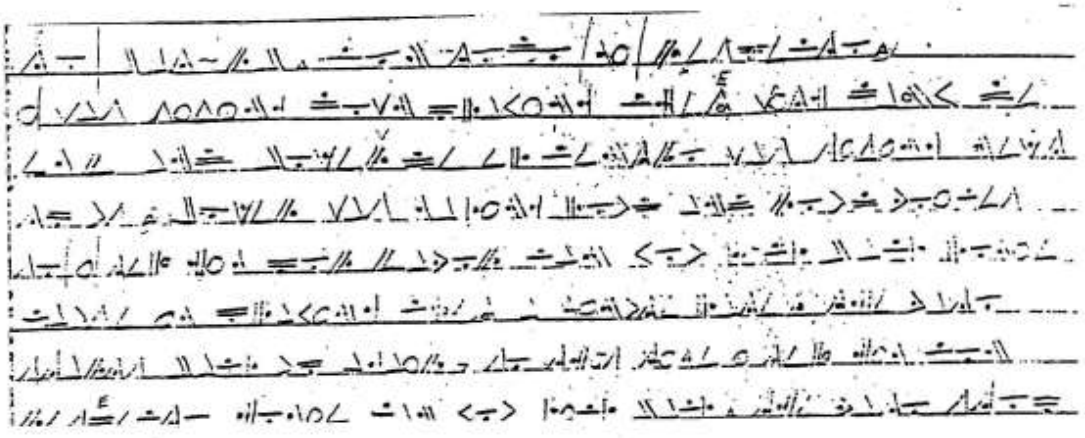
Carta interseptada pelos guardas da Prisão Estadual da Califórnia e enviada ao departamento de psicologia da Universidade de Stanford.

Créditos: Monte Carlo Markov Chain Revolution - Persi Diaconis



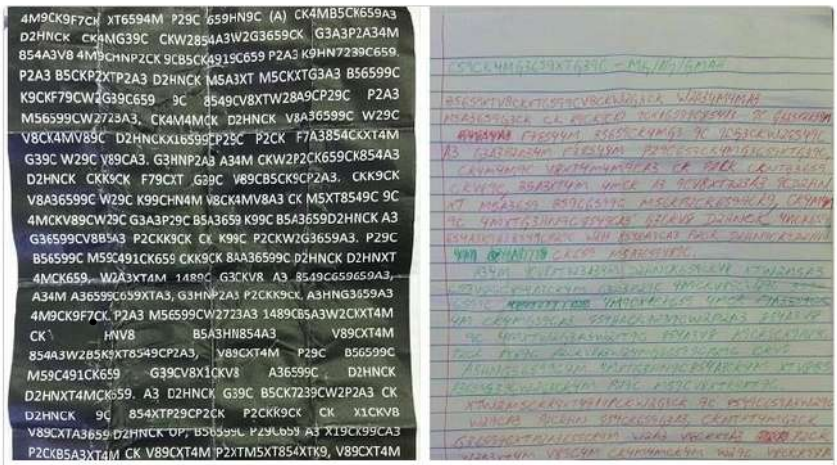
Assuma:

- encriptação por substuição, isto é, cada símbolo representa uma letra do alfabeto latino.

- Alfabeto com 40 caracteres (pontuação, espaço etc).

Pergunta-se: COMO QUEBRAR ESSE CÓDIGO?

Outros exemplos de cartas...



Cartas com supostas ameaças ao promotor de Justiça têm código secreto do PCC
DIVULGAÇÃO

A - 9C
B - X1
C - 854
D - P2
E - CK
F - M5
G - 723
H - 8A
I - XT
J - 148
L - K9
M - V8
N - W2
O - A3
P - B5
Q - D2
R - 659
S - 4M
T - G3
U - HN
V - F7
W - OP
X - NT
Z - 491

Como desencriptar as cartas?

Tentativa:

Análise de frequência de aparecimentos de letras em textos de uma língua.

Problemas:

- Que língua? Inglês?

- Que tipo de textos? Poemas? Verbetes de dicionários?

- Baixa amostragem para inferência.

- Qual o modelo probabilistico usado? Processo de Bernoulli?

