Thompson Sampling

Alvaro J. Riascos Villegas
Universidad de los Andes y Quantil

Septiembre de 2024

Contenido

1. Introducción
2. Análisis Bayesiano en una Cáscara de Nuez
3. Thompson Sampling
   - Bandido Bernoulli
   - TS General

Introducción

- Thompson sampling es una forma general de incorporar incertidumbre sobre los parámetros que definen las variables de interés en un problmea de bandidos multibrazos.
- Esta inspirado en el análisis Bayesiano que ofrece una interpretación alternativa del concepto de probabilidad como una cuantificacion de la incertidumbre.
- Desde un punto de vista operativo permite incorporra estas características del análisis Bayesiano:
  - Usualmente existe información inicial sobre los parámetros de un modelo estructural.
  - Permite condicionar a los datos observados. En el análisis clásico se promedia sobre los datos, aun los no observados.

Contenido

1. Introducción
2. Análisis Bayesiano en una Cáscara de Nuez
3. Thompson Sampling
   - Bandido Bernoulli
   - TS General

Teorema de Bayes

- La definición de probabildad condicional es:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{1}$$

- El teorema de Bayes establece que: $P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$.
- Si θ es una variable aleatoria que parametriza una **distribución muestral** de una variable observada y, $p(y \mid \theta)$ entonces:

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} \tag{2}$$

donde $p(\theta)$ es la **prior**, una medida de la incertidumbre de θ y $p(\theta \mid y)$ se llama la **posterior**.

Estadística Bayesiana

- Toda la estadistica Bayesiana esta basada en el teorema de Bayes y el cálculo de la posterior.
- La interpretacion es la siguiente: Observamos muestras de y de una distribución muestral de $p(y \mid \theta)$,
- θ no se observa y queremos aprender de este parámetro.
- Nuestra incertidumbre sobre θ la modelamos con la prior.
- Una vez observamos los datos y actualizamos la prior a la distribución posterior: esta resume el conocimiento que se tiene de los parámetros dado los datos observados.

Ejemplo: Distribución normal

Example (Distribución inicial y muestral normal)

Supongamos que tenemos una muestra de n observaciones $y_1, ..., y_n$, $y_i \backsim_{i.i.d} N(\mu, 1)$ entonces la distribución muestral es:

$$p(y \,|\mu) = (2\pi)^{-\frac{n}{2}} \sigma^{-n} exp(-\frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2) \tag{3}$$

Ahora supongamos que la distribución inicial $p(\mu) \backsim N\left(\mu_0, \sigma_0^2\right)$ donde los parámetros de esta distribución son conocidos (estos se denominan hiperparámetros).

Ejemplo: Distribución normal

Example (continuación)

La distribución expost es:

$$p(\mu \,|y) \propto p(y \,|\mu) \, p(\mu) \tag{4}$$

$$\propto exp(-\frac{1}{2\overline{\sigma}^2} \sum (\mu - \overline{\mu})^2) \tag{5}$$

$$\overline{\mu} = \frac{\frac{n}{\sigma^2}\overline{y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} \tag{6}$$

$$\overline{\sigma}^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} \tag{7}$$

Obsérvese que la prior y la posterior son de la misma familia de distribuciones. Se llaman distribuciones conjugadas.

Ejemplo: Bernoulli

- Supongamos que tenemos tres armas R_i que se distribuyen Bernoulli, $Bern(\theta_i)$, donde θ_i es desconocido y fijo en el tiempo.
- Cuando se dispara una arma se recibe una recompensa de 1 de lo contrario cero.
- Obsérvese que $E[R_i] = \theta_i$.
- Sigiendo un aproximación Bayesiana, supongamos que θ_i es una variable aleatoria. Esto es una estrategia que usaremos para resolver el problema, no queremos decir que en realidad θ_i sea una variable aleatoria.
- El objetivo es maximizar la suma de las recompensas hasta la ronda T.

