



BOURBAKI

COLEGIO DE MATEMÁTICAS

Índice

01. Introducción	_____	pág. 02
02. Lectura de referencia: la aguja de buffon y el método monte carlo	_____	pág. 04
03. La ley de los grandes números	_____	pág. 05
01. La ley de los grandes números	—	pág. 05
02. Estadística uniforme y equidistribuida		
pág. 06		
03. La aguja de Buffon re-visitada	—	pág. 07
04. Máxima verosimilitud	_____	pág. 09
05. Distribuciones infinitas	_____	pág. 10

01 Introducción

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

- Probabilidad y Estadística
- Álgebra Lineal
- Optimización y cálculo diferencial

Todos los módulos tienen una duración de 8 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. Treinta minutos dedicados a estudiar un artículo de referencia que motivará los conceptos matemáticos de esta semana.
2. Dos horas 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. Kolmogorov, independencia y condicionamiento
2. Variables aleatorias discretas y sus momentos
3. Ley de los grandes números y máxima verosimilitud
4. El teorema límite central y los intervalos de confianza
5. Tests estadísticos
6. Inferencia bayesiana
7. Cadenas de markov y muestreros de gibbs
8. Redes bayesianas

02 Lectura de referencia: la aguja de Buffon y el método monte carlo

En 1977 Georges-Louis Leclerc, Comte de Buffon planteó el siguiente problema.

Definition 00.1. (La aguja de Buffon) Supongamos que tenemos un tablero arbitrariamente grande donde los únicos dibujos que hay son líneas paralelas verticales a distancia d entre ellas. Además debemos de imaginar que contamos con una cantidad arbitrariamente grande de agujas a_1, a_2, \dots todas de longitud l .

Si lanzamos una aguja aleatoriamente en nuestro tablero, ¿cuál es la probabilidad de que la aguja toque alguna de nuestras rectas paralelas?

El mismo Buffon resolvió el problema y demostró el siguiente teorema del cual pueden leer más en [1].

Theorem 00.1. (Buffon) *La probabilidad de que una aguja lanzada al azar toque a alguna de las rectas paralelas es igual a $\frac{2l}{\pi d}$*

Un aspecto muy interesante sobre este resultado es que es posible realizar experimentos para comprobarlo (más no demostrarlo). Estos resultados se basan en una versión sencilla del método de monte carlo.

03 La ley de los grandes números

Hasta el momento en este curso hemos hecho una identificación entre las variables aleatorias X y un conjunto de experimentos $S = \{x_1, \dots, x_N\}$, sin embargo no hemos justificado matemáticamente por qué es posible hacerlo. En esta sección hablaremos de uno de los resultados en probabilidad más conocidos el cual permite justificar la identificación que hemos hecho hasta ahora.

La ley de los grandes números

En esta sección hablaremos sobre cómo una familia de experimentos puede aproximar correctamente a una variable aleatoria.

Theorem 01.1. (*Ley fuerte de los grandes números*) Sean (Ω, F, \mathbb{P}) un espacio de probabilidad y $X_n : \Omega \rightarrow \mathbb{R}$ una familia de variables aleatorias independientes e idénticamente distribuidas tales que $\mathbb{E}[|X_i|] < \infty$ entonces existe un subconjunto $N \in F$ tal que $\mathbb{P}_X(N) = 0$ y además:

$$\lim_{n \rightarrow \infty} \left(\frac{X_1(m) + \dots + X_n(m)}{n} \right) = \mathbb{E}[X_1], \forall m \notin N$$

Una de las aplicaciones más importantes de la ley de los grandes números

es el llamado método de Monte Carlo.

Definition 01.1. (Método Monte Carlo) Supongamos que queremos calcular cierto parámetro μ . El método de Monte Carlo es una aplicación inmediata del resultado anterior, consiste en los siguientes pasos:

1. Definimos variables aleatorias i.d.d. cuya esperanza coincide con μ .
2. Construimos un muestreo de tamaño arbitrario de nuestras variables aleatorias.
3. Gracias a la Ley de los grandes números el promedio empírico de este muestreo cada vez se acercará más a μ

Estadística uniforme y equidistribuida

Definition 02.1. Sea $S = \{x_1, \dots, x_N\} \subseteq \mathbb{R}^d$, definimos:

1. El promedio empírico de S como $\mu_S = \frac{1}{N}(x_1 + \dots + x_N)$.
2. El promedio empírico de cada una de las $j \leq d$ variables como $\mu_{S,j} = \frac{1}{N}(x_{1,j} + \dots + x_{N,j})$
3. La varianza empírica de S se define como $Var(S) = \frac{1}{N} \sum_{i \leq N} (x_i - \mu_S)^2$
4. La covarianza empírica de S se define como una matriz simétrica de $d \times d$ entradas $(Cov(S))_{j,i} = (Cov(S))_{i,j} = \frac{1}{N-1} \sum_{k \leq N} (x_{k,i} - \mu_{S,i})(x_{k,j} - \mu_{S,j})$

