

[Esercizio per l'esame] Nei cifrari a blocchi e ridondanza, si cerca di rendere difficile l'analisi statistica del ciphertext. Se $m=1$ (si cifra le singole lettere), la frequenza di alcuni blocchi è molto più alta di altre. Se m aumenta, i blocchi tendono a diventare equiprobabili, e la probabilità del singolo blocco diminuisce. In questo modo, l'analisi statistica è più difficile, e serve una quantità enorme di ciphertext per creare gliistogrammi.

[a] Sia p una distribuzione di prob. sull'alfabeto A . σ è una sequenza di m lettere

$$\sigma = x_1, x_2, \dots, x_m$$

Dato un blocco σ di m lettere, quindi $\sigma \in A^m$; $f(a) \stackrel{\Delta}{=} \#$ occorrenze di a in σ

Un blocco σ è tipico se $\forall a \in A$, $f(a) \approx m p(a)$

- Trovare una formula per $p(\sigma)$:

$$\sigma \text{ è una seq. di lettere} \rightarrow \sigma = x_1, x_2, \dots, x_m \rightarrow p(\sigma) = p(x_1, x_2, \dots, x_m)$$

Dato che ogni lettera di σ è indipendente dalle altre $\rightarrow p(\sigma) = p(x_1) \cdot p(x_2) \cdot \dots \cdot p(x_m)$

Ricordiammo gli elementi x_i in modo che le prob. delle singole lettere siamo "vive",

$$\text{si ottiene: } p(\sigma) = \prod_{a \in A} p(a)^{f(a)}$$

$$\begin{aligned} \text{Esempio: } \sigma = a, d, d, b, c, d &\rightarrow p(\sigma) = p(a) \cdot p(d) \cdot p(d) \cdot p(b) \cdot p(c) \cdot p(d) = \\ &= p(d) \cdot p(d) \cdot p(d) \cdot p(a) \cdot p(b) \cdot p(c) = \\ &= p(d)^3 \cdot p(a) \cdot p(b) \cdot p(c) \end{aligned}$$

Nel caso dei blocchi tipici $\rightarrow f(a) \approx m p(a)$, pertanto

$$p(\sigma) = \prod_{a \in A} p(a)^{f(a)} \approx \prod_{a \in A} p(a)^{mp(a)}$$

Per i blocchi $p(\sigma)$ dipende solo da $p(a)$ e dal parametro fisso

[b] \forall è possibile approssimare $p(\sigma) \approx 2^{-mH(p)}$ dove $H(p) = \sum_{a \in A} p(a) \cdot \log(p(a))$, e' detta

entropia, e misura il "disordine" di una distribuzione. Più gli elementi della distribuzione sono equiprobabili, più l'entropia è alta \rightarrow valore massimo: $\log |A|$

- Dimostrare l'approssimazione (Consiglio: usare i logaritmi e poi l'elevamento a potenza)

$$\text{Nel caso dei blocchi tipici} \rightarrow p(\sigma) \approx \prod_{a \in A} p(a)^{m_p(a)}$$

Se si applica il logaritmo a entrambi i membri, si ha:

$$\log(p(\sigma)) \approx \log\left(\prod_{a \in A} p(a)^{m_p(a)}\right)$$

Per la proprietà dei logaritmi, il logaritmo di un prodotto è pari alla somma dei logaritmi. Nel caso del logaritmo di una produttoria, il risultato è pari alla sommatoria dei logaritmi:

$$\log(p(\sigma)) \approx \sum_{a \in A} \log(p(a)^{m_p(a)})$$

$$-H(p)$$

↑

$$\log(p(\sigma)) \approx \sum_{a \in A} m_p(a) \log(p(a)) \rightarrow \log(p(\sigma)) \approx m \sum_{a \in A} p(a) \log(p(a))$$

Sostituendo la definizione di entropia ed elevando a 2 i due membri in modo da rimuovere il logaritmo e primo membro, si ottiene:

$$p(\sigma) \propto 2^{-mH(p)}$$

[e] Fissato un blocco σ_0 , quanti blocchi, in media, devo guardare prima di trovare σ_0 ?

• Esempio lancio di un dado: La probabilità che si ottenga una delle facce, per esempio la m^a , è pari a $1/6$. Per il teorema di Bernoulli, supponiamo che il numero di lanci medio m che devono essere fatti è pari a $\Rightarrow m = \frac{1}{p(a)} = \frac{1}{1/6} = 6$ lanci

• Possiamo fare lo stesso ragionamento alle singole lettere, che possiamo vedere come blocchi di dimensione 1. Per esempio, nel caso dei testi scritti in lingua inglese, la lettera "e" è la più frequente $\rightarrow p(e) \approx 0,127$

Il numero medio di blocchi da leggere prima di trovare una "e" è pari a:

$$m = \frac{1}{p(e)} = \frac{1}{0,127} \approx 8$$

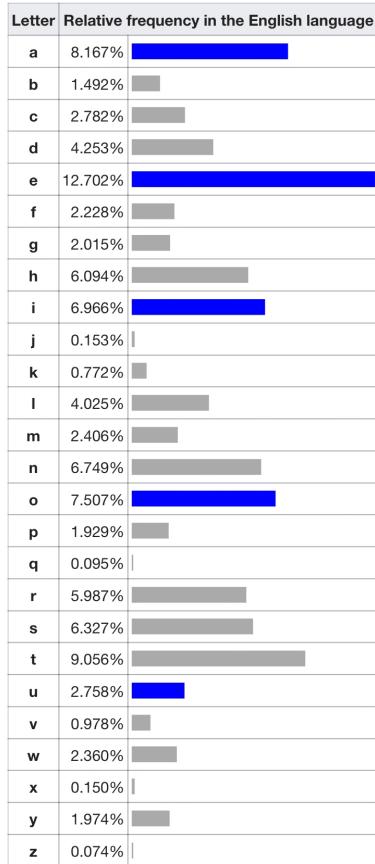
Nel caso di blocchi di dimensione m , la probabilità di un blocco tipico $\bar{\sigma}$ è pari a:

