

1

Probabilidade condicional

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

$$P(B) = \int P(A, B) dA$$

$$P(A \uparrow | B \downarrow) = \frac{0,25}{0,50} = 0,5$$

Produto escalar entre duas V.H.

$$\langle X, Y \rangle = E(XY)$$

sendo $E(x) = \int x P(x) dx$

Definição de variável condicional via produto escalar

$$E(X - E(X|Y), Y) = 0$$

$$E\{(X - E(X|Y))Y\} = 0$$

$$E\{XY\} - E\{XY E(X|Y)\} = 0$$

$$\iint XY P(x, y) dx dy - \iint XY E(X|Y) P(x, y) dx dy = 0$$

Exemplo: duas moedas

$$A = \left\{ \begin{array}{c} \uparrow \\ 0,5 \end{array}, \begin{array}{c} \downarrow \\ 0,5 \end{array} \right\}$$

$$B = \left\{ \begin{array}{c} \uparrow \\ 0,5 \end{array}, \begin{array}{c} \downarrow \\ 0,5 \end{array} \right\}$$

$$P(A, B) = \left\{ \begin{array}{c} \begin{array}{c} A \uparrow \\ \uparrow \\ B \uparrow \end{array}, \begin{array}{c} A \uparrow \\ \uparrow \\ B \downarrow \end{array}, \begin{array}{c} A \downarrow \\ \uparrow \\ B \uparrow \end{array}, \begin{array}{c} A \downarrow \\ \uparrow \\ B \downarrow \end{array} \\ \begin{array}{c} 0,25 \\ 0,25 \\ 0,25 \\ 0,25 \end{array} \end{array} \right\}$$

(2)

Exemplo 2.5 $X \sim \text{Uniforme}[0, 1]$ Encontrar $E(\varepsilon | n)$ sendo

$$\varepsilon = 2x^2$$

e

$$n(x) = \begin{cases} 1 & 0 \leq x \leq 1/3 \\ 2 & 1/3 < x \leq 2/3 \\ 0 & 2/3 < x \leq 1 \end{cases}$$

Seja $h(n) = E(\varepsilon | n)$. Note que $h(n)$ deverá ter apenas 3 valores já que $n = \{1, 2, 0\}$

Resolvendo pela definição:

$$\langle \varepsilon - E(\varepsilon | n) | n \rangle = 0$$

$$\langle \varepsilon - h(n) | n \rangle = 0$$

$$E[(\varepsilon - h(n)) | n] = 0$$

Tomando cada um dos valores de n , temos:

$$n=1$$

$$E((\varepsilon - h(1)) | 1) = 0$$

$$\int (2x^2 - h(1)) P(x) dx = 0$$

$$\int_0^{1/3} (2x^2 - h(1)) dx = 0$$

(3)

$$\int_0^{1/3} 2x^2 dx - h(1) \int_0^{1/3} dx = 0$$

$$2 \frac{x^3}{3} \Big|_0^{1/3} - \frac{h(1)}{3} = 0$$

$$h(1) = 2/3^3 = 2/27$$

Similarmente, teríamos
 $h(2) = 14/27$
 $h(0) = 38/27$

Resolvendo como um problema de minimização

Sabemos que

$$\langle x - E(x|Y), Y \rangle = 0$$

* que $E(x|Y) = h_{opt}(Y)$, no sentido que
 $h_{opt}(Y)$ é o minimum mean squared error MMSE

ou seja

$$h_{opt}(Y) = \min_h \int_{\mathbb{R}^3} (x - h(y))^2 P(x|y) dx$$

Como n assume três valores, vamos tomar

$$h(n) = a + bn + cn^2$$

De modo que a integral fica

$$h_{opt}(n) = \min_h \int (2x^2 - h(n(x)))^2 P(x|n) dx$$

$$= \min_h \int_0^1 (2x^2 - h(n(x)))^2 dx$$

$$h_{\text{opt}}(n) = \min_h f(a, b, c) \quad (4)$$

Impõendo as condições para máximo:

$$\frac{\partial f}{\partial a} = 0, \quad \frac{\partial f}{\partial b} = 0 \quad \text{e} \quad \frac{\partial f}{\partial c} = 0$$

encontramos

$$a = 38/27, \quad b = -20/9 \quad \text{e} \quad c = 8/9$$

Lembrando que $h_{\text{opt}}(n) = E(\varepsilon|n)$ e que

$$h_{\text{opt}}(n) = \frac{38}{27} + \left(-\frac{20}{9}\right)n + \left(\frac{8}{9}\right)n^2$$

encontramos

$$E(\varepsilon|n) = \begin{cases} 2/27 & 0 \leq n \leq 1/3 \\ 14/27 & 1/3 < n < 2/3 \\ 38/27 & 2/3 \leq n \leq 1 \end{cases}$$

Resolvendo usando a ortogonalidade

Usando novamente que $h(n) = a + b n + c n^2$ e
impõendo as condições:

