MPEI — resumo

Paulo J S G Ferreira

4 de Novembro de 2014

Conteúdo

1 Fundamentos

1.1 Que definição?

A probabilidade de um evento é um número no intervalo $[0, 1]$ que exprime "grau de incerteza" quanto à ocorrência do evento. O valor 0 exprime impossibilidade e o valor 1 certeza absoluta.

Mas como definir probabilidade? Um processo poderia ser o seguinte. Considerem-se n experiências das quais pode resultar o evento em causa. Seja m o número de vezes em que se observou o evento (a sua *frequência absoluta*). O limite quando $n \to \infty$ da *frequência relativa* m/n, caso exista, seria a probabilidade do evento.

É preferível dispensar o limite e o recurso à "realização de n experiências". Poderíamos definir probabilidade de um evento como o quociente m/n, sendo m o número de "casos favoráveis" (o número de resultados em que ocorre o evento) e n o "total de casos" (o número total de resultados possíveis). Mas isto requer que todas os resultados possíveis sejam igualmente prováveis.

Há um método mais geral. Os acontecimentos são vistos como subconjuntos de um conjunto U, que representa o universo de possibilidades, e a probabilidade de cada acontecimento determinada pelo conjunto de princípios seguinte.

1.2 Princípios

Seja U o universo de acontecimentos, que assumo não vazio. A probabilidade $P(A)$ de qualquer acontecimento $A \subset U$ respeita (i) $P(A) \geq 0$ (ii) $P(U) = 1$ (iii) a probabilidade da união[1] de eventos disjuntos é a soma das probabilidades:

$$\boxed{P(A+B) \;=\; P(A)+P(B)} \qquad \text{se } AB = \emptyset.$$

Em geral, tem-se $P(A+B) \leq P(A)+P(B)$.

Para poder determinar a probabilidade da intersecção, união e complemento de qualquer evento é preciso que o conjunto U seja fechado com respeito a essas operações, finitas ou infinitas. Dados os eventos $A_1, A_2, \ldots$ pertencentes a U, basta assumir que a sua união e intersecção está ainda em U.

Resulta destes princípios que o conjunto vazio $\emptyset$ pertence a U e tem probabilidade zero, porque se $A \in U$ então $A^c \in U$, e $AA^c = \emptyset$. Tem-se $P(\emptyset) = 0$ porque $P(A) = P(A+\emptyset) = P(A)+P(\emptyset)$.

1.3 Independência

Dois eventos dizem-se independentes se a ocorrência de um não tiver impacto na ocorrência do outro. Uma análise de frequência sugere a definição

$$\boxed{P(AB) \;=\; P(A)P(B)} \qquad A,\ B \text{ independentes.}$$

A probabilidade de A dado que B ocorreu, que se representa por $P(A|B)$, deve reduzir-se a $P(A)$ se A e

[1] Designo união de A e B por $A+B$ ou $A \cup B$ e a intersecção por AB ou $A \cap B$. O complemento de A costuma ser representado por A^c ou $\bar{A}$.

B forem independentes. Um argumento baseado em frequência sugere a definição[2]

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

Se A e B forem independentes tem-se $P(AB) = P(A)P(B)$ pelo que $P(A|B) = P(A)$, como seria de esperar.

1.4 Informação

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, $I(1) = 0$. Quanto mais improvável o acontecimento, maior a informação associada. Logo $I(p)$ deve crescer à medida que p decresce.

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

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

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

$$I(p) = \log \frac{1}{p}$$

que satisfaz $I(1) = 0$ e cresce quando p diminui, qualquer que seja a base do logaritmo. A informação associada a eventos independentes A e B, com probabilidade conjunta pq, é

$$\begin{aligned} I(pq) &= \log \frac{1}{pq} \\ &= \log \frac{1}{p} + \log \frac{1}{q} \\ &= I(p) + I(q), \end{aligned}$$

como se pretendia. Quando se toma a base 2 exprime-se a informação em *bits*.

A quantidade de informação que se transfere ao comunicar o resultado de uma experiência que pode dar dois resultados equiprováveis (logo, de probabilidade $1/2$) é, em bits,

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

Para armazenar 1 bit de informação não é sempre necessário 1 bit de memória.

[2] Como B ocorreu faz sentido assumir $P(B) \neq 0$, conforme aliás a definição sugere.

1.5 Variáveis aleatórias

Uma variável aleatória (ou simplesmente "variável", no contexto certo) faz corresponder a cada evento um número real. Isto permite substituir resultados não-numéricos do mais diverso tipo (caras e coroas, por exemplo) por números reais (0 e 1, por exemplo). É habitual designar as variáveis aleatórias por maiúsculas, como X, por exemplo.

Substituir eventos por números reais torna possível o uso de expressões como $P(X \leq 1)$ ou $P(a < X \leq b)$. Como X assume valores reais, as expressões $X \leq 1$ ou $a < X \leq b$ têm significado e determinam subconjuntos dos reais.

Uma variável que assume um conjunto finito ou contável de valores diz-se *discreta*; caso contrário, diz-se *contínua*.