$$p(\bar{\sigma}) \approx 2^{-mH(p)}$$

dove $m = \text{dim. blocchi}$; $H(p) = \text{entropia della dist. } p$

Data una distribuzione di prob. p , l'entropia è data da:

$$H(p) = - \sum_{\alpha \in A} p(\alpha) \log(p(\alpha))$$



Più gli elementi della distribuzione sono equiprobabili, più il valore dell'entropia è alto

Nel caso della lingua inglese, la distribuzione delle singole lettere è data in figura. L'entropia è pari a:

$$H(p) = 4,17$$

Tale valore è stato calcolato tramite uno script python
(entropy.py)

Pertanto si ottiene

$$p(\bar{\sigma}) \approx 2^{-mH(p)} \Rightarrow m = \frac{1}{p(\bar{\sigma})} = 2^{mH(p)} = 2^{4,17} \approx 2^{54} \approx$$

Se le lettere fossero tutte equiprobabili, quindi $\forall \alpha \in A, p(\alpha) = 1/26$,

il valore dell'entropia è praticamente pari al $\log_2 |A| = \log_2(26) \approx 4,7004$

[d] Nel caso di blocchi $\bar{\sigma}$ generici (anche non tipici $\bar{\sigma}$) di dimensione m :

$$\forall \bar{\sigma} \in A^m \Rightarrow p(\bar{\sigma}) = 2^{-m(H(q) + D(q||p))}$$

dove:

$q(\alpha) \stackrel{\Delta}{=} \frac{f(\alpha)}{m} \rightarrow$ Nei blocchi $\bar{\sigma}$ tipici si ha $q(\alpha) \approx p(\alpha)$

$D(q||p) \stackrel{\Delta}{=} \sum_{\alpha \in A} q(\alpha) \log(q(\alpha)/p(\alpha))$ è detta Divergenza KL, e rappresenta la "diversità" delle distribuzioni q e p . Se q e p sono molto simili, la divergenza è molto bassa, fino a diventare nulla nel caso in cui q e p sono uguali

Dim: la probabilità di un blocco $\bar{\sigma}$ generico è data da:

$$p(\bar{\sigma}) = \prod_{\alpha \in A} p(\alpha)^{f(\alpha)}$$

Ricorrendo alle proprietà dei logaritmi, possiamo trasformare la produttoria in sommatoria. Si ottiene

$$\log(p(\tau)) = \sum_{\alpha \in A} \log(p(\alpha)) f(\alpha)$$

Pec definizione $q(\alpha) \stackrel{\Delta}{=} \frac{f(\alpha)}{m} \rightarrow f(\alpha) \stackrel{\Delta}{=} m q(\alpha)$. Tramite le prop. dei logaritmi, si ottiene:

$$\begin{aligned} \log(p(\tau)) &= \sum_{\alpha \in A} \log(p(\alpha))^{mq(\alpha)} = \sum_{\alpha \in A} m q(\alpha) \log(p(\alpha)) = m \sum_{\alpha \in A} q(\alpha) \log\left(\frac{p(\alpha)}{q(\alpha)} \cdot q(\alpha)\right) = \\ &= m \sum_{\alpha \in A} q(\alpha) \left(\log\left(\frac{p(\alpha)}{q(\alpha)}\right) + \log(q(\alpha)) \right) = m \sum_{\alpha \in A} q(\alpha) \underbrace{\log\left(\frac{p(\alpha)}{q(\alpha)}\right)}_{-D(q||p)} + q(\alpha) \underbrace{\log(q(\alpha))}_{-H(q)} \end{aligned}$$

Pertanto, si ottiene:

$$p(\tau) = 2^{-m(H(q) + D(q||p))}$$

[e] Scimmia di Eddington:

- La scimmia batte i testi a caso:
- Distribuzione q dell'opera di Shakespeare: lingua inglese $\rightarrow H(q) = 4,17$
- Distribuzione p della scimmia: $\forall \alpha \in \mathbb{Z}_{26} \rightarrow p_\alpha(\alpha) = 1/26 \rightarrow H(p) = \log_2 |A| = 4,70$

Si ha una divergenza non nulla tra q e p , perciò α :

$$D(q||p) = \sum_{\alpha \in A} q(\alpha) \log\left(\frac{q(\alpha)}{p(\alpha)}\right)$$

Possiamo prendere l'opera di Shakespeare come un blocco tipico di lunghezza $m = 10^6$

$$q(\alpha) \approx p_\alpha(\alpha) \rightarrow D(q||p) = \sum_{\alpha \in A} p_\alpha(\alpha) \log\left(\frac{p_\alpha(\alpha)}{1/26}\right) = 0,364$$

Si ottiene:

$$m = \frac{1}{p(\tau)} \approx 2^{m(H(q) + D(q||p))} = 2^{10^6(4,70 + 0,364)} = 7,91 \cdot 10^{1524415}$$

La scimmia è ammirestata: In questo caso si ha una divergenza trascurabile:

$$m = \frac{1}{p(\tau)} \approx 2^{m+H(q)} = 2^{10^6 \cdot 4,71} = 1,90 \cdot 10^{1417851}$$