5. La correlación empírica de S se define como una matriz simétrica de

$$d \times d \text{ entradas } (\text{Corr}(S))_{j,i} = (\text{Corr}(S))_{i,j} = \frac{(\text{Cov}(S))_{j,i}}{\sqrt{(\text{Cov}(S))_{j,j}} \sqrt{(\text{Cov}(S))_{i,i}}}$$

Gracias a la Ley Fuerte de los Grandes Números (y quizás algún argumento de continuidad) podemos concluir lo siguiente:

Proposition 02.1. *Supongamos que $S = \{x_1, \dots, x_N\} \subseteq \mathbb{R}^d$ está generado por N variables aleatorias X_1, \dots, X_N independientes e idénticamente distribuidas tales que $\mathbb{E}[|X_i|] < \infty$, entonces existe un subconjunto medible N tal que $\mathbb{P}_X(N) = 0$ y además:*

$$1. \lim_{n \rightarrow \infty} \mu_S(x) = \mathbb{E}[X_1], \forall x \notin N$$

$$2. \lim_{n \rightarrow \infty} (\text{Var}(S)(x)) = \text{Var}(X_1), \forall x \notin N$$

$$3. \lim_{n \rightarrow \infty} (\text{Cov}(S)_{i,j}) = \text{Cov}(X_i, X_j)$$

$$4. \lim_{n \rightarrow \infty} (\text{Corr}(S)_{i,j}) = \text{Corr}(X_i, X_j)$$

La aguja de Buffon re-visitada

Usualmente se nos presenta al número π como la proporción que hay entre el diámetro de un círculo y la longitud de su perímetro, Arquímedes fue la primera persona en aproximar el valor de π .

En 1812 el matemático francés Laplace observó que era posible aproximar el valor de π utilizando el Teorema de Buffon mediante un resultado

conocido en la teoría moderna de la probabilidad como la Ley de los Grandes Números .

El experimento tipo Monte Carlo que propone Laplace es muy sencillo:

1. Reunir tantas agujas de la misma longitud l como sea posible, digamos N .
2. Dibujar en una hoja de papel tantas rectas paralelas verticales como sea posible con la siguiente indicación: la distancia entre las rectas debe de ser exactamente igual a $2l$.
3. Lanzar las agujas en nuestra hoja siguiendo las siguientes indicaciones:
 - a) (Independencia) No podemos lanzar más de una aguja al mismo tiempo.
 - b) (Idénticamente distribuido) Debemos lanzarlas de manera aleatoria y asegurarnos de que para cualquier lugar A de la hoja dos agujas tienen la misma probabilidad de caer ahí.
 - c) (Uniforme) Debemos asegurarnos de no priorizar los lanzamientos en ninguna zona de la hoja, idealmente la aguja podría caer en cualquier lado.
4. Contar el número de agujas que intersectan alguna de las rectas paralelas que dibujamos, llamémosle n .
5. Calcular la división $\frac{n}{N}$

Gracias a la ley de los grandes números lo siguiente es cierto.

Theorem 03.1. *Si seguimos las reglas anteriores en el experimento de Laplace entonces la cantidad $\frac{n}{N}$ se acercará al valor de π a medida que N aumente.*

Exercise 03.2. *Pensar en cómo es posible simular las condiciones del experimento de Laplace utilizando por ejemplo Python.*

Máxima verosimilitud

Un acercamiento probabilista a machine learning es vía la máxima verosimilitud, en esta sección exemplificaremos este punto de vista con un ejemplo.

Supongamos que una compañía quiere modelar la satisfacción de un usuario utilizando la siguiente asignación, si el usuario está satisfecho se le asignará el valor uno, de otro modo se le asignará el valor cero.

Notemos que es posible modelar la satisfacción promedio de todos los usuarios mediante un solo valor $\beta^* \in [0, 1]$ (igual a la esperanza de cierta variable aleatoria), sin embargo algunas veces podría ser complicado conocer la información de todos los usuarios y por tanto se considerará solo una muestra de tamaño N , digamos $S = \{x_1, \dots, x_N\}$ de tal forma que esta muestra sea independiente e idénticamente distribuida. En ese caso es posible definir la estimación empírica $\frac{1}{N} \sum_{i \leq N} x_i$. Gracias a la ley de los grandes números $\mathbb{E} \left[\frac{1}{N} \sum_{i \leq N} x_i \right] = \beta^*$.