1.6 Média

A média é um número que dá informação acerca da "grandeza" de um conjunto de números. Ao saber que "o aluno A tem média 15 valores" e "B tem média 19 valores", ficamos com a impressão que A tem "em geral notas mais baixas" que B — apesar de não termos ficado a saber informação específica sobre as classificações a cada disciplina. A média "resume" os dados.

Antes de definir a média de uma variável discreta X convém considerar um exemplo que mostre o significado comum do termo. Se em 5 disciplinas um aluno tiver notas $12, 18, 15, 15$ e 18, a sua média aritmética será

$$\begin{aligned} \frac{12+18+15+15+18}{5} &= \frac{12+2\times 15+2\times 18}{5} \\ &= 12\,\frac{1}{5} + 15\,\frac{2}{5} + 18\,\frac{2}{5}. \end{aligned}$$

Repare-se que $1/5$ e $2/5$ designam as frequências relativas das notas na lista de notas. Em geral, se o aluno tiver N_k notas de valor k num total de D disciplinas, a média μ será

$$\mu = \frac{1}{D}\sum_{k=0}^{20} k\,N_k,$$

ou seja,

$$\mu = \sum_{k=0}^{20} k\,f_k,$$

em que f_k é a frequência relativa da nota k, $f_k = N_k/D$.

Isto sugere, identificando as frequências relativas com probabilidades, que a definição para a média de uma variável aleatória discreta seja

$$\mu = \sum_k k\,p(k),$$

onde $p(k)$ é uma forma sucinta de escrever a probabilidade da variável X assumir o valor k, isto é, $P(X = k)$.

Também se usa a notação $E[X]$ para designar a média de X. O símbolo E pretende sugerir "valor esperado". Esta notação é conveniente para exprimir certas propriedades da média. Por exemplo, se X e Y forem variáveis aleatórias, tem-se $E[X + Y] = E[X] + E[Y]$, e para α real tem-se $E[\alpha X] = \alpha E[X]$.

1.7 Variância

A variância mede dispersão em torno do valor médio. Os conjuntos de notas $\{14, 15, 16\}$ e $\{10, 15, 20\}$ têm a mesma média aritmética, mas o primeiro conjunto é mais "concentrado" em torno da média que o segundo. O conceito de variância permite quantificar a diferença.

Para chegar à definição, considere-se um conjunto de dados x_i com média μ. Como se quer medir a homogeneidade em torno da média, examinemos a diferença de cada um dos dados para a média, $x_i - \mu$. Se os valores $x_i - \mu$ forem "pequenos" a variância deverá ser baixa. O valor absoluto da diferença, $|x_i - \mu|$, é uma medida natural do "tamanho" de $x_i - \mu$, e como tal não é afectada pelo facto da diferença ser positiva ou negativa. O mesmo acontece com o quadrado da diferença, $(x_i - \mu)^2$, que tem uma vantagem: a função "quadrado" é analiticamente mais fácil de tratar que a função "valor absoluto". Estas considerações sugerem a definição seguinte.

A variância de uma variável X, que se representa por $\sigma^2(X)$, σ_X^2, ou simplesmente σ^2, é o valor médio do quadrado dos desvios para a média, ou seja:

$$\sigma^2 = E[(X - \mu)^2].$$

Expandindo o quadrado, conclui-se que

$$\sigma^2 = E[X^2] - \mu^2,$$

ou seja, a variância é a diferença entre a média do quadrado e o quadrado da média. Esta expressão pode ser mais conveniente para efeitos de cálculo. Notar que a variância se reduz a $E[X^2]$ se a variável tiver média nula.

A raiz quadrada da variância, que se representa por σ, chama-se desvio padrão.

1.8 Soma de variáveis independentes

O quadrado de uma soma contém produtos de termos cruzados pelo que a variância de uma soma irá conter termos da forma $E[XY]$. É importante notar que

$$E[XY] = E[X]E[Y]$$

se X e Y forem independentes (notar contraste com a média da soma de duas variáveis aleatórias, que é sempre a soma das respectivas médias, mesmo que as variáveis não sejam independentes).

A média $E[X]$ é uma soma de termos

$$k\,P(X = k),$$

isto é, valores da variável e respectivas probabilidades. Da mesma forma, $E[Y]$ é uma soma de termos

$$\ell P(Y = \ell).$$

O produto $E[X]E[Y]$ será por isso uma soma de termos semelhantes a

$$k\ell\,P(X = k)\,P(Y = \ell),$$

que devido à independência de X e Y se podem escrever como

$$k\ell\,P(X = k, Y = \ell),$$

e a soma de todos estes termos é igual a $E[XY]$. Logo, $E[XY] = E[X]E[Y]$.

