.

Módulo 5: Aprendizaje no supervisado

Métodos de Muestreo, 2da Parte

Diplomatura Cs. de Datos - FaCENA-UNNE

Docentes: Magdalena Lucini, Luis Duarte, Griselda Bóbeda

Generación de distribuciones

Objetivo: Generar realizaciones de alguna variable(vector) aleatorio cuya función de densidad de probabilidad no es fácil de muestrear.
A esta densidad la llamamos densidad objetivo (target density), y la denotamos p_o

Estrategia:
Usar una *densidad propuesta* (proposal density), fácil de muestrear y por medio de alguna transformación, prueba, algoritmo, usar muestras de la densidad propuesta para generar muestras de la densidad objetivo

Diferentes técnicas de muestreo

- Métodos directos ✓
- Métodos de aceptación -rechazo ✓
- Muestreo de importancia y Sampling Importance Resampling ✓
- MCMC: Markov Chain Monte Carlo
  - Técnicas basadas en la construcción de una cadena de Markov que converge a la densidad objetivo.
  - En general son métodos "universales", ya que se pueden aplicar en casi cualquier caso.
  - Las muestras no necesariamente serán independientes

Diferentes técnicas de muestreo

- Métodos directos ✓
- Métodos de aceptación -rechazo ✓
- Muestreo de importancia y Sampling Importance Resampling ✓
- MCMC: Markov Chain Monte Carlo
  - Técnicas basadas en la construcción de una cadena de Markov que converge a la densidad objetivo.
  - En general son métodos "universales", ya que se pueden aplicar en casi cualquier caso.
  - Las muestras no necesariamente serán independientes

MCMC:Markov Chain Monte Carlo

- Los métodos MCMC son una familiad de algoritmos que usan Cadenas de Markov (MC) para realizar estimaciones de Monte Carlo (MC)
- Los métodos de Monte Carlo se usan para generar muestras aleatorias independientes de una distribución y asi aproximar a la cantidad deseada (cálculo de esperanzas de funciones, probabilidades, etc). No siempre se pueden lograr estas muestras, y cuando se logran, no siempre son independientes.
- Se sabe que, bajo ciertas condiciones, las Cadenas de Markov convergen a una distribución estacionaria. Si se hacen simulaciones con esta cadena de Markov para una cantidad suficientemente larga de tiempos se podría, eventualmente, obtener muestras de esta distribución estacionaria.

MCMC:Markov Chain Monte Carlo

- Objetivo: Obtener muestras de una distribución de probabilidades compleja o dificil de muestrear, llamada densidad objetivo o target density, que denotaremos p_o. Se supone que podemos evaluar p_o, pero que no podemos obtener muestras directas de ella.
- Estrategia: Dada la forma funcional de p_o, construir una cadena de Markov que tenga a p_o como distribución estacionaria.
- Se generan muestras de la cadena de Markov de manera tal que la sucesión de muestras $\{x_n\}$ generadas por esta cadena convergen en distribución a la densidad objetivo p_o.

Cadenas de Markov

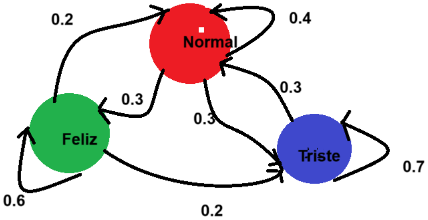
Una Cadena de Markov es un proceso estocástico que evoluciona en el tiempo transicionando en diferentes estados. Esta sucesión de estados se denota por la colección $\{X_i\}$.
La transición entre estados es aleatoria y satisface la propiedad de Markov. Esto es:

$$P(X_t/X_{t-1}, X_{t-2}, \ldots, X_0) = P(X_t/X_{t-1})$$

Esto dice que el proceso no tiene memoria.Así, para determinar la distribución del próximo valor que tome la cadena, sólo necesitamos conocer el estado actual X_t, independientemente de cómo haya sido el camino para llegar a ese estado X_t (independientemente del camino que haya seguido la cadena en el pasado).