Ejemplo: Bernoulli

- Supongamos que después de interactuar con el ambiente, elegir las acciones 1 y 2, 1000 veces y la acción 3, 3 veces, se tiene el siguiente conocimiento sobre la recompensa promedio de cada acción: $E[R_i] = \theta_i$ (i.e., la posterior de θ_i):

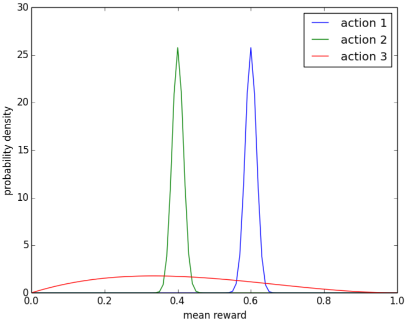


Figura: Densidad de probabilidad sobre recompensas promedio después de elegir las acciones 1 y 2, 1000 veces y la acción 3, 3 veces. Con 600, 400 y 1 ganancia acumulada respectivamente.

Ejemplo: Bernoulli

- En promedio las acción 2 es más alta pero la acción 3 tiene mucha incertidumbre.
- Esta incertidumbre se puede deber a que se ha disparado poco y un algoritmo que no tenga en consideración esto podría no elegir, eventualmente, la mejor acción.
- Un algoritmo ϵ-codicioso explora com la misma probabilidad cada una de la acciones. Esto puede ser ineficiente porque la acción 2 parece estar dominada por la acción 1 mientras que la acción 3 es promisoria.
- Thompson Sampling es una forma de atacar ese problema.

Ejemplo: Caminos más cortos en un grafo

- Una persona desea ir del punto 1 al 12 y los tiempos de desplazamiento son en **promedio** θ_e, donde e es un enlace entre nodos.

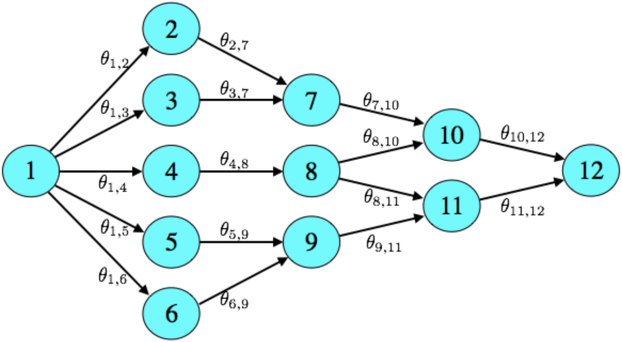


Figura: Camino más corto

- Las acciones son caminos en el grafo entre 1 y 12. Un camino a_t es una sucesion de enlaces $a = (e_1, ..., e_k)$.
- El objetivo es minimizar el valor esperado del tiempo de recorrido: $\sum_{e \in a} \theta_e$

Ejemplo: Caminos más cortos en un grafo

- Las acciones son caminos en el grafo entre 1 y 12. Un camino a_t es una sucesion de enlaces $a = (e_1, ..., e_k)$.
- El objetivo es minimizar el valor esperado del tiempo de recorrido: $\sum_{e \in a} \theta_e$
- Obsérvese que θ_e son los parámetros de interes. La estrategia que usaremos es suponer que son el valor esperado de una variable aleatoria.

Contenido

1. Introducción
2. Análisis Bayesiano en una Cáscara de Nuez
3. Thompson Sampling
    - Bandido Bernoulli
    - TS General

Ejemplo: Bernoulli

- Supongamos que modelamos θ_k como una distribución Beta con parámetros α_k, β_k ($Beta(\alpha_k, \beta_k)$):

$$p(\theta_k) = \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)} \theta_k^{\alpha_k} (1 - \theta_k)^{\beta_k - 1}$$

- Esta distribución juega un papel instrumental en nuestro objetivo. La actualización de esta distribución, usando el Teorema de Bayes, cuantifica la incertidumbre que se tiene de los parámetros de interés.