Podemos agora provar que a variância da soma de variáveis independentes é a soma das respectivas variâncias:

$$\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y).$$

Pode assumir-se sem perda de generalidade (basta subtrair a média) que $E[X] = E[Y] = 0$. Nesse caso, as variâncias de X, Y e $X + Y$ são simplesmente $E[X^2], E[Y^2]$ e $E[(X + Y)^2]$. Tem-se

$$\begin{aligned}
\sigma^2(X + Y) &= E[(X + Y)^2] \\
&= E[X^2 + 2XY + Y^2] && \text{(expandindo)} \\
&= E[X^2] + 2E[XY] + E[Y^2] && \text{(prop. média)} \\
&= E[X^2] + 2E[X]E[Y] + E[Y^2] && \text{(independ.)} \\
&= E[X^2] + E[Y^2] && \text{(médias nulas)} \\
&= \sigma^2(X) + \sigma^2(Y) && \text{(médias nulas)}
\end{aligned}$$

O resultado é válida para somas de mais de duas variáveis (independentes, claro).

2 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.

2.1 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.*

Imagine-se que temos à disposição uma função `rand()` que devolve um real aleatório X, pertencente ao intervalo $[0, 1]$. A função "não tem preferência" por nenhuma gama de valores, ou seja, a variável aleatória em causa tem distribuição uniforme. Isto implica que a probabilidade de $X > 0.5$ e $X < 0.5$ são iguais. Como têm de somar um, serão ambas $1/2$.

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`, podemos facilmente simular o resultado após 100 eventos:

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

Após a execução deste código, n representará o valor do contador após os 100 eventos.

O contador é na realidade uma variável aleatória, determinada por uma sucessão de experiências aleatórias. Qual é o valor médio do contador após k eventos?

Vou associar 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

$$E[X_i] \;=\; 0\times P[X_i=0]+1\times P[X_i=1] = P[X_i=1] = \frac{1}{2}.$$

O valor do contador após k eventos é a soma dos k incrementos,

$$S = X_1 + X_2 + \cdots + X_k,$$

cujo valor médio é

$$\begin{aligned} E[S] &= E[X_1 + X_2 + \cdots + X_k] \\ &= E[X_1] + E[X_2] + \cdots + E[X_k] \\ &= \underbrace{\frac{1}{2} + \cdots + \frac{1}{2}}_{k \text{ termos}} = \frac{k}{2}. \end{aligned}$$

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.*

A variância de um qualquer dos X_i é

$$\sigma^2(X_i) = E[X_i^2] - (E[X_i])^2 = E[X_i^2] - \frac{1}{4}.$$

Como

$$E[X_i^2] \;=\; 0^2 \times P[X_i=0] + 1^2 \times P[X_i=1] = \frac{1}{2},$$

vem

$$\sigma^2(X_i) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.$$

Como as variáveis X_i são independentes, a variância de S é

$$\begin{aligned} \sigma^2(S) &= \sigma^2(X_1 + X_2 + \cdots + X_k) \\ &= \sigma^2(X_1) + \sigma^2(X_2) + \cdots + \sigma^2(X_k) \\ &= \underbrace{\frac{1}{4} + \cdots + \frac{1}{4}}_{k \text{ termos}} = \frac{k}{4}, \end{aligned}$$

o que significa $\sigma = \sqrt{k}/2$. Para $k = 10\,000$ eventos, a média é 5 000 e o desvio padrão 50.

2.2 Distribuição de probabilidade

Pode calcular-se a probabilidade de, após k eventos, o valor do contador ser n. Considere-se o exemplo $k = 4$, e sejam X_1, X_2, X_3 e X_4 as variáveis binárias que descrevem se o contador é incrementado ou não após o evento 1, 2, 3 e 4. Há 16 possibilidades:

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:

$$p(0)=\frac{1}{16},\;\; p(1)=\frac{4}{16},\;\; p(2)=\frac{6}{16},\;\; p(3)=\frac{4}{16},\;\; p(4)=\frac{1}{16}.$$

Quando a probabilidade de incrementar o contador é p e a probabilidade de o não fazer é $1-p$, e probabilidade de observar uma soma igual a n após k experiências é a seguinte:

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

2.3 Trabalho a realizar

2.3.1. Faça um programa para simulação da contagem. Faça várias simulações e calcule a média das contagens obtidas. Compare essa média com $E[S]$.

2.3.2. Mostre que a média da soma de duas variáveis aleatórias é a soma das respectivas médias.

2.3.3. Discutir a expressão *incrementar o contador metade das vezes.*

2.3.4. Qual a distribuição de probabilidades de S? Isto é, qual a probabilidade de S ser zero, um, etc., que se pode representar por $p(k)$?

2.3.5. Qual a variância de S?

2.3.6. Calcule também a variância das estimativas obtidas pelo programa e compare com o obtido em 2.3.5.

2.3.7. Se ocorrer *overflow* no contador haverá diferenças a registar?

2.4 Uma variante

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 1/2. As variáveis X_i seriam agora definidas da seguinte maneira:

$$X_i = \begin{cases} 1, & \text{com probabilidade } 1/64, \\ 0, & \text{com probabilidade } 63/64. \end{cases}$$

O valor médio de cada incremento X_i é agora

$$\begin{aligned} E[X_i] &= 0 \times P[X_i{=}0] + 1 \times P[X_i{=}1] \\ &= 1 \times P[X_i{=}1] = \frac{1}{64}. \end{aligned}$$

Logo, o valor médio do contador após k eventos, representado pela soma dos k incrementos,

$$S = X_1 + X_2 + \cdots + X_k,$$

será dado por

$$E[S] = E[X_1] + \cdots + E[X_k] = \frac{1}{64} + \cdots + \frac{1}{64} = \frac{k}{64}.$$

Neste caso, o número de eventos pode ser estimado por $64n$, onde n é o valor do contador.