~~$\langle \varepsilon - h(n), 1 \rangle = 0$~~

~~$\langle \varepsilon - h(n), n \rangle = 0$~~

$$\left\{ \begin{array}{l} \langle \varepsilon - h(n), 1 \rangle = 0 \\ \langle \varepsilon - h(n), n \rangle = 0 \\ \langle \varepsilon - h(n), n^2 \rangle = 0 \end{array} \right.$$

(5)

$$\Rightarrow \begin{cases} \int_0^1 (2x^2 - h(n(x))) \cancel{\frac{1}{n(x)}} dx = 0 \\ \int_0^1 (2x^2 - h(n(x))) n'(x) dx = 0 \\ \int_0^1 (2x^2 - h(n(x))) n''(x) dx = 0 \end{cases}$$

$$\Rightarrow \begin{cases} -5a/3 - b - c + 2/3 = 0 \\ -3a - 5b/3 - c + 10/27 = 0 \\ -17a/3 - 3b - 5c/3 + 58/81 = 0 \end{cases}$$

$$\Rightarrow a = 0,88\ldots$$

$$b = -2,22\ldots$$

$$c = 1,407\ldots$$

Llevando aos mesmos resultados.

Exercício 2.6

$X \sim \text{Uniforme}(0,1)$

$$\varepsilon = 2x^2$$

$$n(x) = 1 - |2x - 1| = \begin{cases} 2x & 0 < x < 1/2 \\ 2 - 2x & 1/2 < x < 1 \end{cases}$$

Encontre $E(\varepsilon|n)$. Novamente, sabemos que

$$\langle \varepsilon - E(\varepsilon|n), n(x) \rangle = 0$$

$$E[\Gamma(\varepsilon - E(\varepsilon|n)) n(x)] = 0$$

$$\int_0^{1/2} (2x^2 - E(\varepsilon|n)) 2x \, dx + \int_{1/2}^1 (2x^2 - E(\varepsilon|n)) (2-2x) \, dx = 0 \quad (6)$$

Igualando cada termo a zero e desestando

$$E(\varepsilon|n) = h(n)$$

termos

$$\int_0^{1/2} (2x^2 - h(2x)) 2x \, dx = 0$$

$$\text{mudando de variável } n = 2x \rightarrow dn = 2 \, dx \\ x = n/2 \qquad \qquad \qquad dx = dn/2$$

$$\int_0^{1/2} \left(\frac{2n^2}{2^2} - h(n) \right) n \frac{dn}{2} = 0$$

$$\Rightarrow \frac{2n^2}{4} - h(n) = 0$$

$$h(n) = n^2/2$$

$$\text{com } n = 2x \\ 0 \leq n \leq 1$$

Para o outro termo, temos

$$\int_{1/2}^1 (2x^2 - h(2-2x)) (2-2x) \, dx = 0$$

$$\text{usando } n = 2-2x \Rightarrow x = (2-n)/2 \\ dx = -\frac{dn}{2}$$

$$n(1/2) = 1$$

$$n(1) = 0$$

$$\int_1^0 \left(\frac{2n^2}{2^2} - h(n) \right) \frac{-n}{2} \, dn = 0$$

(7)

$$\Rightarrow \frac{(2-n)^2}{2} - h(n) = 0$$

$$\boxed{h(n) = \frac{(2-n)^2}{2}} \quad 1 \leq n \leq 0$$

Como ambas estão no mesmo domínio, a solução é a soma

$$h(n) = \frac{n^2}{2} + \frac{(2-n)^2}{2} = \frac{n^2}{2} + \frac{4-4n+n^2}{2}$$

$$\boxed{h(n) = n^2 - 2n + 2} \quad 0 \leq n \leq 1$$

Entropia de um número de 3-bits

$$(01)(01)(01) \text{ com } P(0) = P(1) = 1/2$$

O Shannon information content é

$$h(x) = \log_2 1/P(x)$$

A entropia de Shannon é

$$H(x) = \sum_x P(x) \log_2 1/P(x)$$

Assim a entropia de um único dígito é

$$H = \sum_{x=\{0,1\}} P(x) \log_2 1/P(x)$$

$$= \frac{1}{2} \log_2 (1/(1/2)) + \frac{1}{2} \log_2 (1/(1/2)) = \log_2 \frac{1}{(1/2)} = 1 \text{ bit}$$

(8)

"gasta-se um bit para representar um número com um bit"

Para um número de 3-bits com $P(0) = P(1) = 1/2$, temos 8 possibilidades igualmente prováveis:

000	100
001	101
010	110
011	111

$$\rightarrow P(\dots) = 1/8$$

De modo que a entropia fica

$$H(X) = \sum_{x \in \{\dots\}} P(x) \log_2 1/P(x)$$

$$= \frac{1}{8} \log_2 \frac{1}{(1/8)} \times 8 = \log_2 2^3 = 3 \text{ bits}$$

Podríamos pensar nesses 3 bits como a quantidade de informação necessária para representar o número. Ou seja, o primeiro dígito é 0 ou 1? O segundo? O terceiro? Gastamos um bit para cada pergunta, logo 3 bits no total.

