Teoria DataSecurity

Jacopo Manetti

April 2023

1 Indici di coincidenza

1.1 Traccia

In questo esercizio vengono approfonditi alcuni aspetti matematici degli indici di coincidenza. Per una sequenza di n caratteri alfabetici $x = (x_1, \ldots, x_n)$, poniamo per ogni carattere $i \in Z_{26}$ f_i = numero di occorrenze del carattere i in $\mathbf{x}$.

Possiamo vedere sia $\mathbf{x}$ che f_i come variabili aleatorie. Assumiamo in particolare che tutte le variabili $x_j (1 \leq j \leq n)$ siano identicamente distribuite secondo una distribuzione di probabilità $\mathbf{p} = (p_1, \ldots, p2s\ )$ sull'alfabeto (per esempio, potrebbe essere $\mathbf{p} = \mathbf{p}E$, la distribuzione tipica della lingua Inglese).

(a) Sin $i \in Z_{2i}$ qualsiasi fissata. Notare che $f_i = \sum_{j=1}^{n} Y_j$, dove Y_j è la variabile Bernoulli che assume il valore 1 se $x_j = i, 0$ altrimenti. Sfruttando la linearità del valore atteso, dimostrare che $E[f_1] = np_i$

(b) Dimostrare quindi che $E\left[f_i^2\right] = \text{var}(f_i) + E[f_i]^2$, dove var= $(f_i) = E\left[f_i^2\right] - E[f_i]^2$ è la varianza di f_{i-}

(c) Assumere ora che i caratteri x_j siano estratti in maniera i.i.d. Notare che f_i è allora una distribuzione binomiale (somma delle n Bernoulli $Y_1, \ldots, Y_n$). Sfruttando la formula per la varianza della distribuzione binomiale e il punto precedente, dimostrare che $E\left|f_i^2\right| = np_i - np_i^2 + n^2p_i^2$ - Concludere che

$$E[f_i(f_i - 1)] = E\left[f_i^2\right] - E[f_i] = n(n-1)p_i^2$$

(d) ricordare che $I_c(\mathbf{x}) = \sum_{i \in Z_{26}} \frac{f_i}{n} \frac{f_i - 1}{n-1}$. sfruttare il risultato del punto precedente per dare una formula esatta di $E[I_c(\mathbf{x})]$. Provare cioè che

$$E[I_c(\mathbf{x})] = \sum_{i=0}^{25} p_i^2.$$

(e) Si consideri la seconda fase dell'attacco a Vigenère, volta a determinare la chiave lettera per lettera. Questa fase implica, dato un certo vettore di probabilità empiriche, $\mathrm{q}_0 = (f_0/n_2 \ldots, f_{25}/n)$, trovare il suo shift circolare di k posizioni $(0 \leq k \leq 25)$, diciamo q. che meglio approssima il vettore delle

probabilità caratteristico della lingua Inglese, p. In altre parole, cerchiamo quel particolare shift circolare q di q_0 tale che vale

$$\mathbf{p} = \mathbf{q}.$$

Trattando per semplicità l'approssimazione di cui sopra come una vera uguaglianza, dimostrare che, tra tutti i 26 shift possibili, questo è il vettore q che massimizzano il prodotto scalare $< \mathbf{p}, \mathbf{q} >$. Allo scopo, servirsi della disuguaglianza di Cauchy-Schwarz, che afferma che, dati due qualsiasi vettori $\mathbf{v}$ e $\mathbf{u}$

$$| < \mathbf{v}, \mathbf{u} > | \leq | \mathbf{u} \|_2 \cdot \| \mathbf{v} \|_2$$

1.2 Svolgimento

(a) Sfruttando la linearità del valore atteso e la definizione di f_i come somma di variabili Bernoulli Y_j con j che assume valori da 1 a n, abbiamo:

$$E[f_i] = E\left[\sum_{j=1}^{n} Y_j\right] \tag{1}$$

Per la linearità del valore atteso, possiamo portare fuori la somma:

$$E[f_i] = \sum_{j=1}^{n} E[Y_j] \tag{2}$$

Ora, poiché Y_j è una variabile Bernoulli, sappiamo che ha valore atteso p_i se $x_j = i$ e $1 - p_i$ altrimenti. Quindi, abbiamo:

$$E[f_i] = \sum_{j=1}^{n} p_i \cdot \Pr(x_j = i) + (1 - p_i) \cdot \Pr(x_j \neq i) \tag{3}$$

