Curso:
Métodos de Monte Carlo.
Unidad 1, Sesión 2: Conceptos básicos

Departamento de Investigación Operativa
Instituto de Computación, Facultad de Ingeniería
Universidad de la República, Montevideo, Uruguay

dictado semestre 1 - 2010

Contenido:

1. Repaso notación elementos básicos de probabilidad.
2. Motivación del Método de Monte Carlo.
3. Ejemplos.
4. Ejercicios.
5. Lectura adicional.

Notación y repaso de elementos básicos de probabilidad

- X variable aleatoria (discreta o continua).
- F_X distribución de probabilidad de X, $F_X(x) = \text{Prob}\,(X \leq x)$.
- Si X es continua, f_X función de densidad de probabilidad de X (tal que $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$).
- $\text{E}\,(X)$ esperanza de X; muchas veces denotamos $\phi = \text{E}\,(X)$ el valor que se desea calcular a través del muestreo de X.
  - Si X es continua, $\text{E}\,(X) = \int_{-\infty}^{\infty} t f_X(t)dt$.
  - Si X es discreta y toma valores en un conjunto C, $\text{E}\,(X) = \sum_{x \in C} x\text{Prob}\,(X = x)$.
- $\text{Var}\,(X) = \text{E}\left((X - \text{E}\,(X))^2\right) = \text{E}\left(X^2\right) - \text{E}\,(X)^2$ varianza de X; muchas veces denotada σ_X^2.

- $\mathrm{DE}(X) = \sqrt{\mathrm{Var}\,(X)}$ desviación estándar de X, muchas veces denotada σ_X.

- $\mathrm{CV}(X) = \mathrm{DE}(X)/\mathrm{E}\,(X)$ coeficiente de variación de X, es una medida de la desviación o dispersión de una distribución de probabilidad, normalizada teniendo en cuenta el valor esperado.

- $\mathbf{X} = (X_1, X_2, \cdots, X_m)$ vector aleatorio de dimensión m (compuesto por m variables aleatorias distintas, dependientes o independientes).

- $X^{(1)}, X^{(2)}, \cdots, X^{(n)}$ - muestra de n variables aleatorias independientes con la misma distribución de X.

- Notar la diferencia entre un vector aleatorio de variables distintas, y una muestra de n variables equidistribuidas.

Distribuciones básicas

Se recuerda las siguientes distribuciones, que serán empleadas en la discusión subsiguiente y en algunos ejemplos y ejercicios.

- Distribución uniforme entre a y b, $U(a,b)$:
  - p.d.f $f_U(x) = 0$ si $x < a$ o $x > b$; $f_U(x) = 1/(b-a)$ si $a \leq x \leq b$;
  - $F_U(x) = 0$ si $x < a$; $F_U(x) = (x-a)/(b-a)$ si $a \leq x \leq b$, $F_U(x) = 1$ si $x > b$

- Distribución exponencial de parámetro $\lambda > 0$, $E(\lambda)$:
  - p.d.f $f_E(x) = 0$ si $x < 0$; $f_E(x) = \lambda e^{-\lambda x}$ si $x \geq 0$;
  - $F_E(x) = 0$ si $x < 0$; $F_E(x) = 1 - e^{-\lambda x}$ si $x \geq 0$.

- Distribución normal de parámetros μ y$\sigma > 0$, $N(\mu, \sigma)$:
  - p.d.f $f_N(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Esquema básico de un Método Monte Carlo

Supongamos que deseo calcular un cierto valor ϕ, y conozco una variable aleatoria X con distribución F_X tal que $\phi = \mathsf{E}(X)$.

El método de Monte Carlo en su versión más simple consiste en

1. *sortear* valores para un conjunto $X^{(1)}, X^{(2)}, \cdots, X^{(n)}$, de variables aleatorias i.i.d. (independientes e identicamente distribuidas) a X.

2. Calcular $S_n = X^{(1)} + \ldots + X^{(n)}$, la suma de los n valores sorteados.

3. Calcular $\widehat{X} = S_n/n$.

4. Calcular $\widehat{V} = \sum_{i=1}^{n} (X^{(i)})^2/(n(n-1)) - \widehat{X}^2/(n-1)$.

Se dice que $\widehat{X}$ es un *estimador* de ϕ; discutiremos en las próximas transparencias los argumentos que llevan a pensar que con alta probabilidad sus valores son cercanos.

Comentarios

- Por sortear entendemos generar aleatoriamente, siguiendo la distribución de probabilidad F_X.

