Teoria Data Security 3

Jacopo Manetti

June 2023

1 Sequenze tipiche

1.1 Traccia

Fissiamo $n \geq 1$. Sia $\sigma = x_1, x_2, \ldots, x_n$ una sequenza di n lettere estratte da un alfabeto finito e non vuoto X. Il tipo della sequenza σ, denotato con t_σ, è la distribuzione empirica che essa determina su X. Ovvero, per ogni $x \in X$, denotando con $f(x)$ la frequenza di x in σ, si pone:

$$t_\sigma(x) \stackrel{def}{=} \frac{f(x)}{n}$$

Chiameremo t_σ un n-tipo se vogliamo ricordare che è stato ottenuto da una sequenza lunga n. (a) Sia p una distribuzione su X, e supponiamo che σ sia la realizzazione di n variabili $X_1, \ldots, X_n$, con le X_i estratte i.i.d. secondo p. Dimostrare che la probabilità dell'intera sequenza σ sotto p puo essere scritta come:

$$p(\sigma) = 2^{-n(H(t_\sigma)+D(t_\sigma \| p))}.$$

(b) Sia t il tipo di una certa sequenza. Consideriamo tutte le sequenze di lunghezza n che hanno lo stesso tipo t, ovvero

$$\mathcal{B}_i \stackrel{def}{=} \{\sigma : t_\sigma = t\}$$

Sfruttando il punto precedente, dimostrare che

$$|\mathcal{B}_t| \leq 2^{nH(t)}$$

(Si noti che deve essere $1 \geq t(\mathcal{B}_t)$, da cui …). (c) Dato un certo tipo t, e una qualsiasi distribuzione di probabilità $p(\cdot)$ su X, dimostrare che è possibile maggiorare la probabilità che venga estratta una qualsiasi sequenza di tipo t come segue

$$p(\mathcal{B}_t) \leq 2^{-nD(t||p)}.$$

(d) Si denoti con $\mathcal{T}_n$ l'insieme di tutti gli n-tipi su X. Si dimostri che $|\mathcal{T}_n| \leq (n+1)^{|X|}$. Dunque, una volta fissato l'alfabeto X, l'insieme degli n-tipi cresce a velocità polinomiale rispetto a n (NB: per il calcolo, ragionare sui modi possibili di distribuire n palline in $|X|$ urne; la quantità $(n+1)^{|X|}$ rappresenta una limitazione superiore molto grossolana, ma sufficiente per il seguito). (e) Sfruttando i due punti precedenti, di dimostri che la probabilità di estrarre una qualsiasi sequenza il cui tipo dista più di un $\varepsilon > 0$ fissato da p, in termini di divergenza KL, è $\leq (n+1)^{|X|} 2^{-n\varepsilon}$. Ovvero

$$p\left(\bigcup\{\mathcal{B}_t : D(t||p) \geq \varepsilon\}\right) = \sum_{t:D(t||p)\geq\varepsilon} p\left(\mathcal{B}_t\right) \leq (n+1)^{|X|} 2^{-n\varepsilon}$$

L'utilità di questa stima risiede nel fatto che, per n abbastanza grande, il fattore polinomiale $(n+1)^{|X|}$ diventa trascurabile rispetto all'esponenziale negativo $2^{-n\varepsilon}$. (f) Sfruttado la stima del punto precedente, dare una limitazione superiore alla probabilità del seguente evento: in $n = 500$ lanci di un dado, il tipo della sequenza risultante dista più di 0,15 (in divergenza KL) dalla distribuzione uniforme. Verificare se un qualsiasi n-tipo che assegna più del 50% di probabilità alla faccia ' 6 ' dista più o meno di 0,15 dalla distribuzione uniforme.

1.2 Svolgimento

(a) Si deve dimostrare che

$$p(\sigma) = 2^{-n(H(t_\sigma)+D(t_\sigma|p))} \tag{1}$$

Dove $H(t_\sigma)$ è l'entropia di Shannon di t_σ e $D(t_\sigma|p)$ è la divergenza di Kullback-Leibler tra t_σ e p.

Per la definizione di probabilità su una sequenza di variabili aleatorie i.i.d., abbiamo:

$$p(\sigma) = p(x_1) \cdot p(x_2) \cdot \ldots \cdot p(x_n) \tag{2}$$

Ora, applicando il logaritmo e la negazione a entrambi i lati dell'equazione, otteniamo:

$$-n \log(p(\sigma)) = \sum_{i=1}^{n} -\log(p(x_i)) \tag{3}$$

Ora si può applicare il principio del valore atteso. Dato che le variabili aleatorie X_i sono estratte i.i.d. secondo p, l'equazione diventa:

$$-n \log(p(\sigma)) = n \sum_{x\in X} t_\sigma(x) \cdot -\log(p(x)) \tag{4}$$

Questa è la definizione di entropia incrociata $H(t_\sigma, p)$. Ma per definizione, $H(t_\sigma, p) = H(t_\sigma) + D(t_\sigma|p)$, e quindi abbiamo:

$$-n \log(p(\sigma)) = n\left(H\left(t_\sigma\right) + D\left(t_\sigma|p\right)\right) \quad (5)$$

Tornando alla notazione esponenziale, otteniamo:

$$p(\sigma) = 2^{-n(H(t_\sigma)+D(t_\sigma|p))} \quad (6)$$

Che è esattamente quello che volevamo dimostrare.