Ma poiché abbiamo supposto che tutte le variabili x_j siano identicamente distribuite secondo la stessa distribuzione di probabilità $\mathbf{p}$, abbiamo che $\Pr(x_j = i) = p_i$ e $\Pr(x_j \neq i) = 1 - p_i$. Quindi, possiamo semplificare la sommatoria come segue:

$$E[f_i] = \sum_{j=1}^{n} p_i + \sum_{j=1}^{n} (1 - p_i) = np_i + n(1 - p_i) = np_i \tag{4}$$

Quindi, abbiamo dimostrato che $E[f_i] = np_i$ come richiesto.

(b) Per dimostrare che:

$$E\left[f_i^2\right] = \operatorname{var}(f_i) + E[f_i]^2 \tag{5}$$

basta considerare il valore di $var(f_i)$ che corrisponde a:

$$var(f_i) = E\left[f_i^2\right] - E\left[f_i\right]^2 \tag{6}$$

andando a sostituire quindi il valore della varianza nella formula iniziale otteniamo che:

$$E\left[f_i^2\right] = E\left[f_i^2\right] - E\left[f_i\right]^2 + E\left[f_i\right]^2 \tag{7}$$

quindi semplificando otteniamo che:

$$E\left[f_i^2\right] = E\left[f_i^2\right] \tag{8}$$

e quindi abbiamo dimostrato l'uguaglianza desiderata.

(c) Per dimostrare che:

$$E\left[f_i^2\right] = np_i - np_i^2 + n^2p_i^2 \tag{9}$$

notiamo che f_i è distribuito come una binomiale, ovvero $\mathbf{f} \sim \text{Bin}(n, p_i)$, quindi la sua varianza è $\text{var}(f_i) = np_i(1-p_i)$ e il suo valore atteso è $E\left[f_i\right] = np_i$. Possiamo quindi scrivere:

$$\begin{aligned} E\left[f_i^2\right] &= \text{var}(f_i) + E\left[f_i\right]^2 \\ &= np_i(1-p_i) + n^2p_i^2 \\ &= np_i - np_i^2 + n^2p_i^2 \end{aligned}$$

Inoltre, notiamo che f_i rappresenta il numero di occorrenze del carattere i nella sequenza $\mathbf{x}$. Quindi $f_i(f_i-1)$ rappresenta il numero di coppie di caratteri nella sequenza che sono entrambi uguali a i. Quindi possiamo scrivere:

$$\begin{aligned} E\left[f_i(f_i-1)\right] &= E\left[f_i^2\right] - E\left[f_i\right] \\ &= np_i - np_i^2 + n^2p_i^2 - np_i \\ &= n(n-1)p_i^2 \end{aligned}$$

(d) Per calcolare $E\left[I_c(\mathbf{x})\right]$, dobbiamo calcolare il valore atteso di $I_c(\mathbf{x})$, cioè

$$E\left[I_c(\mathbf{x})\right] = E\left[\sum_{i \in Z26} \frac{f_i}{n}\frac{f_i-1}{n-1}\right] = \sum i \in Z_{26} \frac{E\left[f_i(f_i-1)\right]}{n(n-1)} \text{(linearità del valore atteso)}$$

$$= \sum i \in Z_{26} \frac{n(n-1)p_i^2}{n(n-1)} \text{(risultato del punto (c))} = \sum i = 0^{25} p_i^2. \tag{10}$$

Quindi abbiamo dimostrato che $E\left[I_c(\mathbf{x})\right] = \sum_{i=0}^{25} p_i^2$

Per dimostrare che il vettore q che massimizza il prodotto scalare $< \mathbf{p}, \mathbf{q} >$ è il vettore q che approssima $\mathbf{p}$, possiamo utilizzare la disuguaglianza di Cauchy-Schwarz.