2.5 Trabalho a realizar

2.5.1. O que se ganha e o que se perde quando se usa 64 em vez de 2?

2.5.2. Qual a variância de S?

2.5.3. Faça um programa para simulação da contagem. Simule várias vezes e calcule a média das estimativas obtidas. Compare essa média com $E[S]$.

2.5.4. Calcule também a variância das estimativas obtidas pelo programa e compare com o obtido em 2.5.2.

2.5.5. Qual a distribuição de S?

2.6 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} (ver Fig. 1). Logo, o incremento *médio* é 2^{-n}. São por isso necessários 2^n eventos para que o valor médio da contagem suba uma unidade, como na sequência seguinte:

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

Repare-se que $2^n - 1$ parece ser uma estimativa adequada para o número de eventos. Na verdade, o valor médio de 2^n, sendo n a leitura obtida do contador após k eventos, é $k + 1$, o que corrobora essa impressão.

Só há duas formas do contador apresentar o valor n após k eventos:

- Ter o valor $n - 1$ após $k - 1$ eventos e ser incrementado.
- Ter o valor n após $k - 1$ eventos e não ser incrementado.

Estas duas alternativas traduzem-se na seguinte equação, em que P_k^n é a probabilidade do contador apresentar o valor n após k eventos:

$$\boxed{P_k^n = P_{k-1}^{n-1}\, 2^{-(n-1)} + P_{k-1}^n\, (1 - 2^{-n}).}$$

A equação menciona o evento $k-1$, pelo que é natural assumir $k \geq 1$. O contador passa a conter 1 após o primeiro evento, ou seja, $P_1^1 = 1$ e $P_1^0 = 0$ (na verdade, $P_k^0 = 0$, $k \geq 1$). Tem-se ainda

$$\sum_{n=1}^{k} P_k^n = 1$$

para qualquer $k \geq 1$, porque após k eventos o valor do contador tem de estar compreendido entre 0 e k, inclusive.

Como o valor n do contador após k eventos é uma variável aleatória, 2^n também o é. A sua média pode calcular-se da seguinte forma:

$$\begin{aligned} \sum_{n=1}^{k} 2^n P_k^n &= \sum_{n=1}^{k} 2^n\, P_{k-1}^{n-1}\, 2^{-(n-1)} + \sum_{n=1}^{k} 2^n\, P_{k-1}^n\, (1 - 2^{-n}) \\ &= 2 \sum_{n=1}^{k} P_{k-1}^{n-1} + \sum_{n=1}^{k} P_{k-1}^n\, (2^n - 1) \\ &= 2 + \left(\sum_{n=1}^{k-1} 2^n\, P_{k-1}^n \right) - 1 \\ &= 1 + \sum_{n=1}^{k-1} 2^n\, P_{k-1}^n. \end{aligned}$$

Logo, o valor médio de 2^n após k eventos é igual ao seu valor médio após $k-1$ eventos, acrescido de uma unidade; ou igual ao valor médio após $k-2$ eventos, acrescido de duas unidades; e assim sucessivamente até ao valor médio após um evento, acrescido de $k-1$ unidades. O valor do contador após o primeiro evento é $n = 1$, pelo que o valor médio de 2^n após um evento é 2. Logo,

$$\sum_{n=1}^{k} 2^n P_k^n = k + 1.$$

Após k eventos, o valor médio de 2^n, em que n é o valor lido no contador, é $k + 1$. Isto confirma o que a tabela acima sugeria: $2^n - 1$ é uma estimativa para k, o número de eventos.

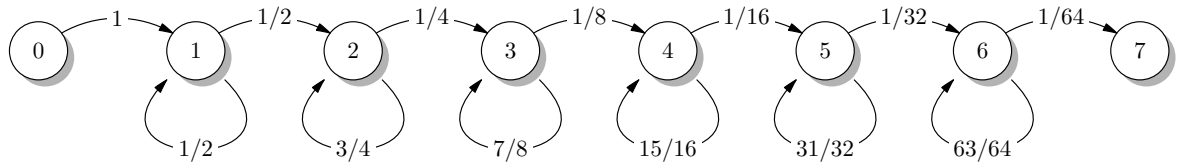


Figura 1: O contador é incrementado com probabilidade cada vez menor à medida que o seu valor aumenta. Neste caso, quando o valor do contador é n, a probabilidade de um incremento é 2^{-n}.