Problemas das 9 bolas

Temos um conjunto de 9 bolas, sendo uma delas mais pesada que as outras. A incerteza em descobrir qual a bola mais pesada ou a informação gasta para isso pode ser estimada via $H(X)$. Como a probabilidade de ser a mais pesada é igual para todas, $P_i(\text{ser pesada}) = \frac{1}{9}$.

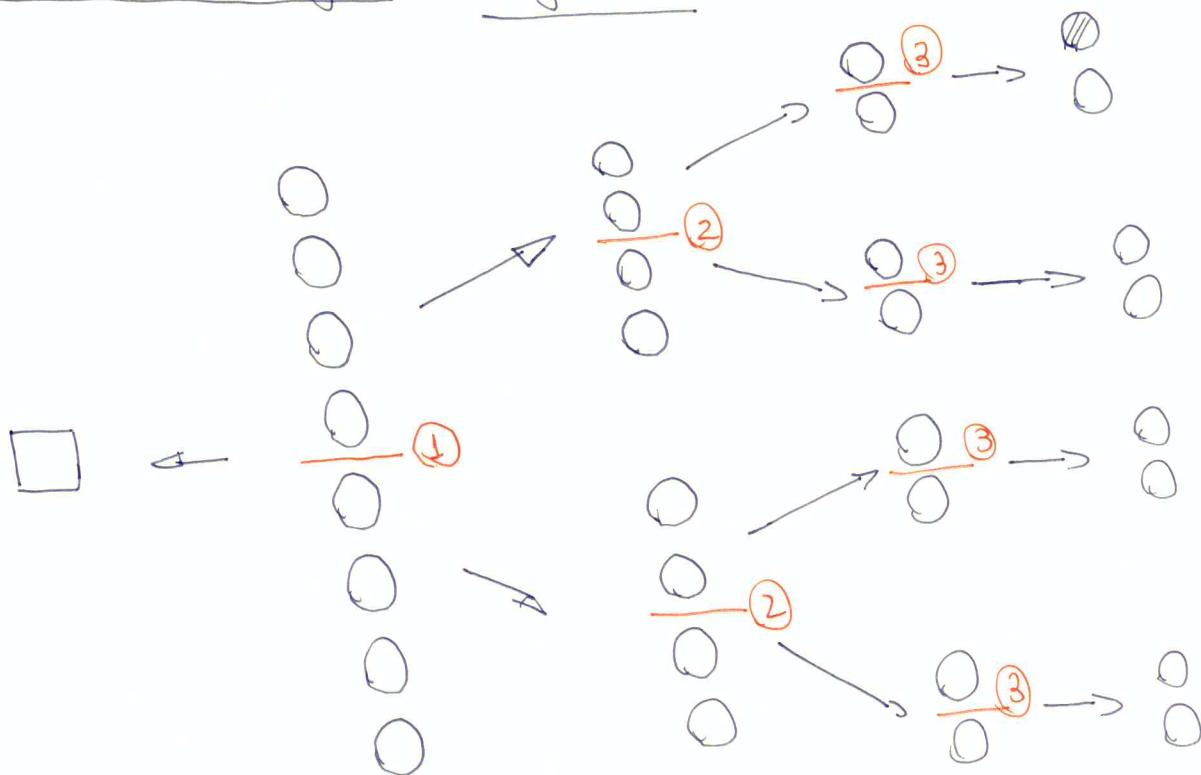
9

Logo a entropia fica

$$H(x) = \sum_{x \in \{\dots\}} P(x) \log_2 1/P(x)$$

$$= \frac{1}{g} \log_2 \frac{1}{1/g} \times g = \log_2 g$$

Primeira solução Fig. 2.9



Se a bola pesada estiver no quadrado, o valor informacional é

$$h(x) = \log_2 1/P(x)$$

$$= \log_2 1/(1/g) = \log_2 g$$

Caso as bolas pesadas estejam entre os círculos, o valor informacional é

$$\textcircled{1} \quad P = 8/9$$

$$\textcircled{2} \quad P = 4/4 \quad \textcircled{3} \quad P = 1/2$$

$$h_1(x) = \log_2 g/8$$

$$h_2(x) = \log_2 4$$

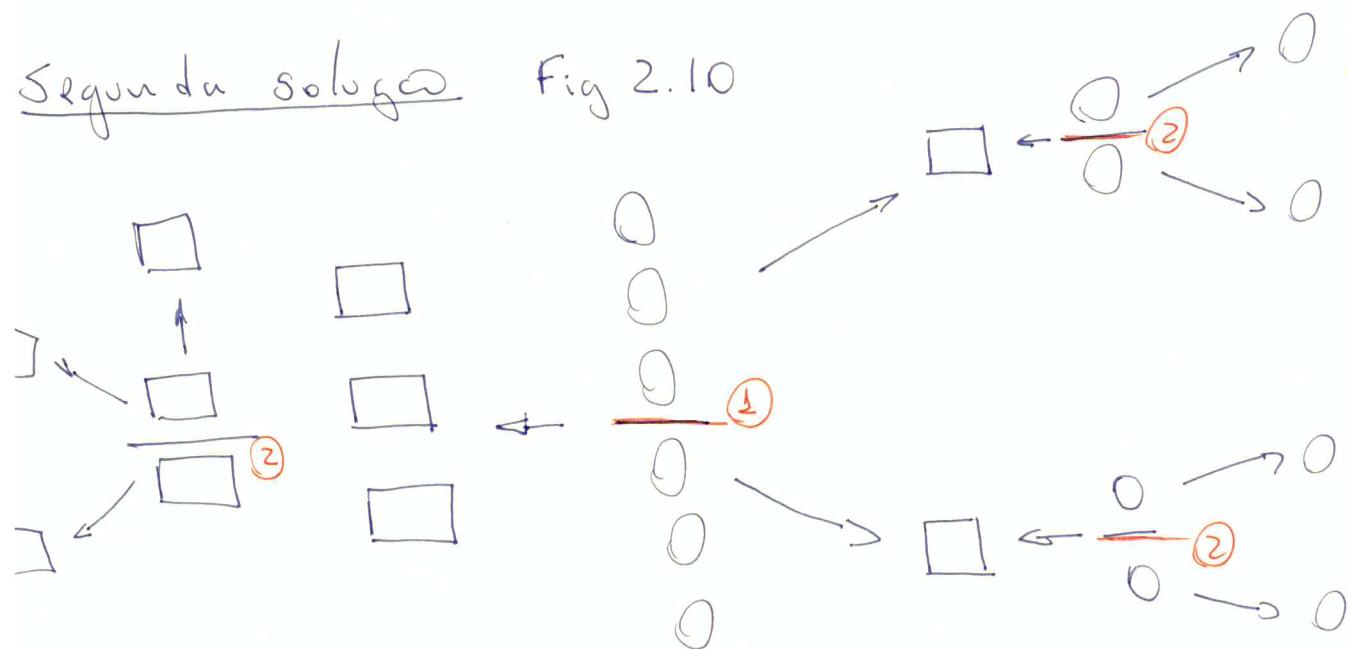
$$h_3(x) = \log_2 2$$

(10)

Sómandos, temos

$$H(x) = \log_2 [(9/8) \cdot 4 \cdot 2] = \log_2(9)$$

O mesmo valor da entropia.

Segunda solução Fig 2.10

Se a bola pesada estiver à esquerda, temos

$$P_1 = 3/9 \quad P_2 = 1/3$$

$$h_1 = \log_2 \frac{1}{3/9} \quad h_2 = \log_2 \frac{1}{1/3}$$

$$h_1 = \log_2 3 \quad h_2 = \log_2 3$$

$$\text{Logo } H = \log_2 3 + \log_2 3 = \log_2 9$$

Se a bola estiver à direita, temos

$$P_1 = 6/9 = 2/3$$

$$h_1 = \log_2 3/2$$

$$P_2 = 1/3$$

$$h_2 = \log_2 3$$

$$P_3 = 1/2$$

$$h_3 = \log_2 2$$

$$\text{Logo } H = \log_2 3/2 + \log_2 3 + \log_2 2$$

$$H = \log_2 9$$

o mesmo valor de antes.

Propriedades da entropia informacional

(11)

$$H(x) \geq 0$$