Sappiamo che il prodotto scalare tra $\mathbf{p}$ e $\mathbf{q}$ è dato da:

$$< \mathbf{p}, \mathbf{q} >= \sum_{i=0}^{25} p_i q_i \tag{11}$$

Sappiamo inoltre che $|\mathbf{u} \cdot \mathbf{v}| \leq |\mathbf{u}|_2 \cdot |\mathbf{v}|_2$, per ogni coppia di vettori $\mathbf{u}$ e $\mathbf{v}$.

Applichiamo questa disuguaglianza al vettore $\mathbf{p}$ e al vettore circolato $\mathbf{q}$, ottenuto applicando uno shift di k posizioni al vettore $\mathbf{q}0$:

$$|\mathbf{p} \cdot \mathbf{q}| \leq |\mathbf{p}|2 \cdot |\mathbf{q}|2 \ = \sqrt{\sum i = 0^{25} p_i^2} \cdot \sqrt{\sum i = 0^{25} q(i-k) \mod 26^2} =$$

$$= \sqrt{\sum_{i=0}^{25} p_i^2} \cdot \sqrt{\sum_{i=0}^{25} q_i^2} \ = |\mathbf{p}|_2 \cdot |\mathbf{q}_0|_2$$

Dato che il prodotto scalare è massimo quando la disuguaglianza di Cauchy-Schwarz diventa un'uguaglianza, possiamo affermare che il prodotto scalare è massimo quando $\mathbf{q}$ è un vettore che rispetta l'uguaglianza, cioè un vettore che approssima $\mathbf{p}$. Inoltre, il vettore $\mathbf{q}$ che approssima meglio $\mathbf{p}$ è quello che ha la massima norma 2, dato che la disuguaglianza di Cauchy-Schwarz diventa un'uguaglianza solo quando i due vettori hanno la stessa direzione.

Pertanto, il vettore circolato $\mathbf{q}$ che massimizza il prodotto scalare $< \mathbf{p}, \mathbf{q} >$ è quello che approssima meglio $\mathbf{p}$, ovvero quello che differisce da $\mathbf{q}0$ di uno shift circolare di k posizioni, dove k è tale che la norma 2 di $\mathbf{q}$ sia massima:

$$k = \operatorname{argmax} i = 0^{25} \sum_{j=0}^{25} q_{(j-i) \mod 26}^2 \tag{12}$$

2 Un crittogramma Vigenere

2.1 Traccia

Si decifri il seguente testo cifrato, ottenuto tramite un cifrario di Vigenere a partire da un plaintext in lingua Inglese.

```
OKZARVGLNSLFOQRVVBPHHZAMOMEVHLBAITLZOWSXCSZFEQFICOOVDXCIISOOVXEIYWNHHLVQHSOWD
BRPTTZZOWJIYPJSAWQYNOYRDKBQKZPHHTLIHDEMICGYMSEVHKVXTQPBWMEWAZZKHLJMOVEVHJYSJR
ZTUMCVDGLZVBUIWOCPDZVEIGSOGZRGGOTAHLCSRSCXXAGPDYPSYMECRVPFHMYWZCYHKMCPVBPHYIF
WDZTGVIZEMONVYQYMCOOKDVQIMSOKLBUEBFZISWSTVFEWVIAWACCGHDRVVZOOBANRYHSSQBUIMSDW
VBNRXSSOGLVWKSCGHLNRYHSSQLVIYCFHVWJMOVEKRKBQMOOSVQAHDGLGWMEOMCYOXMEEIRTZBCFLZ
BVCVPRQVBLUHLGSBSEOUWHRYHSSEIEVDSCGHZRGOSOPBBUIQWNHRZFEIRPBWMEXCSPASBLXZFCWWW
```

```
EMZGLDDBUIOWNTHVPICOOTRZOMYRPBKMEIIHCOQKRWCSNFRAFIYWEKLBUSPHEVHAYMBVESVBGVZAZ
FVPRAJIWRQMIIMUZPDKXXJHSSRBUIMGTRHBUIMSHCXTQFZBZFHBHVIOYRWPRXCFPSRNGLZAVBHEVX
OVPMZMEIAIWZBIJEMSEVDBGLZMHSUMGVVWWWQOGLZCCPLARWYSNZLVRXCOEHKMLAZFPGLVXMIUHWW
PVXDBECWPRJDBLZQQTLOALFHBUIKOEVZWHPYPPRLNSMXIWHWPNXOCZHKMLOISH
```