- El conjunto de valores sorteados para $X^{(1)}, X^{(2)}, \cdots, X^{(n)}$ se llama *muestra*, o equivalentemente *conjunto de replicaciones de* X.

- n es el *tamaño de la muestra*, también llamado *número de replicaciones*.

- $\widehat{X}$ es en sí misma una variable aleatoria, de esperanza igual a ϕ. Resulta interesante estimar (si existe) la varianza de $\widehat{X}$. Formalmente, si la varianza de X existe (y la denotamos $\sigma_X^2 = \mathrm{Var}\,(X)$), aplicando las hipótesis de equidistribución e independencia de las observaciones $X^{(i)}$,

sabemos que

$$\mathsf{Var}\left(\widehat{X}\right) = \mathsf{Var}\left(S_n/n\right) = \mathsf{Var}\left(\frac{\sum_{i=1}^{n} X^{(i)}}{n}\right) =$$

$$= \sum_{i=1}^{n} \frac{\mathsf{Var}\left(X^{(i)}\right)}{n^2} = \frac{n\mathsf{Var}\left(X\right)}{n^2} = \frac{\mathsf{Var}\left(X\right)}{n}.$$

Sin embargo, en general $\mathsf{Var}\left(X\right)$ no se conoce. Como alternativa, podemos emplear la propia muestra para obtener un estimador de $\mathsf{Var}\left(X\right)$, el estimador insesgado más habitual es

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X^{(i)} - \widehat{X})^2.$$

Realizando manipulaciones llegamos al siguiente estimador de la

varianza de $\widehat{X}$,

$$\widehat{V} = \sum_{i=1}^{n} (X^{(i)})^2/(n(n-1)) - \widehat{X}^2/(n-1).$$

Motivación del método

Es claro que el método de Monte Carlo no provee el valor exacto deseado, sino una aproximación, con un cierto error; este tema será discutido más a fondo en sesiones siguientes.

La justificación inicial del uso de Monte Carlo proviene de dos teoremas centrales de la probabilidad y la estadística, la Ley Débil de los Grandes Números y el Teorema del Límite Central (o Teorema Central del Límite).

Sea $X_1, X_2, \ldots$ un conjunto de variables aleatorias i.i.d. (independientes e identicamente distribuidas).

Sea $S_n = X_1 + \ldots + X_n$. Si existe la esperanza $\mu = \mathrm{E}\,(X_i)$, entonces la Ley Débil de los Grandes Números indica que, para todo $\epsilon > 0$,

$$\lim_{n \to \infty} \mathrm{Prob}\left(\left| \frac{S_n}{n} - \mu \right| > \epsilon \right) = 0.$$

La interpretación es que si se suma n muestras independientes de X_i, la probabilidad que la suma (normalizada por n) esté lejos del valor exacto a estimar μ tiene a 0 con n.

Si, adicionalmente, existe la varianza $\sigma^2 = \mathrm{E}\left((X_i - \mu)^2\right)$, el Teorema del Límite Central implica que

$$\lim_{n\to\infty} \mathrm{Prob}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} < a\right) = (2\pi)^{-1/2} \int_{-\infty}^{a} e^{-x^2/2} dx.$$

Dado que el término de la derecha es la distribución de probabilidad de una variable aleatoria normal de media 0 y varianza 1, este teorema indica cuál es el comportamiento asintótico de la distribución del error cometido al emplear S_n como estimador de μ.

Ambos resultados proveen la motivación para aplicar Monte Carlo, ya que indican que con un número suficientemente alto de experimentos, es posible estimar el parámetro deseado incurriendo en pequeño error con alta

probabilidad, y permiten cuantificar asintóticamente la relación entre estos dos valores (error y probabilidad) a través de la distribución normal.

Sin embargo, es preciso ser cauteloso en la aplicación práctica del método, ya que las implementaciones reales no verifican las hipótesis de estos dos teoremas. Por un lado, las limitaciones computacionales imponen una cota superior a los valores de n que se pueden emplear (y cuando se emplean números seudo-aleatorios, la naturaleza cíclica de estos hace que no sea posible obtener una cantidad arbitraria de muestras independientes). Por otra parte, las estimaciones de errores a partir del Teorema del Límite Central sólo son válidas asintóticamente, pero su calidad para un valor de n dado depende de la velocidad de convergencia de la distribución de $S_n - n\mu$ a la distribución normal, lo que introduce una nueva fuente de error.

Es posible usar como alternativa otras fórmulas para derivar información sobre el error cometido por el método, tal como se discute en próximas sesiones.