Letter	Relative frequency in the English language	
	Texts	Dictionaries
A	8.2%	7.8%
B	1.5%	2%
C	2.8%	4%
D	4.3%	3.8%
E	13%	11%
F	2.2%	1.4%
G	2%	3%
H	6.1%	2.3%
I	7%	8.6%
J	0.15%	0.21%
K	0.77%	0.97%
L	4%	5.3%
M	2.4%	2.7%
N	6.7%	7.2%
O	7.5%	6.1%
P	1.9%	2.8%
Q	0.095%	0.19%
R	6%	7.3%
S	6.3%	8.7%
T	9.1%	6.7%
U	2.8%	3.3%
V	0.98%	1%
W	2.4%	0.91%
X	0.15%	0.27%
Y	2%	1.6%
Z	0.074%	0.44%

Fonte: Wikipedia

Modelando a linguagem usando probabilidade

1948 - Claude Shannon - Mathematical Theory of Communication

Introdução do modelo de cadeias de Markov para textos.

O que é uma cadeia de Markov?

Resposta: Uma sequência $\{X_0, X_1, X_2, \dots\}$ de variáveis aleatórias com valores em um conjunto finito $\{s_1, s_2, \dots, s_K\}$ (chamados de estados) tal que

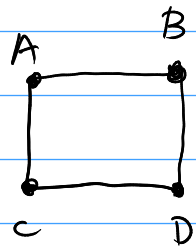
$$P(X_{n+1} = s_j \mid X_0 = s_{i_0}, X_1 = s_{i_1}, \dots, X_n = s_i) = P(X_{n+1} = s_j \mid X_n = s_i)$$

$P = (p_{ij})_{K \times K}$ é o principal ingrediente de uma cadeia de Markov.

Note que: (a) $p_{ij} \geq 0$ (b) $\sum_{j=1}^{K} p_{ij} = 1$ (P é estocástica)

Exemplo [Passeio aleatório no □] $S = \{A, B, C, D\}$

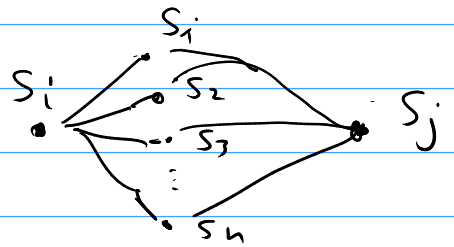
X_0 uniforme em S, P dada por:



$$\begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \end{array} = P$$

Passeio simulado: A B A C D B A B D C D C A ...

O que P^2 representa?



$$p_{ij}^2 = \sum_{\ell=1}^{K} p_{i\ell}\, p_{\ell j} =$$

$$= \sum_{\ell=1}^{K} P(X_{n+1} = s_\ell \mid X_n = s_i) \cdot P(X_{n+2} = s_j \mid X_{n+1} = s_\ell)$$

p_{ij}^2 é a prob. de estarmos em s_j no "tempo" $n+2$ dado que estamos em s_i no "tempo" n. Formalmente:

Teorema: Para uma cadeia de Markov (X_0, X_1,...) Com espaço de estado {s_1,. . . , s_k}, distribuição inicial μ_0 e matriz de transição P, temos para qualquer n que a distribuição μ_n no tempo n satisfaz μ_n = μ_0 P^n.

Prova: Segue por indução com o argumento anterior, com pequenas modificações.

Exemplo 1 simulado: $\mu_0 = (p_A^0, p_B^0, p_C^0, p_D^0)$, $p_A + p_B + p_C + p_D = 1$

$\mu_n = (p_A^n, p_B^n, p_C^n, p_D^n) = \mu_0 P^n$. Quem é P^n?

$$P^n \to \theta$$

```
P^20:
[[0.27777958 0.22222042 0.33333325 0.16666675]
 [0.27777934 0.22222066 0.33333325 0.16666675]
 [0.27777433 0.22222567 0.33333349 0.16666651]
 [0.27777958 0.22222042 0.33333325 0.16666675]]
```

$$\theta = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_k \\ \pi_1 & \pi_2 & & \pi_k \\ \vdots & & & \\ \pi_1 & \pi_2 & \cdots & \pi_k \end{pmatrix}$$

$\mu_0 P^n \to \mu_0 \theta = (\pi_1, \pi_2, \ldots, \pi_n) = \pi$

- NÃO IMPORTA o μ_0 inicial!!!
- A convergência é exponencial

Exemplo 2 (Linguagem):

→ (2A) Gerador aleatório de palavras razoáveis

(2B) Gerador aleatório de textos razoáveis

(2C) Análise sintática de textos (verbos, adjetivos, ...)