O valor médio do quadrado de 2^n após k eventos é dado por:

$$\begin{aligned}\sum_{n=1}^{k} 2^{2n} P_k^n &= \sum_{n=1}^{k} 2^{n+1} P_{k-1}^{n-1} + \sum_{n=1}^{k} P_{k-1}^n (2^{2n} - 2^n) \\ &= \sum_{n=1}^{k} 2^{n+1} P_{k-1}^{n-1} + \sum_{n=1}^{k} 2^{2n} P_{k-1}^n \\ &\quad - \sum_{n=1}^{k} 2^n P_{k-1}^n. \qquad (1)\end{aligned}$$

No primeiro termo, substitua-se $u = n - 1$:

$$\begin{aligned}\sum_{n=1}^{k} 2^{n+1} P_{k-1}^{n-1} &= \sum_{u=0}^{k-1} 2^{u+2} P_{k-1}^u \\ &= 4 \sum_{u=1}^{k-1} 2^u P_{k-1}^u = 4k,\end{aligned}$$

invocando o resultado anterior para o valor médio. O terceiro termo simplifica-se de forma semelhante:

$$\sum_{n=1}^{k} 2^n P_{k-1}^n = \sum_{n=1}^{k-1} 2^n P_{k-1}^n = k,$$

usando de novo o resultado para o valor médio. Voltando à equação (1),

$$\sum_{n=1}^{k} 2^{2n} P_k^n = 3k + \sum_{n=1}^{k-1} 2^{2n} P_{k-1}^n.$$

Significa esta expressão que o valor médio do quadrado de 2^n após k eventos é igual ao valor médio do mesmo quadrado após $k-1$ eventos, acrescido de $3k$ unidades. Reparar na estrutura da equação:

$$\begin{aligned}F(k) &= 3k + F(k-1) \\ F(k-1) &= 3(k-1) + F(k-2) \\ F(k-2) &= 3(k-2) + F(k-3) \\ &\vdots \quad \vdots \\ F(2) &= 3 \cdot 2 + F(1).\end{aligned}$$

É agora fácil chegar à conclusão

$$\sum_{n=1}^{k} 2^{2n} P_k^n = \frac{3k(k+1)}{2} + 1.$$

Para obter a variância falta apenas subtrair o quadrado da média, média essa obtida anteriormente:

$$\sigma^2 = \frac{3k(k+1)}{2} + 1 - (k+1)^2 = \frac{k(k-1)}{2}.$$

2.7 Trabalho a realizar

2.7.1. Faça um programa para simulação da contagem. Faça várias simulações e calcule a média de $2^n - 1$, sendo n o valor lido do contador. Compare com o valor teórico.

2.7.2. Até que valor consegue contar com contadores de 4 bits? E com contadores de 8 bits?

3 Pertença e cardinalidade

Considere-se a seguinte pergunta: o elemento s (uma cadeia de bits de tamanho arbitrário) pertence ou não a um dado conjunto S? A resposta é fácil quando o conjunto é "pequeno". Para conjuntos enormes *(big data)*, há obstáculos significativos.

3.1 Numeração *(hashing)*

Interpretando uma cadeia de bits s como um inteiro, e esse inteiro como um índice de um vector v, tem-se uma método simples de pesquisa. Primeiro, colocam-se todos os elementos de v a zero, e para cada s pertencente ao conjunto faz-se $v[s] = 1$. Depois, para verificar se uma dada cadeia de bits t pertence ao conjunto, interpreta-se a mesma como um inteiro, e verifica-se o correspondente elemento do vector, $v[t]$. O elemento t pertencerá ao conjunto se e só se $v[t] = 1$.

Não há dificuldades quando os elementos de S se podem representar com poucos bits (por exemplo, se tiverem 20 bits basta usar um vector com cerca de um milhão de bits).

Contudo, quando os elementos de S têm extensões arbitrárias, convém considerar um passo intermédio: uma função *(hashing function)* que *numera* qualquer cadeia que lhe seja entregue, ou seja, calcula a partir dessa cadeia um inteiro de *tamanho fixo*. Um exemplo[3] (DJB31MA):

```
uint hash(const uchar* s, int len, uint seed)
{
    uint h = seed;
    for (int i=0; i < len; ++i)
        h = 31 * h + s[i];
    return h;
}
```

Cada elemento do conjunto S pode ser submetido a esta função e assim associado a um inteiro; esse inteiro

[3] Na verdade, é uma família de exemplos, uma vez que o argumento `seed` pode variar.

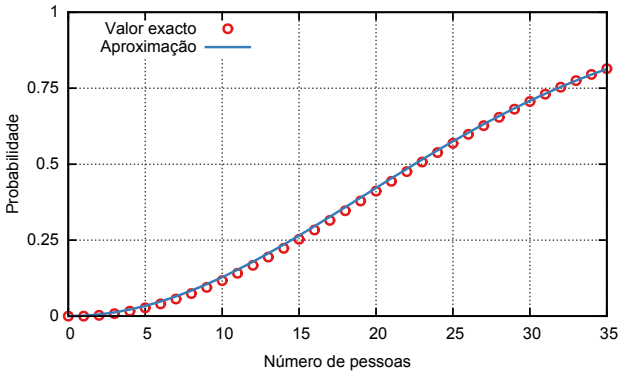


Figura 2: Probabilidade de num grupo de m pessoas existirem pelo menos duas com o mesmo aniversário, para vários valores de m. A probabilidade excede $1/2$ para $m \geq 23$.