La colección de estados que puede tomar una cadena de Markov se denomina Espacio de los estados (state space) y el objeto que gobierna la probabilidad que la cadena se mueva de un estado a otro se llama kernel de transición o matriz de transición

Cadenas de Markov



Representación de estados emocionales

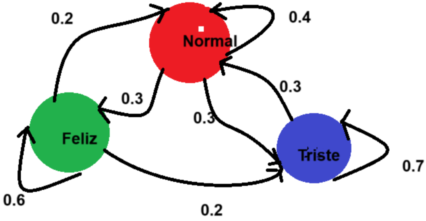
P(Estar Triste hoy/Ayer Feliz) = 0.2

P(Estar Triste hoy/Ayer Feliz, Antes de ayer Normal) = 0.2

- P(Hoy Triste/Ayer Feliz) = $\mathrm{P}(X_t/X_{t-1})$= 0.2
- P(Hoy Triste/Ayer Feliz, Antes de Ayer Feliz) = $\mathrm{P}(X_t/X_{t-1}, X_{t-2})$= 0.2
- P(Hoy Feliz/Ayer triste) = 0

Cadenas de Markov

En el ejemplo hay tres estados, las flechas idican a qué estado la cadena se puede mover desde el estado actual, junto a las probabilidades de transición de un estado al otro



Representación de estados emocionales

1. Feliz, 2. Normal, 3.Triste

Matriz de transición

$$P = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.3 & 0.7 \end{bmatrix}$$

$$P(X_{t+1} = j / X_t = i) = P_{ij}$$

SI la cadena de Markov comienza en el estado 3, con probabilidad 1 $\Rightarrow$ la probabilidad incial sobre los tres estados es $\pi_0 = (0, 0, 1)$

Cadenas de Markov

$$P = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.3 & 0.7 \end{bmatrix}$$

- Supongamos cadena empieza en estado 3, la dist. de probabilidad sobre tres estados es $\pi_0 = (0, 0, 1)$
- La dist de probabilidades sobre los tres estados luego de una iteración es $\pi_1 = \pi_0 P = (0, 0.3, 0.7)$
- Luego de n iteraciones la distribución de probabilidades sobre los tres estados es

$$\pi_n = \pi_0 P^n$$

- Luego de 5 iteraciones, comenzando en el estado 3, con $\pi_o = (0, 01)$, $\pi_5 = (0.2085, 0.3128, 0.4787)$
- Comenzando en el estado 3, $\pi_{100} = (0.23, 0.31, 0.46)$
- Comenzando en el estado 3, $\pi_{101} = (0.23, 0.31, 0.46)$
- Comenzando en el estado 2, con $\pi_o = (0, 1, 0)$, $\pi_5 = (0.2433, 0.30487, 0.45183)$
- Comenzando en el estado 2, con $\pi_o = (0, 1, 0)$, $\pi_{100} = (0.23, 0.31, 0.46)$!
- P(Feliz) $=$ 0.23, P(Normal) $=$ 0.31, P(Triste) $=$0.46

Propiedades de las cadenas de Markov

1. Consideremos una cadena de Markov con un espacio de estados discreto y matriz de transición P. Sea π tal que $\pi = \pi P$. Entonces se dice que tal cadena es estacionaria y π es la distribución estacionaria . Bajo ciertas condiciones (existencia, aperioricidad, irreducibilidad) una cadena de Markov es estacionaria, esto es, $lim_{n\to\infty}\pi_n(i) = \pi(i), \forall i$ en el espacio de estados.
2. Una cadena de Markov es reversible en el tiempo si

$$(X_0, X_1, \ldots, X_n) \stackrel{D}{=} (X_n, X_{n-1}, \ldots, X_0)$$

Esto implica que $(X_0, X_1) \stackrel{D}{=} (X_1, X_0) \Rightarrow X_0 \stackrel{D}{=} X_1$ y por lo tanto $\pi_1 = \pi_0$. Como $\pi_1 = \pi_0 P$, con P matriz de transición, entonces $\pi = \pi_0$ y la cadena es estacionaria.

Cadenas de Markov reversibles en el tiempo

$$
\begin{aligned}
(X_0, X_1) &\stackrel{D}{=} (X_1, X_0) \\
P(X_0 = i, X_1 = j) &= P(X_1 = i, X_0 = j) \\
P(X_0 = i)P(X_1 = j \mid X_0 = i) &= P(X_0 = j)P(X_1 = i \mid X_0 = j)
\end{aligned}
$$

Esta última linea puede escribirse como

$$\pi(i)P(i, j) = \pi(j)P(j, i)$$

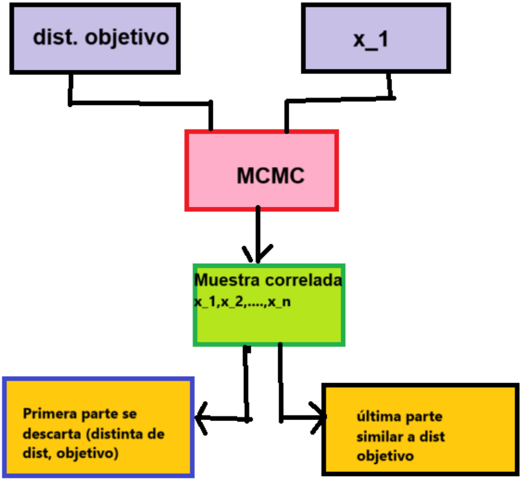
que son las ecuaciones de balance local.