(2D) Análise Semântica de Homônimos (cadeias de Markov escondidas)

(2E) Corretor ortográfico

(2F) Corretor de endereços

GERADOR BERNOULLI NO PAPER DE CLAUDE SHANNON EM 1948:

REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

GERADOR MARKOVIANO NO PAPER DE CLAUDE SHANNON EM 1948:

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

Teorema Fundamental das Cadeias de Markov

Def: Uma cadeia de Markov é dita REGULAR se existe $L \in \mathbb{N}$ tal que $P^L > 0$, i.e., $p_{ij}^L > 0$, $\forall i,j$.

Ex 1 é regular.

Teorema: Se P é uma cadeia de Markov regular, então existe um vetor de probabilidades $\tilde{\pi}$ tal que dado um vetor de probabilidades μ_0, então $\mu_0 P^n \to \tilde{\pi}$.

Equivalentemente, $P^n \to \begin{pmatrix} \tilde{\pi} \\ \tilde{\pi} \\ \vdots \\ \tilde{\pi} \end{pmatrix} = W_{k \times k}$ matriz com k linhas $= \tilde{\pi}$.

Além disso, a convergência acima é exponencial, i.e.,

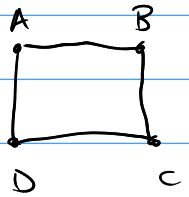
$\exists C, 0 < \lambda < 1$ t.q. $\|P^n - W\| \leq C\lambda^n$, $\forall n \geq 0$

Ex3: Passeio aleatório em um grafo $G=(V,A)$

Estados = Vértices

$$P_{ij} = \begin{cases} \frac{1}{d_i}, & (i,j) \in A \\ 0, & (i,j) \notin A \end{cases}$$

Ex1:



Obs1: Muitos exemplos interessantes são modelados por passeio aleatório em grafos.

obs2: O grafo pode ser bem grande

obs3: No problema inicial de descriptação da carta dos prisioneiros, podemos considerar o seguinte modelo:

$$V = \{f;\ f: S \longrightarrow A \text{ permutação}\}$$

A = alfabeto
S = símbolos

$$\#V \cong 40!$$

$$A = \{(f_1, f_2);\ f_1 = f_2 \circ \sigma,\ \sigma \text{ transposição}\}$$

$\longrightarrow$ $G=(V,A)$ $\quad d(v) = \binom{40}{2}$

$\longrightarrow$ G tem diametro pequeno, i.e., existe um caminho

$$f_{i_0} = f_i \to f_{i_1} \to \cdots \to f_{i_{100}} = f_j$$

$s_1 s_2 s_3 \ldots s_t \longrightarrow$ mensagem codificada.

$$P: V \longrightarrow \mathbb{R}$$

$$f \longrightarrow P(f) = \prod_{\ell=1}^{t} M\left(f(s_\ell), f(s_{\ell+1})\right)$$

$M(\alpha,\beta)$ = prob. de aparecer β dado que aparecem α na língua (inglesa?)

M é calculado usando uma base textual externa.

Método para maximizar P (MCMC)

(1) Escolha f inicial qualquer

(2) Calcule $P(f)$

(3) Escolha f^* vizinho de f em G

(4) Calcule $P(f^*)$. Se $\frac{P(f^*)}{P(f)} > 1$, aceite f^*

(5) Caso contrário, jogue uma moeda com chance $p = \frac{P(f^*)}{P(f)}$ de cara. Se der cara, aceite f^*

(6) Se der coroa, fique com f.

SIMULAÇÃO: Mensagem original

```
ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS
NOBLER IN THE MIND TO SUFFER THE SLINGS AND ARROWS OF OUTRAGEOUS
FORTUNE OR TO TAKE ARMS AGAINST A SEA OF TROUBLES AND BY OPPOSING END
```

É aplicada uma permutação e o algoritmo é rodado.

