MPEI 2015-2016

Aula 14

Aplicação interessante de soma de variáveis aleatórias

Variáveis aleatórias e informação

Aplicação – Contagem

Motivação

- Evitar contadores grandes quando o volume de dados é grande.
- Como um contador de n bits contará no máximo até 2^n eventos, será este o limite a ultrapassar.

Primeira solução

- Para duplicar o número de eventos que se podem contar, incrementa-se o contador com probabilidade 1/2 cada vez que ocorre um evento.

- A ideia *é incrementar o contador metade das vezes.*

…

- Com base na função rand() podemos agora tomar decisões aleatórias com probabilidade 1/2 e portanto construir uma função para incrementar (ou não) o contador:

```
if (rand() < 0.5) then
    incrementar_contador
endif
```

Em octave/ Matlab

- Podemos facilmente simular o resultado após 100 eventos:

```
% gera 100 var aleatórias indep em [0,1]
x = rand(1, 100);
% calcular quantas são < 0.5
n = sum(x < 0.5);
```

DEMO

- N representará o valor do contador após os 100 eventos
- O contador é uma variável aleatória, determinada por uma sucessão de experiências aleatórias

Qual é o valor médio do contador após k eventos?

- Associando uma variável aleatória a cada evento, de forma a representá-lo probabilisticamente.
- Seja X_i a variável aleatória que representa o incremento i, com valor 1 se o contador foi incrementado, e valor zero caso contrário.
- Como $P(X_i = 0)$ e $P(X_i = 1)$ são iguais a 1/2, tem-se
- $E[X_i] = 0 \times P(X_i = 0) + 1 \times P(X_i = 1) = \frac{1}{2}$

Valor médio

- O valor do contador após k eventos é a soma dos k incrementos, $S = X_1 + X_2 + \cdots + X_k$

- E o valor médio:
- $E[S] = E[X_1 + X_2 + \cdots + X_k]$
- $= E[X_1] + E[X_2] + \cdots + E[X_k]$
- $= \frac{1}{2} + \cdots + \frac{1}{2} = \frac{k}{2}$

- Como o valor médio do contador após k eventos é $k/2$, o número de eventos pode ser estimado através do dobro do número registado pelo contador.

Variância

- A variância de um qualquer dos X_i é
- $Var(X_i) = E\left[{X_i}^2\right] - (E[X_i])^2$

- $E\left[{X_i}^2\right] = 0^2 \times P(X_i = 0) + 1^2 \times P(X_i = 1)$
- $= \frac{1}{2}$

- $Var(X_i) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$

Variância (continuação)

- Como as variáveis X_i são independentes, a variância de S é
- $Var(S) = Var(X_1 + X_2 + \cdots + \ X_k)$
- $= Var(X_1) + Var(X_2) + \cdots + \ Var(X_k)$
- $= \frac{1}{4} + \cdots + \frac{1}{4} = \frac{k}{4}$
- O que implica $\sigma = \frac{\sqrt{k}}{2}$
- Para n=10000 teremos:
  - média 5000
  - desvio padrão 50

Distribuição de probabilidade

- Pode calcular-se a probabilidade de, após k eventos, o valor do contador ser n.

- Fixemos $k = 4$:
- Teremos $X_1, X_2, X_3\ e\ X_4$
  – Variáveis binárias que descrevem se o contador é incrementado ou não após o evento 1,2,3 e 4
- O que nos dá 16 possibilidades (2^4)

X_1	X_2	X_3	X_4	valor do contador
0	0	0	0	0
0	0	0	1	1
0	0	1	0	1
0	0	1	1	2
0	1	0	0	1
0	1	0	1	2
0	1	1	0	2
0	1	1	1	3
1	0	0	0	1
1	0	0	1	2
1	0	1	0	2
1	0	1	1	3
1	1	0	0	2
1	1	0	1	3
1	1	1	0	3
1	1	1	1	4

- É agora fácil determinar as probabilidades, por contagem:
- $p(0) = \frac{1}{16}$
- $p(1) = \frac{4}{16}$
- $p(2) = \frac{6}{16}$
- $p(3) = \frac{4}{16}$
- $p(4) = \frac{1}{16}$

Generalizando

- Sendo p a probabilidade de incrementar e $1 - p$ a probabilidade de não incrementar …
- A probabilidade de observar uma soma igual a n após k experiências é:

$$p(n) = \binom{k}{n} p^n \ (1-p)^{k-n}$$

Simulações

- Efectue várias simulações e calcule a média das contagens obtidas.
  - Comparar essa média com E[S].