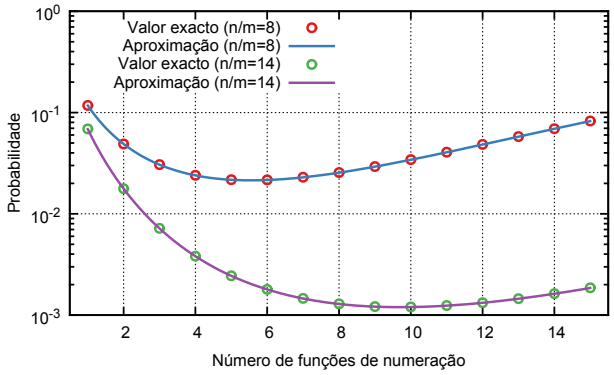


Figura 3: Probabilidade de um erro (falso positivo) em função do número de funções de numeração.

pode ser depois mapeado num vector com um número apropriado de elementos, n. O método é flexível, mas há uma possibilidade nova a considerar.

Quando dois elementos distintos são mapeados na mesma posição do vector, diz-se que há uma *colisão*. A análise simplifica-se assumindo que o mapeamento é realizado de forma aleatória, com uma distribuição uniforme (todos os resultados são igualmente prováveis).

A probabilidade de não haver colisões num vector de tamanho n depois de inserir m elementos é

$$\begin{aligned} p &= \frac{n-1}{n} \times \frac{n-2}{n} \times \cdots \times \frac{n-(m-1)}{n} \\ &= \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{m-1}{n}\right) \\ &\approx e^{-1/n} e^{-2/n} \cdots e^{-(m-1)/n} \\ &= e^{-m(m-1)/2n} \approx e^{-m^2/2n}. \end{aligned}$$

A probabilidade de haver pelo menos uma colisão é

$$q \approx 1 - e^{-m^2/2n}.$$

Tem-se $q = 1/2$ quando

$$\frac{m^2}{2n} = \log 2 \approx 0.693,$$

ou seja

$$m \approx \sqrt{1.386\, n} \approx 1.177\sqrt{n}.$$

Exemplo (Fig. 2): para $n = 365$, a probabilidade de colisão é pelo menos $1/2$ para $m > 22.49$ (o paradoxo do aniversário).

3.2 Trabalho opcional

3.2.1. Faça um programa para falsificar assinaturas digitais baseadas na função de numeração dada, para 32 bits.

3.3 Filtro de Bloom *(Bloom filter)*

Um filtro de Bloom usa k funções de numeração sobre o mesmo vector de bits. Assume-se que qualquer das funções de numeração produz n resultados uniformemente distribuídos, mutuamente independentes.

Seja b_i o valor do bit i. Inicialmente, todos os bits são colocados a zero. Para inserir o primeiro elemento no filtro, usam-se as k funções de numeração, que determinam que bits têm de ser colocados a 1. A probabilidade de $b_i = 1$ após usar a primeira função de numeração é $1/n$, pelo que a probabilidade de $b_i = 0$, nas mesmas condições, é $1-1/n$. A operação tem de ser repetida um total de k vezes, uma vez que há k funções de numeração. Finalmente, uma vez colocado o primeiro elemento no filtro, o processo tem de ser repetido para colocar os restantes elementos no filtro.

Como há k funções de numeração e m elementos para colocar, a probabilidade de $b_i = 0$ após processamento dos m elementos é (assumindo independência)

$$\left(1-\frac{1}{n}\right)^{km} = \Bigg[\underbrace{\left(1-\frac{1}{n}\right)^{m}}_{a}\Bigg]^{k} = a^k.$$

A probabilidade de $b_i = 1$, nas mesmas condições, é

$$1 - \left(1-\frac{1}{n}\right)^{km} = 1 - a^k.$$

A probabilidade de um erro (falso positivo) após a inserção de m elementos é a probabilidade de encontrar k bits a um, ou seja,

$$p = \left[1-\left(1-\frac{1}{n}\right)^{km}\right]^{k} = \left(1-a^k\right)^k. \qquad (2)$$

Para determinar o valor de k que minimiza a probabilidade de erro p (ver Fig. 3) é preferível minimizar $\log p$, que tem estrutura mais favorável:

$$\log p = k \log(1-a^k).$$

Derivando e igualando a zero vem

$$(1-a^k)\log(1-a^k) - a^k \log a^k = 0,$$

cuja solução é $a^k = 1/2$. O melhor valor de k seria, portanto,

$$k_{\text{opt}} = \frac{\log 1/2}{\log a} = \frac{\log 1/2}{m \log(1-1/n)} \approx \frac{n \log 2}{m} \approx \frac{0.693\, n}{m}.$$

Contudo, como este valor não é inteiro, terá de se usar o inteiro mais próximo. Se fosse possível ter $a^k = 1/2$ para k inteiro, a equação (2) conduziria a

$$p_{\text{opt}} = 2^{-k_{\text{opt}}},$$

que se pode tomar como limite inferior para a probabilidade de erro.