Resumen:

- Se quiere generar muestras de una distribución compleja p_o (densidad objetivo)
- Se sabe que una cadena de Markov aperiódica e irreducible con distribución estacionaria p_o eventualmente converge a esa distribución estacionaria (nuestra distribución objetivo)
- Si una cadena de Markov con matriz de transición P es reversible en el tiempo con respecto a p_o, entonces p_o es la distribución estacionaria de esa cadena de Markov.
- Si una cadena e Markov tiene matriz de transición P, entonces se pueden realizar simulaciones (Monte Carlo) de esta cadena por un período largo de tiempo y eventualmente se estará simulando desde p_o.

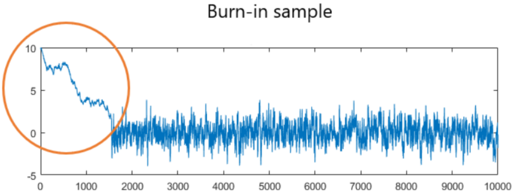
Estos son los principios básicos de un MCMC

MCMC



- MCMC funciona como un método de MonteCarlo, aunque las muestras generadas no son independientes, son realizaciones de una sucesión de v.a. que forman una cadena de Markov '
- A medida que la cantidad de muestras aumenta, ($n \to \infty$), estas muestras se tornan independientes, y la distribución estacionaria.
- La distribución estacionaria es la distribución objetivo
- Bloque MCMC: Se inicia una cadena de Markov con una dist. de probabilidad aleatoria sobre los estados y se mueve gradualmente la cadena hacia la distribución objetivo aplicando alguna condicion (ecuaciones de balance) para asegurar que la dist. estacionaria se parezca a la dist objetivo.

MCMC - Burn in sample



- Debido a las discrepancias entre la distribución objetivo y las distribuciones de los primeros elementos de la cadena, es práctica habitual descartar las primeras realizaciones de una muestra MCMC
- Una sugerencia es descartar el 10% inicial, y quedarse con el resto,
- A este conjunto de muestras descartadas se lo llama Burn in sample
- Al descartar este conjunto se retienen aquellas muestras cuyas distribuciones son más similares a la dist objetivo

MCMC - Correlación y tamaño de muestra efectivo

- Luego de descartar la Burn in sample, se obtiene un conjunto de muestras muy similar a la distribución objetivo. Sin embargo estas muestras NO son independientes.
- Debemos usar el concepto de Tamaño de muestra efectivo: T muestras dependientes son equivalentes a un menor número de muestras independientes (1000 muestras dependientes podrian ser equivalentes a 100 muestras independientes, en este caso el tamaño de muestra efectivo es 100)
- Cuanto mayor sea la correlación entre muestras vecinas, menor será el tamaño de muestra efectivo, y menos precisa la aproximación MCMC.

MCMC: Metropolis-Hastings

- Sea $q(Y \mid X)$: una densidad de transición (densidad propuesta) para X e Y de dimensión p, a partir de la cual se pueda muestrear facilmente.
- $p_o(X)$: es nuestra densidad objetivo, es decir, la distribución estacionaria a la que la cadena de Markov converge.

Supongamos estamos en un estado x.

Algorithm 1 Metropolis Hastings

1.Simular $y \sim q(Y \mid x)$, donde y depende del estado actual x (y vector candidato)

2. Calcular la razón de aceptación $\alpha(y \mid x) = \min \left\{ \dfrac{p_o(y)q(x \mid y)}{p_o(x)q(y \mid x)}, 1 \right\}$

3, Generar $u \sim \mathcal{U}(0,1)$.

if $u \leq \alpha(y \mid x)$ **then**
  aceptar y como próximo estado
else
  permanecer en el estado x
end if

Metropolis Hastings

- Este proceso de 3 pasos representa la matriz de transición de la cadena de Markov a partir de la cual se generan las simulaciones.
- Esta cadena de Markov debería converger a la distribución estacionaria y eventualmente podriamos suponer que las muestras generadas por este proceso son muestras de la distribución estacionaria p_o.