- Qual a variância de S   (valor teórico) ?
- Calcule também a variância das estimativas obtidas pelo programa e compare com o valor anterior.

Variante 1

- Como proceder para alargar mais ainda a gama do contador?
- Imaginemos, por exemplo, que se quer multiplicar por 64 essa gama. A solução natural é incrementar com probabilidade $1/64$ em vez de ½
- O valor médio de X_i será agora $\frac{1}{64}$
- E $E[S] = \quad \dots . = \frac{k}{64}$
- Neste caso, o número de eventos pode ser estimado por $64\,n$, sendo n o valor do contador

Segunda solução

- Neste caso o contador é incrementado com probabilidade cada vez menor à medida que o seu valor aumenta:
- quando o contador contém n, a probabilidade de um incremento é 2^{-n}

Eventos	*Valor do contador*	*Número de eventos*
x	1	1
x		
x	2	3
x		
x		
x		
x	3	7
x		
x		
x		
x		
x		
x		
x		
x	4	15

Para mais informação

- Ver notas do ano lectivo anterior de autoria do Prof. Paulo Jorge Ferreira
  - (disponíveis no elearning da UC)

Informação

Probabilidade e informação são conceitos relacionados

Exemplo

- Suponhamos que temos um conjunto de letras, que designamos por E
- E={a, b, c, d, e, f, g, h}
- Podemos codificar E por:
  - {000,001,010,011,100,101,110,111}

- Encontrar uma letra em E pode ser feito com 3 perguntas, tantas quantas os bits necessários para codificar cada elemento de E

…

- Por exemplo, encontrar "c":
- Questão 1: Encontra-se na metade esquerda ?
  – Resposta: Sim (tinhamos a,b,c,d,e,f,g)
- Questão 2: Encontra-se na metade esquerda (do domínio anterior) ?
  – Resposta: Não (tinhamos a,b,c,d)
- Questão 3: Encontra-se na metade esquerda (do domínio anterior) ?
  – Resposta: Sim (tinhamos c,d)

Informação e probabilidade

- Probabilidade e informação são conceitos relacionados.
- Ao comunicar a ocorrência de um acontecimento transfere-se uma quantidade de informação que depende da probabilidade do acontecimento,
  - o que sugere que se escreva a informação como função da probabilidade, $I(p)$.
- Se o acontecimento for certo, não se comunica qualquer informação.
  - Logo, $I(p) = 0$
- Quanto mais improvável o acontecimento, maior a informação associada.
  - Logo $I(p)$ deve crescer à medida que p decresce.

Informação de 2 acontecimentos indep.

- A informação a associar a dois acontecimentos independentes deverá ser a soma da informação associada a cada acontecimento em separado.
- Como a probabilidade da ocorrência de dois eventos independentes, com probabilidades p e q, é o produto pq, a informação deverá satisfazer

$$I(pq) = I(p) + I(q)$$

logaritmo

- A função logaritmo satisfaz $\log pq = \log p + \log q$, o que sugere que se tome $I(p) = \log p$.
- Contudo, esta função decresce quando p diminui.
- Trocando-lhe o sinal obtém-se a função
  $$I(p) = \log\left(\frac{1}{p}\right)$$
- que satisfaz $I(1) = 0$ e cresce quando p diminui, qualquer que seja a base do logaritmo

Informação de 2 acontecimentos indep. (continuação)

- A informação associada a eventos independentes A e B, com probabilidade conjunta pq , é
- $I(pq) =$
- $= log\,\frac{1}{pq}$
- $= log\,\frac{1}{p} + log\,\frac{1}{q}$
- $= I(p) + I(q)$
  - Como se pretendia

- Quando se toma a base 2 exprime-se a informação em bits

Exemplo

- A quantidade de informação que se transfere ao comunicar o resultado de uma experiência que pode dar dois resultados equiprováveis (logo, de probabilidade ½) é, em bits,

- $I\left(\frac{1}{2}\right) = \log\frac{1}{1/2} = \log 2 = 1 \ bit$

- Para armazenar n bits de dados nem sempre é necessário utilizar n bits de memória.