sendo a igualdade verificada apenas se $P(x)=1$. Nesse caso, não temos nenhuma incerteza. Além disso, $H(x)$ é máxima quando os elementos de P são equiprováveis.

Prova (discreto) Seja $P = \{P_1, P_2, \dots, P_N\}$ com

$$\sum_{i=1}^N P_i = 1$$

Vamos maximizar

$$H(P) = \sum_{i=1}^N P_i \log \frac{1}{P_i}$$

Para isso usamos multiplicadores de Lagrange

$$f(P_1, P_2, \dots, P_N, \lambda) = - \sum_{i=1}^N P_i \log P_i + \lambda \left\{ \sum_{i=1}^N P_i - 1 \right\}$$

impondo

$$\frac{\partial f}{\partial P_j} = 0 \Rightarrow -P_j \frac{1}{P_j} - \log P_j + \lambda = 0$$

$$\log P_j = \lambda - 1$$

$$P_j = e^{\lambda-1} = cte = c \quad \forall j$$

Note que P_j é uma constante, logo

$$\sum_{i=1}^N P_i = 1 \Rightarrow Nc = 1$$

$$c = 1/N$$

Logo $P_j = 1/N$

(12)

Exemplo $P = \{P, 1-P\}$

$$H = - \sum P_i \log_2 P_i$$

$$H = -P \log_2 P - (1-P) \log_2 (1-P)$$

$$H = -P \log_2 P - \log_2(1-P) + P \log_2(1-P)$$

$$\frac{\partial H}{\partial P} = 0 \Rightarrow P = 1/2$$

Entropia em duas variáveis

$$x, y \rightsquigarrow P(x,y)$$

seja $H(x,y) = - \sum_{x,y} P(x,y) \log_2 P(x,y)$

Note que $H(x,y) = H(x) + H(y)$

explicar se $P(x,y) = P_x(x) P_y(y)$

Prova:

$$H(x,y) = - \sum_{x,y} P_x(x) P_y(y) \log_2 P_x(x) P_y(y)$$

$$= - \sum_{x,y} P_x(x) P_y(y) \log_2 P_x(x)$$

$$- \sum_{x,y} P_x(x) P_y(y) \log_2 P_y(y)$$

$$= - \sum_x P_x(x) \log_2 P_x(x) - \sum_y P_y(y) \log_2 P_y(y)$$

$$H(x,y) = H(x) + H(y)$$

13

Divergência de Kullback-Leibler

Sejam P e Q , duas distribuições, a divergência K-L é dada por

$$D_{KL}(P||Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)}$$

Propriedades:

$$D_{KL}(P||Q) \geq 0$$

sendo a igualdade válida apenas se $P = Q$.