3.4 Estimação de frequências

A substituição de cada bit de um filtro de Bloom por um contador permite estimar o número de ocorrências de cada elemento do conjunto. O algoritmo é uma extensão óbvia do anterior: usam-se as k funções de numeração para determinar outros tantos contadores, e incrementam-se os mesmos por uma unidade. Para estimar a frequência de um dado símbolo, usam-se as mesmas k funções de numeração e toma-se o mínimo das contagens encontradas (como é possível haver colisões, o menor dos valores é sempre o menos afectado por erros).

3.5 Trabalho a realizar

3.5.1. Implemente um *Bloom filter* (para testes de pertença ou para contagem, à sua escolha).

3.5.2. Há uma descrição mais precisa do trabalho e um exemplo em Java em `e-learning.ua.pt`, mas a linguagem de programação fica à sua escolha.

4 Minhash e aplicações

4.1 Detecção de duplicados

É frequentemente necessário identificar elementos duplicados (ou muito parecidos) num conjunto de objectos (por exemplo livros, páginas de web ou fotografias digitais). A solução óbvia é comparar todos os pares de objectos:

```
for (i=0; i < n; ++i)
  for (j=0; j < i; ++j)
     if (obj(i) == obj(j))
       Print obj(i) e obj(j) são duplicados
```

Considere-se que há um milhão de objectos para comparar, isto é, $n = 10^6$. É fácil ver que teríamos de fazer cerca de 10^{12} comparações (tantas quanto os pares distintos de objectos). Se pudermos comparar 1 milhão de objectos por segundo (recordemos que os objectos podem ser livros ou imagens), o tempo necessário seria cerca de 10^6 segundos, o que é mais de 10 dias. Por outro lado, se o número de objectos duplicar, o tempo necessário quadruplica. Esta solução não é viável, pelo que é necessário procurar outra abordagem.

4.2 Minhash

Para isso repare-se, em primeiro lugar, que é conveniente caracterizar o conteúdo do objecto (livro, página de web, fotografia digital, música digital, etc.) de forma o mais simples possível. Depois de identificar os elementos $s_1, s_2, \ldots, s_n$ constituintes de um qualquer objecto A, é vantajoso substituí-los por $h(s_1), h(s_2), \ldots, h(s_n)$, onde h é uma função de numeração. Para abreviar, usarei a notação $h(A) := \{h(s_1), h(s_2), \ldots, h(s_n)\}$.

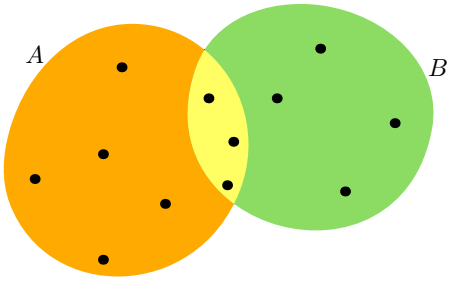


Figura 4: $|A \cap B| = 3$, $|A \cup B| = 12$, logo o índice de Jaccard de A e B é $J(A, B) = 3/12$.

Em segundo lugar, é necessário decidir como medir a semelhança entre objectos. Usarei o *índice de Jaccard*:

$$J(A, B) := \frac{|A \cap B|}{|A \cup B|}.$$

A notação $|A|$ designa o número de elementos de A. O índice de Jaccard assume valores no intervalo $[0, 1]$, tendo-se $J(A, B) = 0$ para conjuntos disjuntos e $J(A, B) = 1$ para conjuntos iguais (Fig. 4).

Em terceiro lugar, não é necessário comparar todos os números $h(A)$ e $h(B)$ para ter uma indicação acerca da semelhança de A e B. Basta comparar os dois menores, que se chamam "minhash". Por exemplo, o minhash de A é

$$\min h(A) = \min\{h(s_1), h(s_2), \ldots, h(s_n)\}.$$

O facto importante é o seguinte: a probabilidade do minhash de A ser igual ao de B é igual ao índice de Jaccard de A e B:

$$P[\min\, h(A) = \min\, h(B)\,] \quad = \quad J(A, B).$$

A demonstração baseia-se no seguinte argumento: qualquer elemento da união de A e B tem igual probabilidade de ser o elemento de valor mínimo, e a probabilidade pretendida aumenta proporcionalmente com o tamanho de $A \cap B$.