Resultado:

```
 100 ER ENOHDLAE OHDLO UOZEOUNORU O UOZEO HD OITO HEOQSET IUROFHE HENO ITORUZAEN
 200 ES ELOHRNDE OHRNO UOVEOULOSU O UOVEO HR OITO HEOQAET IUSOPHE HELO ITOSUVDEL
 300 ES ELOHANDE OHANO UOVEOULOSU O UOVEO HA OITO HEOQRET IUSOFHE HELO ITOSUVDEL
 400 ES ELOHINME OHINO UOVEOULOSU O UOVEO HI OATO HEOQRET AUSOWHE HELO ATOSUVMEL
 500 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 600 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 900 ES ELOHANME OHANO UODEOULOSU O UODEO HA OITO HEOQRET IUSOWHE HELO ITOSUDMEL
1000 IS ILOHANMI OHANO RODIORLOSR O RODIO HA OETO HIOQUIT ERSOWHI HILO ETOSRDMIL
1100 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1200 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1300 ISTILOHARMITOHAROT ODIO LOS TOT ODIOTHATOENOTHIOQUINTE SOWHITHILOTENOS DMIL
1400 ISTILOHAMRITOHAMOT OFIO LOS TOT OFIOTHATOENOTHIOQUINTE SOWHITHILOTENOS FRIL
1600 ESTEL HAMRET HAM TO CE OL SOT TO CE THAT IN THE QUENTIOS WHETHEL TIN SOCREL
1700 ESTEL HAMRET HAM TO BE OL SOT TO BE THAT IN THE QUENTIOS WHETHEL TIN SOBREL
1800 ESTER HAMLET HAM TO BE OR SOT TO BE THAT IN THE QUENTIOS WHETHER TIN SOBLER
1900 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
2000 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
```

No problema original (após apertar alguns parafusos...)

to bat-rb. con todo mi respeto. i was sitting down playing chess with danny de emf and boxer de el centro was sitting next to us. boxer was making loud and loud voices so i tell him por favor can you kick back homie cause im playing chess a minute later the vato starts back up again so this time i tell him con respecto homie can you kick back. the vato stop for a minute and he starts up again so i tell him check this out shut the f**k up cause im tired of your voice and if you got a problem with it we can go to celda and handle it. i really felt disrespected thats why i told him. anyways after i tell him that the next thing I know that vato slashes me and leaves. dy the time i figure im hit i try to get away but the c.o. is walking in my direction and he gets me right dy a celda. so i go to the hole. when im in the hole my home boys hit doxer so now "b" is also in the hole. while im in the hole im getting schoold wrong and

Caso Geral: Método de Monte Carlo Markov Chain (MCMC)

Objetivo: simular um processo de Bernoulli Y com estados $\{s_1, \dots, s_K\}$ e distribuição $\pi = (\pi_1, \pi_2, \dots, \pi_K)$.

$K=2$ Para cada n, tome $x \in [0,1]$ com prob. uniforme.

Se $x \in [0, \pi_1] \rightarrow Y(x) = s_1$

$x \in (\pi_2, 1] \rightarrow Y(x) = s_2$

O mesmo é viável para k pequeno. Contudo, se $K=40!$ não dá para guardar na memória do PC toda informação.

SAÍDA: Simular uma cadeia de Markov X_n que aproxima Y. Isto é, X_n tem vetor estacionário π. Ou seja,

$\mu_n = \mu_0 P^n \rightarrow \pi$

Vantagem: P é escolhida esperta mente de modo que seja esparsa. Logo, P^n pode ser calculado.

A cadeia de Markov é um passeio aleatório sobre um grafo!

Receita: $G = (V, A)$ onde $V = \{s_1, s_2, \dots, s_k\}$ e A é arbitrário, mas que:

- Bem conectado (garante $P^L > 0$ para algum $L \in \mathbb{N}$)
- Grau pequeno de cada vértice (P^M fácil de calcular)

$$P_{ij,i'} = \begin{cases} \frac{1}{d_i} \min\left\{\frac{\pi_j d_i}{\pi_i d_j}, 1\right\}, & s_i \sim s_j \quad i \neq j \\ 0, & s_i \not\sim s_j, \quad i' \neq j \\ c, & i' = j \end{cases}$$

P tem vetor estacionário π!!

Obs: Às vezes, π_i é difícil de calcular, mas $\frac{\pi_i}{\pi_j}$ é fácil de calcular.

Prove que $\pi_i P_{ij} = \pi_j P_{ji}$ (P é reversível).

Isso implica que $\pi P = \pi$. $\left(\sum_j \pi_j \cdot P_{ji} = \pi_i\right)$

Finalizando.... Podemos considerar o prob. de descriptação via MCMC.

Não dá p/ calcular!

$$\pi(f) = \frac{1}{Z} \prod_i M(f(s_i), f(s_{i+1}))$$

$$Z = \sum_f \prod_i M(f(s_i), f(s_{i+1}))$$

$\frac{\pi(f)}{\pi(f^*)}$ cancela Z!