En la última línea no es claro por qué es necesario calcular la esperanza de la estimación $\frac{1}{N} \sum_{i \leq N} x_i$. Esta última cantidad es enrealidad una variable aleatoria. Ahora vamos a explicar qué significa esta cantidad en términos de verosimilitud. Notemos que si queremos calcular la probabilidad de obtener el muestreo S tenemos la siguiente fórmula:

$$\mathbb{P}(S) = \prod_{i \leq N} (\beta^*)^{x_i} (1 - \beta^*)^{1-x_i} = (\beta^*)^{\sum_{i \leq N} x_i} (1 - \beta^*)^{\sum_{i \leq N} (1-x_i)}$$

$$\text{Eso implica que } \log(\mathbb{P}(S)) = \left(\sum_{i \leq N} x_i \right) \log(\beta^*) + \left(\sum_{i \leq N} (1-x_i) \right) \log(1 - \beta^*)$$

Como \log es una función creciente, definimos el maximizador de la verosimilitud de la siguiente manera:

Definition 04.1. ■ Dado un parámetro $\beta \in [0, 1]$, definimos la cantidad

$L(S, \beta) = \left(\sum_{i \leq N} x_i \right) \log(\beta) + \left(\sum_{i \leq N} (1-x_i) \right) \log(1 - \beta)$ como la verosimilitud logarítmica del parámetro β dada la base de datos S .

■ Dada una base de datos S , definimos el parámetro que maximiza la verosimilitud de la siguiente manera:

$$\beta_{MLE} = \max_{\beta \in [0, 1]} L(S, \beta)$$

Proposition 04.1. *Bajo las hipótesis de esta sección se tiene que $\beta_{MLE} = \frac{1}{N} \sum_{i \leq N} x_i$*

Hasta ahora solo 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 05.1. Si E es la experiencia aleatoria de lanzar un dado justo hasta obtener el número seis, definimos la siguiente variable aleatoria: $X(\omega_1, \omega_2, \dots) =$

$$\inf_{j \geq 1} \{j : \omega_j = 6\}$$

Example 05.2. Sea $\lambda > 0$, definamos $\mathbb{P}_{Poisson, \lambda}(i) = e^{-\lambda} \frac{\lambda^i}{i!}$, ese es un ejemplo de una experiencia aleatoria numerable, llamada ley de Poisson.

Remark 05.3. 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: si $\lim_{n \rightarrow \infty} n \cdot p_n = \lambda$ entonces $\lim_{n \rightarrow \infty} \mathbb{P}_{Bin, p_n, n}(i) = \mathbb{P}_{Poisson, \lambda}(i)$.

Exercise 05.4. Comparemos la distribución de Poisson con la ley binomial que calculamos en el ejemplo de las olimpiadas en las primeras notas, en este caso nuestro parámetro λ será igual a la esperanza de la variable aleatoria de Bernoulli, lo cuál corresponde con 1.535 gracias al cálculo de las primeras notas, de esa forma $\mathbb{P}_{Poisson, 1.535}(0) = 0.215$, $\mathbb{P}_{Poisson, 1.535}(1) = 0.33$, $\mathbb{P}_{Poisson, 1.535}(2) = 0.253$, y $\mathbb{P}_{Poisson, 1.535}(3) = 0.129$.

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 altamente probables, una pregunta inmediata es qué pasa si deseamos calcular eventos cuya probabilidad no se ni muy pequeña ni muy alta, para ello es necesario utilizar el célebre Teorema Límite Central de Lévy.

05.1 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 general solo hablaremos de las llamadas leyes continuas con densidad.

Definition 05.1.

La ley de probabilidad uniforme sobre el conjunto $[-1,1]$ es la ley de probabilidad tal que $\mathbb{P}_{unif}((\epsilon_1, \epsilon_2)) = \frac{1}{|\epsilon_1 - \epsilon_2|}$

Definition 05.2. La ley de probabilidad Gaussiana o normal con parámetros (μ, σ^2) se define para los intervalos $(-\infty, x]$ de la siguiente manera:

$$\mathbb{P}_{Gauss, \mu, \sigma^2}(-\infty, x] = \frac{1}{\sigma \cdot \sqrt{2\pi}} \int_{-\infty}^x \left(\exp \left(-\frac{(t - \mu)^2}{2 \cdot \sigma^2} \right) \right) dt$$

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

Proposition 05.5. Si X es una variable aleatoria tal que $\mathbb{P}_X = \mathbb{P}_{Gauss,\mu,\sigma}$ entonces $\mathbb{E}(X) = \mu$ y $Var(X) = \sigma^2$.

Exercise 05.6. (Distribución gaussiana multi-variada) Sea $\mu \in \mathbb{R}^d$ y $S \in \mathbb{R}^{d \times d}$ tal que $S = S^T$ y $xSx^T > 0$ para cualquier $x \in \mathbb{R}^d \setminus \{\bar{0}\}$. Definimos la ley de probabilidad gaussiana d dimensional con parámetros μ, S de la siguiente forma:

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

B O U R B A K I
COLEGIO DE MATEMÁTICAS

Bibliografía

- [1] L. Badger, *Mathematics Magazine* **1994**, 67, 83-91.



escuela-bourbaki.com