4.3 Matriz de assinaturas

Considere-se agora a *matriz de assinaturas* dos objectos A_i, para as funções de numeração $h_1, h_2, \ldots, h_k$:

	A_1	A_2	$\cdots$	A_n
h_1	$\min h_1(A_1)$	$\min h_1(A_2)$	$\cdots$	$\min h_1(A_n)$
h_2	$\min h_2(A_1)$	$\min h_2(A_2)$	$\cdots$	$\min h_2(A_n)$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
h_k	$\min h_k(A_1)$	$\min h_k(A_2)$	$\cdots$	$\min h_k(A_n)$

Suponhamos que A_1 e A_2 são objectos muito parecidos, com índice de Jaccard $J(A_1, A_2) = 0.9$. Qual a probabilidade de todos os minhash de A_1 e A_2 serem iguais? A probabilidade pretendida é 0.9^k, e tende para zero

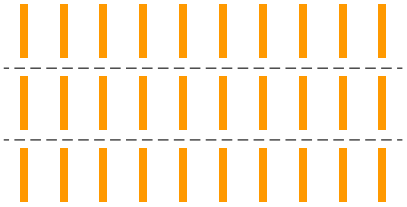
à medida que k aumenta. Comparar colunas completas da matriz de assinaturas não conduz a resultados interessantes.

4.4 LSH

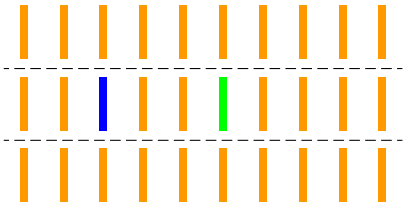
Há vantagem em dividir a matriz de assinaturas em faixas (*bands*). O exemplo seguinte, para três objectos A_1, A_2 e A_3, apresenta três faixas, cada uma com duas linhas:

	A_1	A_2	A_3
h_1	$\min h_1(A_1)$	$\min h_1(A_2)$	$\min h_1(A_3)$
h_2	$\min h_2(A_1)$	$\min h_2(A_2)$	$\min h_2(A_3)$
h_3	$\min h_3(A_1)$	$\min h_3(A_2)$	$\min h_3(A_3)$
h_4	$\min h_4(A_1)$	$\min h_4(A_2)$	$\min h_4(A_3)$
h_5	$\min h_5(A_1)$	$\min h_5(A_2)$	$\min h_5(A_3)$
h_6	$\min h_6(A_1)$	$\min h_6(A_2)$	$\min h_6(A_3)$

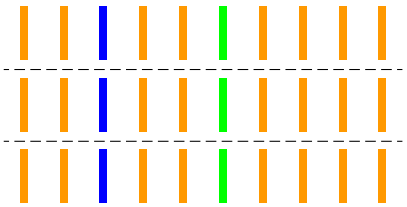
É conveniente representar as faixas e os blocos por faixa da forma abreviada que se segue (o exemplo refere-se a 10 objectos, 3 faixas):



Considerem-se os dois blocos assinalados na figura seguinte:



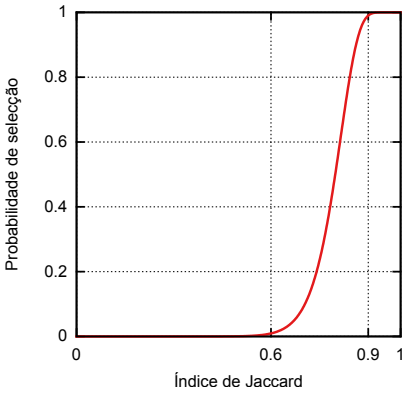
A probabilidade de todos os elementos do bloco azul e verde serem iguais é J^r, onde J é o índice de Jaccard para os dois objectos em causa e r é o número de linhas em cada faixa. A probabilidade disso não acontecer é $1 - J^r$. A probabilidade disso não suceder em nenhuma faixa, conforme as faixas assinaladas na figura seguinte



é dada por $(1 - J^r)^b$, onde b é o número de bandas, ou faixas. Logo, a probabilidade de haver pelo menos uma faixa onde todos os elementos do bloco azul e verde são iguais é

$$P = 1 - (1 - J^r)^b.$$

Esta equação é a base de um algoritmo rápido para detecção de objectos semelhantes (LSH, ou *locally sensitive hashing*). Consideram-se objectos próximos todos



aqueles para os quais se verifica a igualdade de todos os minhash numa faixa (pelo menos). O princípio não é infalível (há falsos positivos e falsos negativos) mas a probabilidade desses falsos positivos e negativos pode controlar-se escolhendo b e r adequadamente.

Imagine-se que se pretende que objectos com índice de Jaccard 60% ou inferior sejam seleccionados com probabilidade inferior a 0.01, mas que objectos com índice de Jaccard superior a 90% sejam seleccionados com probabilidade 0.99. É possível calcular r e b de forma a que isto se verifique, obtendo-se $b \approx 20$ e $r \approx 15$. A forma da curva $P(J) = 1 - (1 - J^r)^b$ para estes valores de b e r é apresentada na Fig. 4.4.

4.5 Trabalho opcional

4.5.1. Obtenha o ficheiro `ratings.dat` do conjunto de dados "MovieLens 10M dataset" (google: movielens dataset). O ficheiro tem 10 000 054 linhas. Cada linha tem um número que identifica um utilizador (há 69 878 utilizadores) e outro número que identifica um filme classificado pelo utilizador (há 10 677 filmes).

4.5.2. Há utilizadores com conjuntos exactamente iguais de filmes. Ache possíveis candidatos em menos de um segundo. Verifique se são de facto duplicados ou não.