Si impieghi il metodo degli indici di coincidenza, eventualmente accoppiato con il metodo di Kasiski. Illustrare i vari passaggi, in particolare:

1. le ripetizioni nel testo, le loro distanze e i valori di m (lunghezza della chiave) da esse suggeriti;
2. i valori degli indici di coincidenza che si ottengono per il valore corretto di m;
3. per ciascuna delle m lettere della chiave, come ‘e stato individuato lo shift che d‘a il valore della chiave.

Si puo far uso di strumenti disponibili online per il computo delle statistiche rilevanti del testo (ripetizioni, frequenze, etc.), come per esempio quelli offerti dalla pagina Black Chamber di Simon Singh:

http://www.simonsingh.net/The_Black_Chamber/vigenere_cracking_tool.html

2.2 Svolgimento

Iniziamo cercando le ripetizioni nel testo cifrato e calcolando le loro distanze, per facilitarci il compito utilizziamo il sito disponibile alla pagina Black Chamber di Simon Singh:

`http://www.simonsingh.net/The_Black_Chamber/vigenere_cracking_tool.html`

Da qui passando il cyphertext proposto otteniamo:

Vigenere Repeat Distance Repeated Sequence	Spacing	Possible length of key (or factors)																	
		02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
LNS	725				x															
RVV	272	x		x				x								x	x			
VBP	208	x		x				x					x			x				
BPH	208	x		x				x					x			x				
PHH	88	x		x				x			x									
EVH	96	x	x	x		x		x				x				x				
EVH	24	x	x	x		x		x				x								
EVH	376	x		x				x												
ZOW	48	x	x	x		x		x				x				x				
XCS	408	x	x	x		x		x				x					x			
ZFE	395				x															
ICO	432	x	x	x		x		x	x			x				x		x		
COO	200	x		x	x			x		x										x

Questa è solo una parte dei risultati ottenuti, analizzando questa prima parte si nota che i tri-grammi che si ripetono più spesso sono:

- VBP distanza: 208
- BPH distanza: 208
- EVH distanza: 24

- PHH distanza: 88
- ZOW distanza: 48
- XCS distanza: 408
- ICO distanza: 432
- COO distanza: 200

Calcolando l'MCD tra le distanze otteniamo $MCD(208, 24, 88, 48, 408, 432, 200) = 8$, dall'analisi completa delle ripetizioni si nota infatti che i valori più probabili per m (lunghezza della chiave) sono 2,4,8 e 16.

Ipotizziamo quindi 8 per la lunghezza della chiave.

Per accertarci che la lunghezza della chiave sia 8, possiamo utilizzare il metodo degli indici di coincidenza.

Per fare ciò, dividiamo il testo cifrato in m sotto-testi, in cui ogni sotto-testo è costituito dalle lettere cifrate utilizzando la stessa lettera della chiave. Per ogni sotto-testo, calcoliamo l'indice di coincidenza utilizzando la formula:

$$I.C. = \frac{\sum_{i=1}^{n} f_i(f_i - 1)}{N(N-1)}$$

Dove n è il numero di possibili valori che ogni elemento della sequenza può assumere (nel nostro caso 26), f_i è il numero di volte che l'i-esimo valore appare nella sequenza, e N è il totale di elementi nella sequenza.