Algumas vezes é chamada de distância, embora não seja uma métrica, isto é,

$$D_{KL}(P||Q) \neq D_{KL}(Q||P)$$

Podemos definir uma versão simétrizada:

$$D_{KL}^*(P||Q) = \frac{1}{2} D_{KL}(P||Q) + \frac{1}{2} D_{KL}(Q||P)$$

Intrigas sobre a D_{KL}

Suponha que devemos transmitir um conjunto de mensagens

$$\{(x_1, P(x_1)), (x_2, P(x_2)), \dots, (x_n, P(x_n))\}$$

Baseado na ideia de entropia informacional, podemos escolher o tamanho da mensagem com base em $\log_2 1/P(x)$. Assim, mensagens mais frequentes irão consumir o menor número de bits.

Já seja, quanto maior $P(x)$, menor será $-\log_2 P(x)$. (14)

Nesse caso, a entropia envolvida na transmissão de K mensagens fica

$$H(x) = \sum_k P(x_k) \log_2 1/P(x_k)$$

Suponha agora que vamos transmitir a mesma mensagem, mas com probabilidades diferentes $Q(x)$. Nesse caso, a entropia (cruzada) fica

$$H_q(x) = \sum_k P(x_k) \log_2 1/Q(x_k)$$

Note que apenas a probabilidade associada ao comprimento da mensagem muda, não a probabilidade das mensagens em si. A diferença entre as duas

$$D_{KL}(P, Q) = H_q(x) - H(x)$$

$$= \sum_x P(x) \log_2 \frac{P(x)}{Q(x)}$$

Assim a KL divergência é a diferença média na codificação do comprimento do mesmo conjunto de mensagens entre dois regimes de probabilidade.

Isso explica a ausência de simetria, ou seja, por si só, P e Q podem fornecer a melhor estratégia para codificar o tamanho das mensagens, mas pode não haver simetria entre as duas estruturas de arquitetura para transmitir as mensagens.

Funções geradoras de momentos (característicos) (15)

São definidas como:

$$M(t) = E\{e^{tx}\}$$

Assim, o primeiro momento é

$$\frac{dM}{dt} = \frac{d}{dt} E\{e^{tx}\} = E\left\{\frac{d}{dt} e^{tx}\right\}$$

$$\frac{dM}{dt} = E\{x e^{tx}\}$$

Tomando $t=0$, temos

$$\left. \frac{dM}{dt} \right|_{t=0} = E\{x\} = M^{(1)}(0)$$

Similarmente

$$\begin{aligned} \frac{d^n}{dt^n} M &= E\left\{\frac{d^n}{dt^n} e^{tx}\right\} \\ &= E\{x^n e^{tx}\} \end{aligned}$$

Logo, para $t=0$, temos

$$\left. \frac{d^n}{dt^n} M \right|_{t=0} = M^{(n)}(0) = E\{x^n\}$$

Desse modo, por exemplo, a variância fica

$$V(x) = E(x^2) - E(x)^2 = M^{(2)}(0) - M^{(1)}(0)$$

Exemplo com a Binomial

$$P_{K;n,p} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = B(n,p)$$

é probabilidade de se obter k sucessos em n tentativas.

$$M(t) = E\{e^{tK}\}$$

$$= \sum_{k=0}^n e^{tk} \binom{n}{k} p^k (1-p)^{n-k}$$

$$= \sum_{k=0}^n \binom{n}{k} (pe^t)^k (1-p)^{n-k}$$

Usando que

$$(x+y)^n = \sum_{k=0}^n \frac{n!}{k!(n-k)!} x^k y^{n-k}$$

$$= \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}$$

temos:

$$M(t) = [pe^t + 1 - p]^n$$

$$M(t) = [p(e^t - 1) + 1]^n$$

Teorema da unicidade: duas variáveis x e y tendo a mesma função geradora de momentos, tem a mesma distribuição de probabilidade.