(b) Dobbiamo dimostrare che

$$|\mathcal{B}_t| \leq 2^{nH(t)} \quad (7)$$

Ogni sequenza in $\mathcal{B}_t$ ha lunghezza n e tipo t. Quindi, tutte le sequenze in $\mathcal{B}_t$ hanno la stessa distribuzione empirica, ovvero la stessa distribuzione di frequenze delle lettere $x \in X$. In altre parole, tutte le sequenze in $\mathcal{B}_t$ sono permutazioni l'una dell'altra.

Il numero di queste permutazioni è dato da:

$$|\mathcal{B}t| = \frac{n!}{\prod x \in X(n \cdot t(x))!} \quad (8)$$

Usando l'approssimazione di Stirling $n! \approx n^n e^{-n}$, otteniamo:

$$|\mathcal{B}t| \approx \frac{n^n e^{-n}}{\prod x \in X((n \cdot t(x))^{nt(x)} e^{-nt(x)})} \quad (9)$$

Semplificando, otteniamo:

$$|\mathcal{B}_t| \approx 2^{nH(t)} \quad (10)$$

(c) Dobbiamo dimostrare che

$$p\left(\mathcal{B}_t\right) \leq 2^{-nD(t||p)} \quad (11)$$

Usiamo il risultato del punto (a). La probabilità di una singola sequenza $\sigma \in \mathcal{B}t$ sotto p è $2^{-n(H(t\sigma)+D(t_\sigma|p))}$. Dal momento che tutte le sequenze in $\mathcal{B}_t$ hanno lo stesso tipo t, tutte hanno la stessa probabilità. Quindi, la probabilità totale di tutte le sequenze in $\mathcal{B}_t$ è data da:

$$p\left(\mathcal{B}_t\right) = |\mathcal{B}_t| \cdot 2^{-n(H(t)+D(t|p))} \quad (12)$$

Ma dal punto (b) sappiamo che $|\mathcal{B}_t| \leq 2^{nH(t)}$. Quindi:

$$p\left(\mathcal{B}_t\right) \leq 2^{nH(t)} \cdot 2^{-n(H(t)+D(t|p))} = 2^{-nD(t||p)} \quad (13)$$

Che è esattamente quello che volevamo dimostrare.

(d) Per dimostrare che $|\mathcal{T}_n| \leq (n+1)^{|X|}$, si può utilizzare il principio della combinazione con ripetizioni. In questo caso, stiamo distribuendo n palline (le frequenze delle lettere nelle sequenze) in $|X|$ urne (le possibili lettere). Il numero di modi in cui questo può essere fatto è dato dal coefficiente binomiale multinomiale, che è delimitato da $(n+1)^{|X|}$.

(e) Dobbiamo dimostrare che

$$p\left(\bigcup \mathcal{B}t : D(t||p) \geq \varepsilon\right) = \sum t : D(t||p) \geq \varepsilon p\left(\mathcal{B}_t\right) \leq (n+1)^{|X|} 2^{-n\varepsilon} \quad (14)$$

Usiamo i risultati dei punti (c) e (d). Il lato sinistro dell'equazione è una somma su tutti i tipi t che hanno una divergenza di Kullback-Leibler da p maggiore o uguale a ε. Per ogni tipo t, la probabilità di $\mathcal{B}_t$ è limitata da $2^{-nD(t||p)}$. Quindi, la somma totale è limitata da:

$$\sum_{t:D(t||p)\geq\varepsilon} 2^{-nD(t||p)} \leq \sum_{t:D(t||p)\geq\varepsilon} 2^{-n\varepsilon} \quad (15)$$

Il lato destro dell'equazione è il prodotto del numero di tipi $|\mathcal{T}_n|$ e il termine esponenziale $2^{-n\varepsilon}$. Dal punto (d), sappiamo che $|\mathcal{T}_n| \leq (n+1)^{|X|}$. Quindi otteniamo:

$$\sum_{t:D(t||p)\geq\varepsilon} 2^{-n\varepsilon} \leq (n+1)^{|X|} 2^{-n\varepsilon} \quad (16)$$

Che è esattamente quello che volevamo dimostrare.

(f) Utilizzando la stima del punto (e), possiamo calcolare la probabilità dell'evento richiesto. In questo caso, $n = 500$, $|X| = 6$ (poiché stiamo lanciando un dado a 6 facce), e $\varepsilon = 0.15$. Quindi la probabilità dell'evento è limitata da $(501)^6 \cdot 2^{-500 \cdot 0.15}$, pertanto, la probabilità di ottenere una sequenza di lanci del dado il cui tipo dista più di 0.15 dalla distribuzione uniforme è molto bassa.

Per la seconda parte della domanda, consideriamo un n-tipo che assegna più del 50% di probabilità alla faccia '6'. Questo significa che la distribuzione del tipo è molto diversa dalla distribuzione uniforme su X. In particolare, la divergenza di Kullback-Leibler tra il tipo e la distribuzione uniforme sarà molto grande. Pertanto, è molto probabile che la divergenza di Kullback-Leibler sia maggiore di 0.15.