Random Walk Metropolis-Hastings

- La densidad de transición $q(Y \mid X = x)$ se define como $Y = X + \varepsilon$, donde $\varepsilon \sim g$ y g es simétrica respecto a 0.
- Esto implica que $q(Y \mid X = x) = q(X \mid Y = y) = g(\varepsilon)$,
- Como $q(Y \mid X = x)$ es simétrica en x e y, entonces la razón de aceptación de MH es

$$\begin{aligned} \alpha(y \mid x) &= \min\left\{\frac{\pi(y)q(x \mid y)}{\pi(x)q(y \mid x)}, 1\right\} \\ &= \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\} \end{aligned}$$

Algorithm 2 Random Walk Metropolis Hastings

1. Simular $\varepsilon \sim g$ y calcular $y = x + \varepsilon$
2. Calcular la razón de aceptación $\alpha(y \mid x)$

3, Generar $u \sim \mathcal{U}(0, 1)$.

if $u \leq \alpha(y \mid x)$ **then**
    aceptar y como próximo estado
else
    permanecer en el estado x
end if

Ejemplos y ejercicios: Notebook Clase6_ANS.ipynb

Muestreo de Gibbs

- Variante de Metropolis Hastings para generar muestras de distribuciones conjuntas complejas
- Dada una densidad objetivo $p_o(X_1, X_2, \ldots, X_n)$, el algoritmo va generando secuencialmente muestras de las distribuciones condicionales $p_o(X_i/X_{-i})$ de cada variable, manteniendo fijas las demás variables. (X_{-i} denota la n-1 - upla formada por todas las componentes X_j con $j \neq i$
- No hay un paso de aceptación-rechazo, sino que se acepta todo.

Muestreo de Gibbs - Caso bidimensional

- Supongamos es difícil obtener muestras del $p_o(x, y)$, pero que no es dificil obtener muestras de $p_o(x/y)$ y de $p_o(y/x)$.
- Muestreo de Gibbs:
  1. Elegir un estado incial (x_0, y_0)
  2. Muestrear $x_1 \sim p_o(x/y_0)$
     Estado actual:(x_1, y_0)
     Muestrear $y_1 \sim p_o(y/x_1)$
     Estado actual:(x_1, y_1)
  3. Muestrear $x_2 \sim p_o(x/y_1)$
     Estado actual:(x_2, y_1)
     Muestrear $y_2 \sim p_o(y/x_2)$
     Estado actual:(x_2, y_2)
  4. $\vdots$

Repetir iteraciones 1 y 2 M veces. Este proceso define una sucesión de pares de v.a $(X_0, Y_0), (X_1, Y_1), (X_2, Y_2), \ldots$ que forman una cadena de Markov.

Muestro de Gibbs - 3 componentes

Objetivo: obtener muestras de $p_o(x, y, z)$. Las distribuciones condicionales asociadas a esa densidad son: $p_o(x/y, z)$, $p_o(y/x, z)$ y $p_o(z/x, y)$.
Si el estado actual en la n-ésima iteración es (x_n, y_n, z_n), entonces se actualizan:

1. Muestrear $x_{n+1} \sim p_o(x/y_n, z_n)$
   Estado actual:(x_{n+1}, y_n, z_n)
2. Muestrear $y_{n+1} \sim p_o(y/x_{n+1}, z_n)$
   Estado actual:(x_{n+1}, y_{n+1}, z_n)
3. Muestrear $z_{n+1} \sim p_o(z/x_{n+1}, y_{n+1}, z_n)$
   Estado actual:$(x_{n+1}, y_{n+1}, z_{n+1})$

Geman y Geman mostraron que si $p(x_n, y_n, z_n)$ es la densidad en la n-ésima iteración,entonces cuando $n \to \infty$,

$$\begin{aligned} p(x_n, y_n, z_n) &\to p_o(x, y, z) \\ p(x_n) &\to p_o(x) \\ p(y_n) &\to p_o(y) \\ p(z_n) &\to p_o(z) \end{aligned}$$

Muestreo de Gibbs

Ejemplos y ejercicios: Notebook Clase6_ANS.ipynb

Bibliografía

[1] Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006

[2] Peng, Roger. Advanced Statistical Computing. Leanpub, 2022