Exemplo 2 – aplicando aos contadores

- No caso do contador inicial :
- Probabilidade de cada incremento igual a 1/2
- Informação:

$$I\left(\frac{1}{2}\right) = 1\ bit$$

- Na variante :
- $I(X_i) = I\left(\frac{1}{64}\right) = \log\frac{1}{64} = 6\ bits$

Entropia

- A entropia é uma medida da imprevisibilidade do conteúdo de informação.
- É dada pela média da informação

$$\mathrm{H}(X) = E[I(X)] = \sum_{k=1}^{n} p(x_k) \log\left(\frac{1}{p(x_k)}\right)$$

Entropia - exemplos

- Se considerarmos uma variável aleatória que toma um de dois dois valores equiprováveis, então:

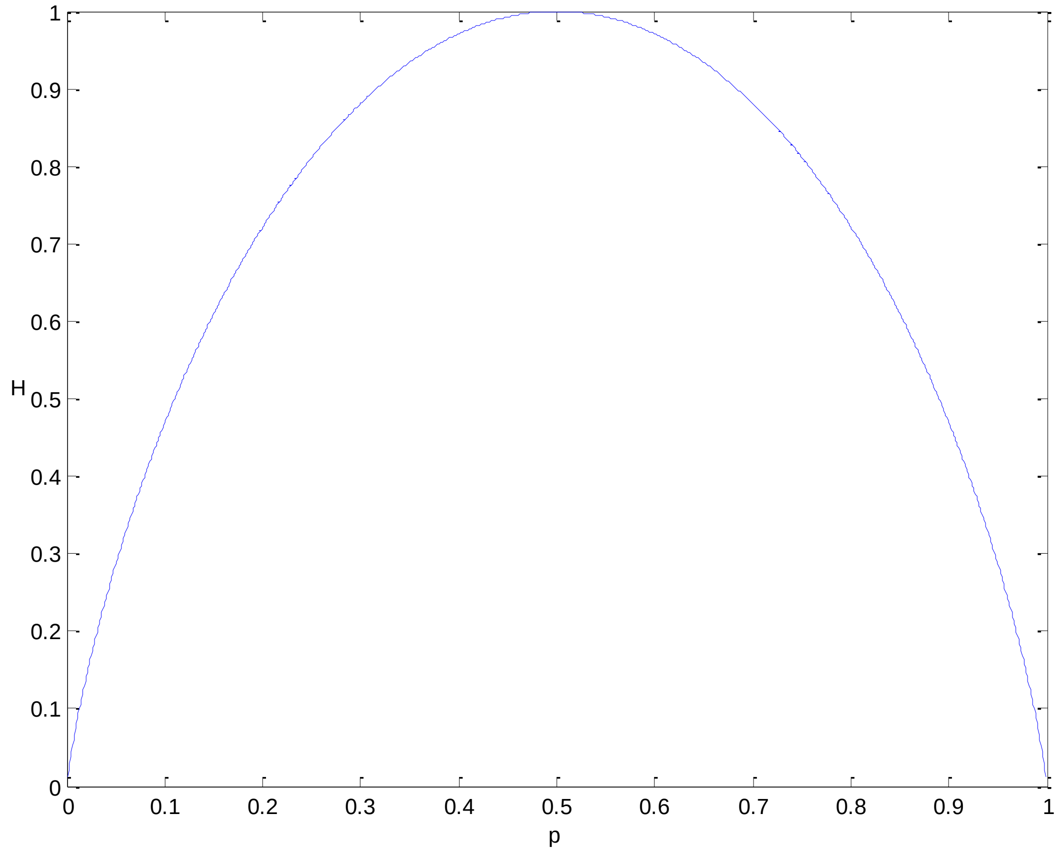
$$\text{H}(X) = \frac{1}{2}\log(2) + \frac{1}{2}\log(2) = 1 \text{ bit}$$

- Por outro lado se as probabilidades dos dois acontecimentos forem p e 1-p então:

$$\text{H}(X) = -p\log(p) - (1-p)\log(1-p)$$

Que varia entre 0 (p=0 ou p=1) e 1 (p= ½).

Entropia - exemplos



Entropia – exemplos

- Se a variável aleatória X tomar um de 2^n valores equiprováveis, então a entropia será:

$$\mathrm{H}(X) = \sum_1^{2^n} 2^{-n} \log(2^n) = 2^n \times 2^{-n} \times n$$
$$= n \text{ bits}$$