Exemplo: suponha conhecida a probabilidade condicional de X dado $V = p$, sendo \hookrightarrow uniforme $[0,1]$

$$P(X|V=p) = \mathcal{B}(n,p)$$

Isso pode representar o número de caras em n lançamentos de moedas cuja probabilidade de sair cara é p . Queremos encontrar a probabilidade não condicionada para X . Sabemos que

$$E(e^{tX}|V=p) = (pe^t + 1-p)^n$$

Vamos ver que V é uniforme, podemos integrar em

$$\begin{aligned} E(e^{tX}) &= \int_0^1 E(e^{tX}|V=p) dp \\ * \sum_{k=0}^n r^k &= \frac{1-e^{n+1}}{1-e} \\ \text{se } r = e^t &= \int_0^1 (pe^t + 1-p)^n dp \\ \frac{1-e^{t(n+1)}}{1-e^t} &= \frac{1}{n+1} \cdot \frac{e^{t(n+1)} - 1}{e^t - 1} = \frac{1}{n+1} \cdot \frac{1 - e^{t(n+1)}}{1 - e^t} \\ * \frac{1-e^{t(n+1)}}{1-e^t} &= \frac{1}{n+1} \left\{ 1 + e^t + e^{2t} + \dots + e^{nt} \right\} \end{aligned}$$

Visto que o momento é o valor esperado de e^{tX} , ou seja,

$$E(e^{tX}) = \sum_{i=0}^n P_i e^{tx_i}$$

Vemos que as variáveis $x = (0, 1, 2, \dots, n)$, todas

com

$$P_i = \frac{1}{n+1} \quad \forall i \in \{0, 1, \dots, n\}$$

ou seja X é um conjunto uniforme de variáveis aleatórias discretas. Num exemplo concreto, se deixarmos cair uma caixa de moedas cuja probabilidade de sair cara não é conhecida e, em seguida, contarmos o número de moedas para, essa distribuição será uniforme.

Funções geradoras de momento e soma de variáveis aleatórias independentes

Seja $Y = X_1 + X_2$, então

$$\begin{aligned} M_Y(t) &= E(e^{tY}) = E(e^{tX_1+tX_2}) \\ &= E(e^{tX_1}) E(e^{tX_2}) \end{aligned}$$

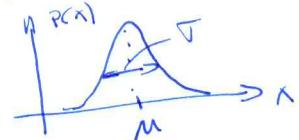
$$\boxed{M_Y(t) = M_{X_1}(t) M_{X_2}(t)}$$

Exemplo: duas gaussianas somadas

$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$X_1 \sim N(\mu_1, \sigma_1^2)$$

$$X_2 \sim N(\mu_2, \sigma_2^2)$$



Queremos $Y = X_1 + X_2$. Podemos usar que

$$M_Y(t) = M_{X_1}(t) M_{X_2}(t)$$

$$M_{X_1}(t) = E(e^{tX_1})$$

$$= \int_{-\infty}^{\infty} \frac{e^{-(x_1 - \mu_1)^2/2\sigma_1^2}}{\sqrt{2\pi\sigma_1^2}} e^{tx_1} dx_1$$

$$= \exp \left\{ t \left(\mu_1 + \frac{\sigma_1^2 t}{2} \right) \right\}$$

Similarmente temos para a outra variável. Logo,

$$M_Y(t) = \exp \left\{ \frac{t}{2} \left(2\mu_0 + 2\mu_1 + \sigma_0^2 t + \sigma_1^2 t \right) \right\}$$

a agrupando as potências em t , temos (veja SymPy)

$$M_Y(t) = \exp \left\{ t(\mu_0 + \mu_1) + t^2 \left(\frac{\sigma_0^2 + \sigma_1^2}{2} \right) \right\}$$

Comparando com a forma para a gaussiana padrão,

vemos que

$$Y \sim N(\mu_0 + \mu_1, \sigma_Y^2 = \sqrt{\sigma_0^2 + \sigma_1^2})$$

Técnicas de amostragem de Monte Carlo

Suponha que desejamos gerar números aleatórios com uma distribuição $f(x)$, supondo que temos um gerador uniforme $[0,1]$. Uma abordagem é investigar como um histograma de X amostra da distribuição se aproxima de $f(x)$.

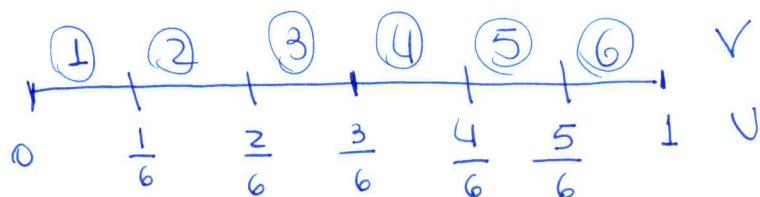
Escrevemos:

$$P(v \in N_\Delta(x)) \approx f(x) \Delta x \quad (*)$$

O que significa que a probabilidade de uma amostra estar nas vizinhanças de x ($v \in N_\Delta(x)$) é aproximada por $f(x) \Delta x$. De outra maneira, podemos pensar nisso como uma condição para gerar amostras v .

Método da CDF inversa

Exemplo Dada de 6 faces a partir da $U[0,1]$.



