X) \sim Y$. De ese hecho viene la relación de la regresión con el aprendizaje supervisado. Si X es una base de datos con variable supervisada Y. Se espera que la predicción $f(X)$ se aproxime a la supervisión Y. El ejemplo más sencillo de este problema es el de las regresiones de Poisson.

En el caso de una **regresión de Poisson** se asumirá la siguiente fórmula:

$$\log(\mathbb{E}[Y|X]) = \langle X, \beta \rangle \tag{05.4}$$

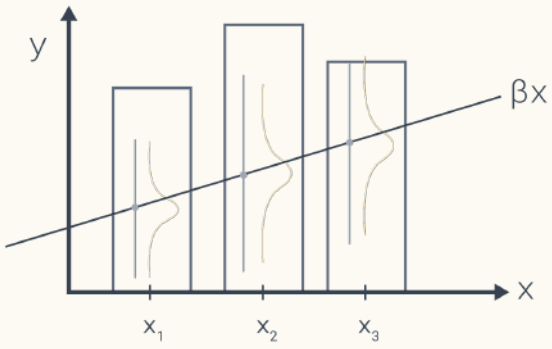
Lo anterior es equivalente a suponer que la variable Y sigue una distribución de Poisson al condicionarla con X.

Example 04.1. *En una base de datos X con d características y variable objetivo Y, en donde $\mathbb{P}(Y = n|(x_1, \ldots, x_d) = \frac{e^{-\lambda} \cdots \lambda^n}{n!}$, es decir, tenemos una distribución de Poisson para cada $(x_1, \ldots, x_d)$, entonces λ está dado por los x_i como $\lambda_{(x_1,\ldots,x_d)} = e^{\beta_1 x_1 + \cdots + \beta_d x_d}$ y además $\mathbb{E}[Y|(x_1, x_2, \ldots, x_d)] = e^{\beta_1 x_1 + \cdots + \beta_d x_d}$. Aquí los betas son precisamente los coeficientes de la regresión.*



Example 04.2. Regresión lineal. *Tradicionalmente en las regresiones lineales se intenta modelar* $\mathbb{E}[Y|X]$ *con una combinación lineal de las variables explicativas* X.

En una base de datos X *con* d *características y variable objetivo* Y *con distribución gaussiana, es decir* $\mathbb{P}(Y = y|(x_1, \dots x_d)) = N[\beta_1 x_1 + \dots + \beta_d x_d, \sigma]$ *es gaussiana, entonces el valor esperado* $\mathbb{E}[Y|(x_1, \dots x_d)] = \beta_1 x_1 + \dots + \beta_d x_d$.



Machine Learning Generativo

Muchos de los modelos matemáticos buscan modelar la distribución $\mathbb{P}\left(y|x\right)$ donde $x \in \mathbb{R}^d$ e $y \in \{-1, +1\}$. A este tipo de modelos los llamaremos modelos discriminativos o modelos de clasificación.

En contraposición a estos modelos están los modelos generativos los cuales buscan modelar la distribución de probabilidad $\mathbb{P}\left(x|y\right)$. Es decir, a partir de supervisiones fijas, se busca generar las características que se apegan a esas supervisiones. Gracias al teorema de Bayes, si conocemos las probabilidades $\mathbb{P}\left(x|y\right)$ y $\mathbb{P}\left(y\right)$ es posible deducir

$$\mathbb{P}\left(y|x\right) = \frac{\mathbb{P}\left(x|y\right)\cdot\mathbb{P}\left(y\right)}{\mathbb{P}\left(x\right)} = \frac{\mathbb{P}\left(x|y\right)\cdot\mathbb{P}\left(y\right)}{\mathbb{P}\left(x|y=-1\right)\mathbb{P}\left(y=-1\right)+\mathbb{P}\left(x|y=+1\right)\mathbb{P}\left(y=+1\right)}$$

La última igualdad ocurre gracias a la igualdad llamada Regla de Bayes, para los eventos mutuamente excluyentes $y = 1$ y $y = -1$.

Proposition 05.1. ***Regla de Bayes****. Si $\{A_1, A_2, \dots A_n\}$ son eventos mutuamente excluyentes con probabilidad distinta de cero, y X es un evento del que se conocen las probabilidades condicionales $\mathbb{P}(X|A_i)$, entonces podemos conocer las probabilidades condicionales*

$$\mathbb{P}(A_i|X) = \frac{\mathbb{P}(X|A_i)\cdot\mathbb{P}(A_i)}{\sum_{k=1}^{n}\mathbb{P}(X|A_k)\cdot\mathbb{P}(A_k)}$$

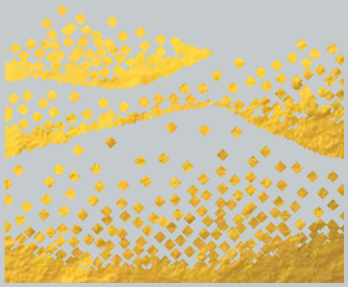
Los modelos generativos no son un cambio radical sin embargo en la prác-

tica para algunos casos resulta sorprendentemente útil modelarlos de esta manera. En lo particular para el problema de topic modeling resulta mucho más eficaz pues dado un prior (por ejemplo $y = -1$) deseamos modelar la frecuencia de las palabras que ahí aparecen.

Bibliografía

[1] Y. Arata, *Journal of Economic Dynamics and Control* **2019**, *103*, 63-82.

[2] P. Sur y E. J. Candè s, *Proceedings of the National Academy of Sciences* **2019**, *116*, 14516-14525.



escuela-bourbaki.com