Ejemplo 1

Supongamos que tenemos un satélite, que para su funcionamiento depende de que al menos 2 paneles solares de los 5 que tiene disponibles estén en funcionamiento, y queremos calcular ϕ la vida útil esperada del satélite (el tiempo promedio de funcionamiento hasta que falla, usualmente conocido en la literatura como MTTF - Mean Time To Failure).

Supongamos que cada panel solar tiene una vida útil que es aleatoria, y está uniformemente distribuída en el rango [1000 hs, 5000 hs] (valor promedio: 3000 hs).

Para estimar por Monte Carlo el valor de ϕ, haremos n experimentos, cada uno de los cuales consistirá en sortear el tiempo de falla de cada uno de los paneles solares del satélite, y observar cual es el momento en el cuál han fallado 4 de los mismos, esta es la variable aleatoria cuya esperanza es el tiempo promedio de funcionamiento del satélite.

El valor promedio de las n observaciones nos proporciona una estimación de ϕ.

Exper. nro.	Tiempo hasta falla de Panel 1	Panel 2	Panel 3	Panel 4	Panel 5	satélite, $X^{(i)}$
1	3027	1738	2376	4685	4546	4546
2	4162	4029	4615	3455	3372	4162
3	3655	2896	1378	4010	4144	4010
4	2573	2649	2117	3956	1281	2649
5	2977	2724	1355	2268	3262	2977
6	3756	4190	1749	3398	2581	3756
Prom.	-	-	-	-	-	$S_n/n = 3683$

Table 1: Una simulación detallada con $n = 6$ experimentos.

De esta simulación, tenemos un valor estimado para la vida útil esperada del satélite de 3683. Un indicador del error que podemos estar cometiendo es la varianza o equivalentemente la desviación estándar de S_n, que en este caso es (haciendo los cálculos) 297.

Seudocódigo básico de un Método Monte Carlo

Supongamos que deseo calcular un cierto valor ϕ, y conozco una variable aleatoria X con distribución F_X tal que $\phi = \mathsf{E}(X)$.

Procedimiento EstimaciónMonteCarlo (integer n, real $\widehat{X}$), real $\widehat{V}$
Parámetro de entrada: n, $tamaño\ de\ la\ muestra$
Parámetros de salida: $\widehat{X}$, $estimador$ de ϕ; $\widehat{V}$, $estimador$ de Var $\left(\widehat{X}\right)$

1. $\widehat{X} = 0.$ /* Inicialización */
2. $\widehat{V} = 0.$
3. For $i = 1, \ldots, n$ do
   - 3.1 Sortear un valor de la variable $X^{(i)}$ con distribución F_X
   - 3.2 $\widehat{X} = \widehat{X} + X^{(i)}$ /* Acumular*/
   - 3.3 $\widehat{V} = \widehat{V} + (X^{(i)})^2$ /* Acumular*/
4. $\widehat{X} = \widehat{X}/n$
5. $\widehat{V} = \widehat{V}/(n * (n-1)) - \widehat{X}^2/(n-1)$

Para la implementación computacional de Monte Carlo, se supone siempre posible el conseguir muestras de variables aleatorias uniformes entre 0 y 1 ($U(0,1)$), y el generar muestras de otras distribuciones a partir de la transformación de las variables uniformes. En la unidad 4 se discutirá más a fondo este tema, esencial en la práctica.

Para dar los elementos necesarios para poder programar implementaciones, se adelantan los siguientes conceptos:

- Bibliotecas para generar números seudo-aleatorios: conjunto de funciones que permiten generar secuencias de números que se comportan de forma razonablemente similar a una secuencia de variables aleatorias independientes con distribución uniforme entre 0 y 1.
  - Semilla: valor dado para inicializar la secuencia, semillas distintas resultan en secuencias distintas.
  - Función de inicialización: inicializa la secuencia con una semilla.
  - Función de sorteo: proporciona el próximo número aleatorio dentro de la secuencia.

- Generación de una v.a. $X = U(a, b)$ a partir de una v.a. $U = U(0, 1)$: se sortea el valor de U, y se calcula $X = a + (b - a)U$.

- Generación de una v.a. $X = E(\lambda)$ a partir de una v.a. $U = U(0, 1)$: se sortea el valor de U, y se calcula $X = -ln(U)/\lambda$.

Ejemplo 2 - implementación

El siguiente ejemplo muestra una implementación en lenguaje C de Monte Carlo aplicado en un caso muy sencillo.