Na linguagem da eq. (*)

$$P(v=2) = P(v \in [1/6, 2/6]) = f(x) \Delta x = 1 \cdot \frac{1}{6}$$

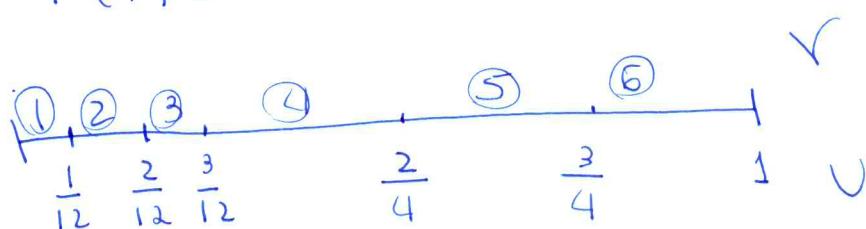
O que vale para todos os demais v .

Exemplo no parâmetro

Suponha um dado não justo

$$P(1) = P(2) = P(3) = 1/12$$

$$P(4) = P(5) = P(6) = 1/4$$



Note que esse método é chamada CDF inversa porque estamos invertendo a CDF de v

$$\{0, 1/12, 2/12, 3/12, 2/4, 3/4, 1\}$$

CDF inversa para o caso contínuo

(21)

No caso contínuo, a condição \star se torna

$$\begin{aligned} P(F(x) < v < F(x + \Delta x)) &= F(x + \Delta x) - F(x) \\ &= \int_x^{x+\Delta x} f(x') dx' \approx f(x) \Delta x \end{aligned}$$

$F(x) = P(X < x)$. Isso que a probabilidade de v estar entre $F(x)$ e $F(x + \Delta x)$ é igual a probabilidade de encontrar um \bar{X} menor que $(x + \Delta x)$ menos a probabilidade de encontrar \bar{X} menor que x .

$$F(x) = P(\bar{X} < x) = \int_{-\infty}^x f(x') dx'$$

$$P(x + \Delta x) = \int_{-\infty}^{x+\Delta x} f(x') dx'$$

$$F(x + \Delta x) - F(x) = \int_{-\infty}^{x+\Delta x} \dots - \int_{-\infty}^x \dots$$

$$= \int_x^{-\infty} \dots + \int_{-\infty}^{x+\Delta x}$$

$$= \int_x^{x+\Delta x} f(x') dx' \approx f(x) \Delta x$$

$$\approx f(u) \Delta u$$

O fruque é considerar $v \sim [0,1]$ e uma CDF que seja inversível ~~de forma fácil~~, no sentido

$$F(F^{-1}(u)) = u$$

Além disso, como v é uniforme, temos

na v.a. uniforme

$$\begin{aligned} P(F(x) < u < F(x+\Delta x)) &= P(x/F'(u) < x+\Delta x) \\ P(x) &= \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{fora} \end{cases} \\ &\stackrel{*}{=} F(x+\Delta x) - F(x) = \int_x^{x+\Delta x} f(x) dx \\ &\stackrel{*}{=} f(x) \Delta x \\ &= P(x \leq F^{-1}(u) \leq x + \Delta x) \end{aligned}$$

Dessse modo vemos que se

$$v = F^{-1}(u)$$

v será distribuído com $f(x)$.

Outro modo mais direto:

$$v \sim U[0,1]$$

$$x \sim F_x(v) \quad (\text{CDF de } x)$$

Podemos procurar por uma transformação tal que

$$T(u) \sim F_x(v)$$

Para u sabemos que

$$\begin{aligned} F_x(x) &= P(x < u) = P(T(v) < u) \\ &= P(v < T^{-1}(u)) = T^{-1}(u) \end{aligned}$$

pois é uniforme

Assim

$$F_x(u) = T^{-1}(u)$$

$$\text{Logo } T(u) = F_x^{-1}(u)$$

Ou seja, se $\bar{X} = F_x^{-1}(u)$, \bar{X} tem CDF igual a $F_x(x)$.

Exemplo exponencial

pdf: $f_x(x) = \alpha e^{-\alpha x} \quad x \in [0, \infty)$

cdf: $F(x) = \int_{0}^x \alpha e^{-\alpha x'} dx' = -e^{-\alpha x'} \Big|_0^x = -\left(e^{-\alpha x} - 1\right)$

$$F(x) = 1 - e^{-\alpha x}$$

Inversa

$$F(F^{-1}(x)) = x$$

$$1 - e^{\alpha F^{-1}(x)} = x$$

$$x - 1 = e^{\alpha F^{-1}(x)}$$

$$\ln(x-1) = \alpha F^{-1}(x)$$

$$F^{-1}(x) = \frac{-1}{\alpha} \ln(x-1)$$

$$F^{-1}(x) = \frac{1}{\alpha} \ln \frac{1}{(x-1)}$$

Assim

$$F^{-1}(u) = \frac{1}{\alpha} \ln \frac{1}{(x-1)}$$

tem pdt $f_x(x) = \alpha e^{-\alpha x}$.