Ejemplo: Bernoulli

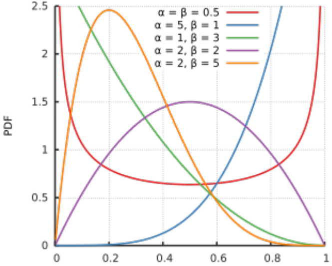


Figura: Beta distribution. By Horas based on the work of Krishnavedala - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=15404515

Ejemplo: Bernoulli

- Por el teorema de Bayes:

$$f(\theta_i \mid R_i) = \frac{f(R_i \mid \theta_i) f(\theta_i)}{f(R_i)}$$

- Si R_i condicional a θ_i se distribuye $Bern(\theta_i)$ y θ_i se distribuye $Beta(\alpha_i, \beta_i)$ entonces:

$$f(\theta_i \mid R_i) \text{ se distribuye } Beta(\alpha_i', \beta_i')$$

  donde: $(\alpha_i', \beta_i') \leftarrow (\alpha_i, \beta_i) + (R_i, 1 - R_i)$ si $a = i$, caso contrario los parámetros permanencen invariantes (a es la acción que el agente toma antes de actualizar la distribución de θ_i).

- Observaciones:
  - El parámetro solo se actualiza si se dispara el arma.
  - La ventaja de cuantificar la incertidumbre con la distribucion Beta es que la posterior sigue siendo Beta. Decimos que la distribución de Bernoulli y Beta son distribuciones conjugadas.

Ejemplo: Bernoulli

- Una distribución *Beta*(α_i, β_i) tiene media $\frac{\alpha_i}{\alpha_i+\beta_i}$
- La figura corresponde a:
  $(\alpha_1, \beta_1) = (601, 401), (\alpha_2, \beta_2) = (401, 601), (\alpha_3, \beta_3) = (2, 3)$

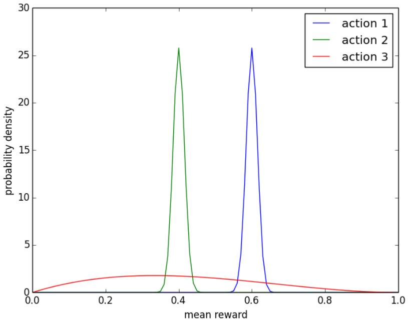


Figura: Densidad de probabilidad sobre recompensas promedio después de elegir las acciones 1 y 2, 1000 veces y la acción 3, 3 veces. Con 600, 400 y 1 ganancia acumulada respectivamente.

Ejemplo: Bernoulli

- Una estrategia greedy (codiciosa) para descubrir la política óptima es:

Algorithm 1 BernGreedy(K, α, β)

for $t = 1, 2, \ldots$ **do**
  #estimate model:
  for $k = 1, \ldots, K$ **do**
    $\hat{\theta}_k \leftarrow \alpha_k / (\alpha_k + \beta_k)$
  end for

  #select and apply action:
  $x_t \leftarrow \text{argmax}_k\, \hat{\theta}_k$
  Apply x_t and observe r_t

  #update distribution:
  $(\alpha_{x_t}, \beta_{x_t}) \leftarrow (\alpha_{x_t} + r_t, \beta_{x_t} + 1 - r_t)$
end for

Algorithm 2 BernTS(K, α, β)

for $t = 1, 2, \ldots$ **do**
  #sample model:
  for $k = 1, \ldots, K$ **do**
    Sample $\hat{\theta}_k \sim \text{beta}(\alpha_k, \beta_k)$
  end for

  #select and apply action:
  $x_t \leftarrow \text{argmax}_k\, \hat{\theta}_k$
  Apply x_t and observe r_t

  #update distribution:
  $(\alpha_{x_t}, \beta_{x_t}) \leftarrow (\alpha_{x_t} + r_t, \beta_{x_t} + 1 - r_t)$
end for

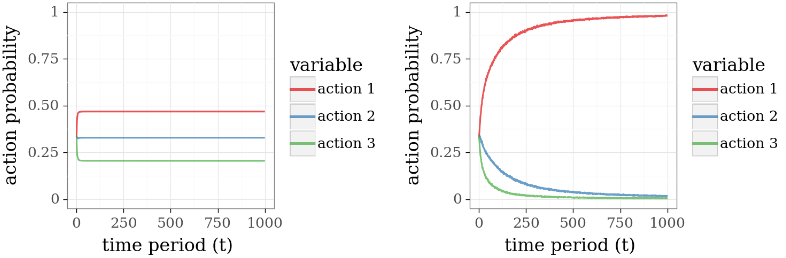
Figura: Bernoulli Codicioso y Bernoulli TS

Ejemplo: Bernoulli

- ¿Por que TS funcionaria?
- La estrategia codiciosa no eligiría la acción 3.
- La estrategia Bernoulli Greddy tampoco eligiría la acción 3.
- Las dos estrategias anteriores con exploración (ϵ codiciosa) le asignariá la misma probabilidad a las tres acciones.
- TS elige las estrategias $1, 2, 3$ con probabilidad $0{,}82, 0, 0{,}18$. Es decir, explora con una probabilidad alta aquellas acciones sobre las que se tiene más incertidumbre.