Problema: se desea calcular la esperanza de la vida útil de un satélite, cuyo equipamiento principal (sujeto a fallos) consiste en dos computadoras y un equipamiento de transmisión. El satélite funciona mientras al menos una de las dos computadoras y el equipo de transmisión funcionen. En el momento del lanzamiento, las computadoras poseen una vida útil aleatoria de distribución uniforme entre 0 y 500 hs; y el equipamiento de transmisión posee una vida útil aleatoria de distribución uniforme entre 0 y 1500 hs.

Solución: Definir un experimento consistente en sortear la vida útil (el tiempo hasta falla) de las dos computadoras y el equipo de transmisión, y en base a estos calcular el tiempo hasta falla del satélite. Para sortear estos valores, se emplea una rutina que genera números seudoaleatorios.

Repetir este experimento un número fijo de veces (por ejemplo 1000), y

calcular la estimación de la esperanza de vida útil y de la desviación de esta estimación utilizando las fórmulas vistas previamente.

El código correspondiente (utilizando la función estándar drand48 para generación de números aleatorios) está disponible en
`http://www.fing.edu.uy/inco/cursos/mmc/codigoC/satelite.c`

También está disponible el código (utilizando el generador "Mersenne Twister", disponible en
`http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html`):
`http://www.fing.edu.uy/inco/cursos/mmc/codigoC/satelite-version2.c`

Ambos códigos fueron compilados con `gcc`, usando la opción `-lm` (biblioteca matemática).

Preguntas para resumen de lectura (ayuda para el estudio y la autoevaluación de lo aprendido.)

- ¿Cuál es el esquema básico del Método de Monte Carlo?
- ¿Qué es el tamaño de la muestra del método?
- ¿Qué es el estimador calculado por el método?
- ¿Qué teoremas son los que motivan el desarrollo del método? ¿Es fácil verificar que las hipótesis de estos teoremas se cumplen en la práctica?

Ejercicio (parte de la evaluación del curso, debe realizarse en grupos y entregarse por mail al docente del curso)

Supongamos que para construir una casa debemos efectuar la siguiente lista de tareas:

- T1 - cimientos - tiempo aleatorio uniforme entre 32 y 48 hs.
- T2 - paredes - tiempo aleatorio uniforme entre 40 y 60 hs.
- T3 - techo - tiempo aleatorio uniforme entre 15 y 25 hs.
- T4 - instalación sanitaria - tiempo aleatorio uniforme entre 10 y 15 hs.
- T5 - instalación eléctrica - tiempo aleatorio uniforme entre 10 y 15 hs.
- T6 - cerramientos - tiempo aleatorio uniforme entre 6 y 10 hs.

- T7 - pintura - tiempo aleatorio uniforme entre 18 y 24 hs.
- T8 - limpieza final - tiempo aleatorio uniforme entre 4 y 8 hs.

Hay ciertas dependencias que implican que una tarea no puede comenzar hasta haberse terminado otra previa:

- T2, T3 dependen de 1.
- T3 depende de 2
- T4 depende de 2
- T5 depende de 2 y 3
- T6 depende de 2 y 3
- T7 depende de 2, 4 y 5

- T8 depende de 4, 5, 6 y 7

Ejercicio 2.1:

1. implementar un programa que reciba como parámetros de línea de comando (o pregunte en pantalla) la cantidad de replicaciones n a realizar, y emplee Monte Carlo para calcular (e imprimir) la estimación del tiempo total desde que se comienza la obra hasta que se finaliza la misma, y la desviación estándar de este estimador.

2. Incluir código para calcular el tiempo de cálculo empleado por el programa.

3. Utilizar el programa con $n = 10, 100, 1000, 10000$, y mostrar en una tabla las estimaciones de media y desviación estándar, así como los tiempos de cálculo. Discutir estos resultados.

Las pautas sobre el informe a entregar estan disponibles en `http://www.fing.edu.uy/inco/cursos/mmc/pautas.htm`.

Fecha entrega: Ver calendario de entregas en página web del curso.

Material adicional de repaso/referencia

- Material
  de repaso sobre probabilidad y estadística para Monte Carlo, Capítulo 2 de
  http://www.ipp.mpg.de/de/for/bereiche/stellarator/Comp_sci/CompScience/csep/csep1.phy.ornl.gov/mc/mc.html
  (especialmente interesantes secciones 2.1, 2.2, 2.3, 2.4.6).

- Otro repaso de principios de probabilidad y estadística para Monte Carlo:
  http://www.ualberta.ca/~cdeutsch/images/Lec02-ProbDist.pdf.

- Discusión de las propiedades de la varianza (incluyendo la derivación del estimador insesgado de la varianza de una muestra) en
  http://en.wikipedia.org/wiki/Variance.