Nel nostro caso, essendo m=8, dividiamo il testo in 8 sotto-testi, ad esempio il primo sotto-testo ottenuto sarà:

```
Sotto-testo 1:
ONVOICCIYHTYNKHMTAOJDWIGRDRZVDMCMBVAVYMXKYYOMDMIZVBYD
OQICZZOCYICYPBZJMJMMZICZOAMZVZYCZIDDOKYXOI
```

il cui indice di coincidenza è: 0.0645

Ripetiamo per tutti i sotto-testi:

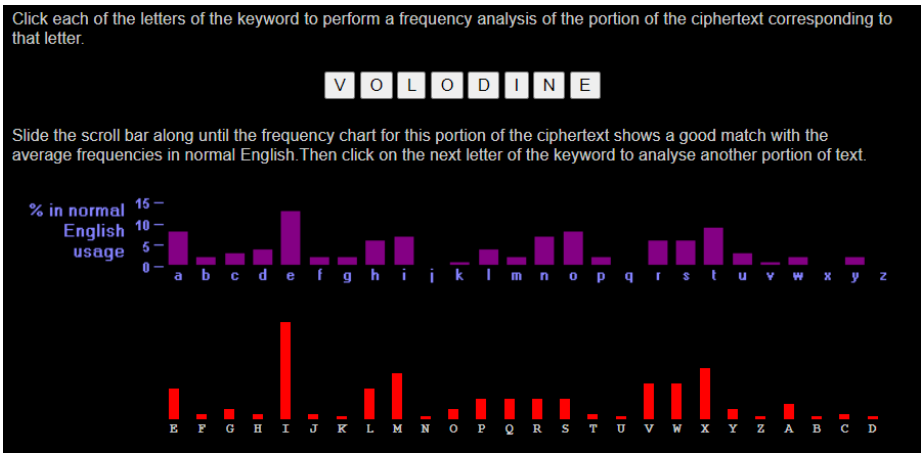
Sotto-testo 1:	0.0645
Sotto-testo 2:	0.0759
Sotto-testo 3:	0.0808
Sotto-testo 4:	0.0578
Sotto-testo 5:	0.0803
Sotto-testo 6:	0.0768
Sotto-testo 7:	0.0594
Sotto-testo 8:	0.0809

I valori ottenuti si avvicinano a 0.065, concludiamo quindi che m=8.

Una volta determinata la lunghezza della chiave, si può procedere alla sua individuazione. Per fare ciò, si può considerare il testo cifrato come una sequenza di m sotto-testi, ognuno dei quali è stato cifrato con uno shift diverso rispetto alla chiave. Per individuare lo shift utilizzato per cifrare ciascun sotto-testo, si può utilizzare l'analisi delle frequenze. In particolare, per ogni sotto-testo

si calcola la frequenza di ogni lettera e si confronta con la distribuzione delle frequenze delle lettere in lingua inglese. Lo shift corrispondente a ciascuna lettera della chiave sarà quello che minimizza la differenza tra la distribuzione delle frequenze delle lettere nel sotto-testo e la distribuzione delle frequenze delle lettere in lingua inglese.

Con il sito visto in precedenza è possibile shiftare le singole lettere manualmente per far combaciare le frequenze del cyphertext con quelle della lingua inglese:



Nell'immagine vengono mostrate le frequenze delle lettere dell'ottavo sotto-testo (in rosso) con le frequanze delle lettere nella lingua inglese (in viola), da questo shift si ottiene l'ultima lettera della chiave, cioè una E.

Ripetendo per tutti i sotto-testi si ottiene che la chiave è:

```
VOLODINE
```

Il messaggio decriptato, applicando gli opportuni spazi è:

```
Two months earlier, an eternity, the downfall of the Orbis E had happened as
predicted. Immediately followed by exodus and a completely empty future. The
city centers flowed with the blood of reprisals. The barbarians had reclaimed
power just like everywhere else on the planet. Vassilis Samarachvili had
wandered with a group of partisans for several days, and then the resistance had
dispersed and then died out. So, with two comrades in disaster, Kronauer and
Ilyushenko, she managed to get around the barriers erected by the victors and
enter the empty territories. Apathetic fence had forbidden her entry.
```