Método da rejeição

Sejam $v_1 \sim [a, b]$
 $v_2 \sim [a, b]$

e suponha que desejamos gerar amostras de
 $x \sim f(x)$

$$F(u_i) = \int_a^{u_i} f(v_i) dv_i$$

$$F(u) = \frac{v_i - a}{b - a}$$

$$f(u_i) = \frac{1}{b-a} \quad a \leq u_i \leq b$$

Para isso, consideramos

$$\begin{aligned} P(v_1 \in N_{\Delta}(x) \wedge v_2 < \frac{f(u_i)}{M}) \\ \cong \frac{\Delta x}{b-a} \cdot \left(\frac{\frac{f(u_i)}{M} - a}{b - a} \right) \end{aligned}$$

$$\cong \frac{\Delta x}{b-a} \frac{f(u_i)}{M}$$

Assim, sorteamos um número aleatório $v_i \in [a, b]$
 usamos em $f(x)$ e se a segunda condição
 for válida, v_i é uma amostra de $f(x)$. Note que
 a probabilidade à esquerda é proporcional a $f(x)$
 A variável M tem o papel de rescalar $f(x)$ para
 que fique no range de v_2 .
 A eficiência desse método está associada com

a probabilidade de acertar v_i como
uma amostra de $f(x)$, ou seja

$$\sim \int_{M(b-a)}^{f(x) dx} = \frac{1}{M(b-a)}$$

Assim, M não pode ser muito grande nem o
range dos valores de v_i .

Exemplo:

$$f(x) = e^{-\frac{(x-\bar{x})^2}{2x}} \cdot \frac{(x+1)}{12}$$

Veja que não temos a forma fechada para a CDF.

Uma melhoria do método é substituir a distribuição
uniforme de v_i por uma outra que seja mais
proxima de $f(x)$. Especificamente

$$v_1 \rightsquigarrow g(x)$$

$$v_2 \rightsquigarrow [0,1]$$

De modo que

$$P(v_1 \in N_\Delta(x) \wedge v_2 < \frac{f(v_1)}{M})$$

$$\equiv g(x) \Delta x \frac{f(v_1)}{M}$$

Veja que esse resultado não é proporcional a
 $f(x)$. Para que fique, devemos tomar
 $f \rightarrow h = f(x)/g(x)$

na condição para v_2 , isto é,

$$P(v_i \in N_\Delta(x) \wedge v_2 < \frac{h(v_i)}{h_{\max}})$$

$$\approx g(x) \Delta x \frac{h(v_i)}{h_{\max}}$$

$$\approx g(x) \Delta x \frac{f(v_i)}{g(v_i) h_{\max}}$$

$$\approx \Delta x \frac{f(v_i)}{h_{\max}}$$

Integrando em v_i , temos a probabilidade de v_i ser acitivo:

$$\sim \int \frac{\Delta x f(v_i) dv_i}{h_{\max}} \sim \frac{1}{h_{\max}}$$

Desigualdades úteis

Desigualdade de Markov \rightarrow (positiva)

Se X uma variável aleatória com média finita. Então para algum t , temos

$$P(X > t) \leq E(X)/t$$

dem:

$$E(X) = \int_0^\infty x f(x) dx = \int_0^t x f(x) dx + \int_t^\infty x f(x) dx$$

$$\geq \int_t^\infty x f(x) dx \geq t \int_t^\infty f(x) dx$$

mais $x > t$

(27)

$$E(x) \geq t \int_t^{\infty} f(x) dx$$

$$E(x) \geq t P(x > t)$$

$$P(x > t) \leq E(x)/t$$

E' considerada uma desigualdade "frouxa" no sentido que existe grande diferença entre os dois lados.

Desigualdade do Chebyshov

Seja x uma variável aleatória e $t > 0$ um parâmetro, então

$$P(|x - \mu| > t) \leq \sigma^2/t^2$$

$$\text{sendo } \mu = E(x) \quad \sigma^2 = V(x) = E((x - E(x))^2)$$

Prova

$$V(x) = \int_{-\infty}^{\infty} (x - E(x))^2 f(x) dx$$

seja $A \rightarrow (-\infty, \infty)$ e A os pontos de \mathbb{R} que satisfazem a condição $|x - E(x)| > t$

$$V(x) = \int_{x \in A} (x - E(x))^2 f(x) dx + \int_{x \notin A} (x - E(x))^2 f(x) dx$$

$$\geq \int_{x \in A} (x - E(x))^2 f(x) dx \geq \int_{x \in A} t^2 f(x) dx$$

pois $(x - E(x)) > t$

$$V(x) \geq t^2 \int_{x \in A} f(x) dx$$

$$V(x) \geq t^2 P(|x - E(x)| \geq t)$$

$$P(|x - E(x)| \geq t) \leq V(x) / t^2$$

Note que se a variável x for normalizada

$$z = \frac{x - \mu}{\sigma}$$

ficamos com

$$P(|z| \geq t) \leq 1/t^2$$

Desigualdade de Hoeffding

Seja X uma variável aleatória, tal que $a \leq X_i \leq b$ e $\mu \in \mathbb{R}$. Então

$$\frac{-2n\epsilon^2}{(b-a)^2}$$

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-\frac{-2n\epsilon^2}{(b-a)^2}}$$

onde $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ e $\mu = E(X)$

corolário se $P(a \leq X_i \leq b) = 1$ com $E(X) = \mu$

$$|\bar{X}_n - \mu| \leq \sqrt{\frac{c}{2n} \log \frac{2}{\delta}}$$

$$\text{com } c = (b-a)^2$$