Ejemplo: Bernoulli

- La siguiente gráfica muestra el desempeño del modelo para los parámetros: $E[\theta_1] = 0{,}9$, $E[\theta_2] = 0{,}8$, $E[\theta_1] = 0{,}7$.



(a) greedy algorithm　　**(b)** Thompson sampling

Figure 3.1: Probability that the greedy algorithm and Thompson sampling selects an action.

Figura: 10,000 mil simulaciones de cada algortimo. Cada simulacion de 1,000 rondas. Cada punto representa la fracción de veces en una ronda específica que el algoritmo seleccionó una acción.

Modelo

- Supongamos que un agente toma una sucesión de acciones $x_1, x_2, \ldots,$ donde cada $x_i \in \Xi$.
- Depués de la acción i el agente observa un resultado y_i, $y_i \backsim q_{\theta_i}(. \mid x_t)$.
- θ es desconocido pero el agente cuantifica su incertidumbre usando una prior $p(\theta)$.
- El agente recibe una recompensa $r_t = r(y_t)$.
- El objetivo del agente es maximizar el valor esperado de la recompensa: $v_{x_t}(\theta) = E_{q_\theta(.|x_t)}[r_t]$
- El algoritmo general es:

Ejemplo: Bernoulli

Algorithm 3 Greedy($\mathcal{X}, p, q, r$)

for $t = 1, 2, \dots$ **do**
    #estimate model:
    $\hat{\theta} \leftarrow \mathbb{E}_p[\theta]$

    #select and apply action:
    $x_t \leftarrow \mathrm{argmax}_{x \in \mathcal{X}} \mathbb{E}_{q_{\hat{\theta}}}[r(y_t)|x_t = x]$
    Apply x_t and observe y_t

    #update distribution:
    $p \leftarrow \mathbb{P}_{p,q}(\theta \in \cdot | x_t, y_t)$
end for

Algorithm 4 Thompson($\mathcal{X}, p, q, r$)

for $t = 1, 2, \dots$ **do**
    #sample model:
    Sample $\hat{\theta} \sim p$

    #select and apply action:
    $x_t \leftarrow \mathrm{argmax}_{x \in \mathcal{X}} \mathbb{E}_{q_{\hat{\theta}}}[r(y_t)|x_t = x]$
    Apply x_t and observe y_t

    #update distribution:
    $p \leftarrow \mathbb{P}_{p,q}(\theta \in \cdot | x_t, y_t)$
end for

Figura: TS General

Ejemplo: Bernoulli

Example

- $\Xi = \{1, 2, ...K\}$.
- $y_t = r_t$.
- $q_\theta(1 \mid k) = \theta_k$
- $p(\theta)$ es Beta.

Forma Equivalente de TS

- La siguiente es una forma equivalente del algortimo.
- Sea:

$$w_{xt} = \int I(x = argmax_{x'} v_{x'}(\theta)) p(\theta \mid y_t) d\theta$$

  Es decir $w_{xt} = p(\theta_k \mid y_k)$ si $x = argmax_{x'} v_{x'}(\theta)$ y cero caso contrario..
- Entonces TS se puede implementar de la siguinete forma:
  1. Para cada x estimar w_{xt} por ejemplo usando Montecarlo: muestrear θ de $p(\theta)$ y calcular el promedio.
  2. Para $t+1$ elegir x con probabilidad w_{xt}.
- Esto es equivalente al algoritmo TS introducido anteriormente.
- Con la nueva formulación se tiene una interpretación: si x es óptimo con una probabilidad de w_{xt} entonces con esa probabilidad se elige en la siguiente ronda.