Richiami di probabilità

Informale

Paolo D'Arco
pdarco@unisa.it

Università di Salerno

Elementi di Crittografia

Contenuti

Richiami di probabilità

Ω: insieme di tutti i possibili risultati di un esperimento (eventi elementari)

Un *evento* E è un sottoinsieme di Ω

Una probabilità è un modo di assegnare ad ogni evento un valore tra 0 e 1 con la condizione che l'evento Ω ha probabilità 1. Precisamente, per ogni $E \subseteq \Omega$, risulta

$$Pr(E) \geq 0 \text{ e } Pr(\Omega) = 1.$$

Inoltre, se E_1 ed E_2 sono mutualmente esclusivi, cioè non hanno risultati "in comune", allora

$$Pr(E_1 \vee E_2) = Pr(E_1) + Pr(E_2).$$

Se $\bar{E} = \Omega \setminus E$ indica il complemento di $E \subseteq \Omega$, allora

$$Pr(\bar{E}) = 1 - Pr(E).$$

Se E_1 ed E_2 sono eventi, allora

$$Pr(E_1 \wedge E_2) \leq Pr(E_1).$$

mentre,

$$Pr(E_1 \vee E_2) \geq Pr(E_1) \qquad \text{e} \qquad Pr(E_1 \vee E_2) \leq Pr(E_1) + Pr(E_2).$$

In generale, dati k eventi, vale il seguente risultato (*union bound*)

$$Pr(\bigvee_{i=1}^{k} E_i) \leq \sum_{i=1}^{k} Pr(E_i).$$

Richiami di probabilità

La probabilità condizionata di E_1 dato E_2, denotata con $Pr(E_1|E_2)$, è definita come

$$Pr(E_1|E_2) \stackrel{def}{=} \frac{Pr(E_1 \wedge E_2)}{Pr(E_2)}, \qquad \text{dove } Pr(E_2) > 0.$$

Segue che:

$$Pr(E_1 \wedge E_2) = Pr(E_1|E_2) \cdot Pr(E_2).$$

Teorema di Bayes. Se $Pr(E_2) \neq 0$, allora

$$Pr(E_1|E_2) = \frac{Pr(E_2|E_1) \cdot Pr(E_1)}{Pr(E_2)}.$$

Dim.

$$Pr(E_1|E_2) = \frac{Pr(E_1 \wedge E_2)}{Pr(E_2)} = \frac{Pr(E_2 \wedge E_1)}{Pr(E_2)} = \frac{Pr(E_2|E_1) \cdot Pr(E_1)}{Pr(E_2)}.$$

Gli eventi E_1 ed E_2 sono probabilisticamente *indipendenti* se

$$Pr(E_1 \mid E_2) = Pr(E_1).$$

Il verificarsi di E_2, cioè, non cambia la probabilità che si verifichi E_1.

Nota che, se E_1 ed E_2 sono indipendenti, risulta

$$Pr(E_1) = Pr(E_1 \mid E_2) = \frac{Pr(E_1 \wedge E_2)}{Pr(E_2)}$$

che implica:

$$Pr(E_1 \wedge E_2) = Pr(E_1) \cdot Pr(E_2).$$

Diremo che gli eventi $E_1, E_2, \ldots, E_n$ costituiscono una *partizione* di Ω se

$$Pr(E_1 \vee E_2 \vee \ldots \vee E_n) = 1 \qquad \text{e per ogni } i \neq j, Pr(E_i \wedge E_j) = 0.$$

In tal caso, per qualsiasi $F \subseteq \Omega$, risulta

$$Pr(F) = \sum_{i=1}^{n} Pr(F \wedge E_i).$$

Nel caso in cui $n = 2$, risulta $E_2 = \bar{E}_1$ e quindi

$$\begin{aligned} Pr(F) &= Pr(F \wedge E_1) + Pr(F \wedge E_2) \\ &= Pr(F \wedge E_1) + Pr(F \wedge \bar{E}_1) \\ &= Pr(F \mid E_1) \cdot Pr(E_1) + Pr(F \mid \bar{E}_1) \cdot Pr(\bar{E}_1). \end{aligned}$$

Prendendo $F = E_1 \vee E_2$, per qualsiasi E_2, otteniamo una limitazione *migliore* dell'union bound

$$\begin{aligned} Pr(E_1 \vee E_2) &= Pr(E_1 \vee E_2 \mid E_1) \cdot Pr(E_1) + Pr(E_1 \vee E_2 \mid \bar{E}_1) \cdot Pr(\bar{E}_1) \\ &\leq Pr(E_1) + Pr(E_2 \mid \bar{E}_1). \end{aligned}$$

Estendendo il risultato agli eventi $E_1, E_2, \ldots, E_n$, vale il seguente

$$Pr(\bigvee_{i=1}^{n} E_i) \leq Pr(E_1) + \sum_{i=2}^{n} Pr(E_i \mid \bar{E}_1 \wedge \ldots \wedge \bar{E}_{i-1}).$$

Problema del compleanno

Se scegliamo q elementi $y_1, \ldots, y_q$ uniformemente a caso da un insieme di taglia N, qual è la probabilità che esistano i e j distinti tali che $y_i = y_j$ (collisione)?

Indichiamo la probabilità dell'evento con $coll(q, N)$.

Problema del compleanno: quanto deve essere numeroso un gruppo di persone affinchè, con probabilità almeno $1/2$, due di esse siano nate lo stesso giorno?

Problema del compleanno

Corrispondenza:

- assumendo che i compleanni siano uniformemente distribuiti e che $N = 365$
- y_i rappresenti il compleanno della persona i-esima nel gruppo $y_1, \ldots, y_q$

la soluzione al problema del compleanno consiste nel trovare

$$\text{il minimo } q \text{ per cui risulta } coll(q, 365) \geq 1/2$$

Sorprendentemente $q = 23$ è sufficiente.

Upper bound

Lemma $A.15$. Sia N un intero positivo fissato, e siano $y_1, \ldots, y_q$ q elementi scelti indipendentemente ed uniformemente da un insieme di taglia N. La probabilità che esistano i e j distinti per cui $y_i = y_j$ è

$$coll(q, N) \leq q^2/2N.$$

Dim. Applichiamo l'union bound. Sia

- $Coll$ l'evento che denota una collisione
- $Coll_{i,j}$ l'evento $y_i = y_j$

Per le assunzioni fatte, $Pr[Coll_{i,j}] = 1/N$ per ogni i e j distinti e $Coll = \bigvee_{i \neq j} Coll_{i,j}$. Pertanto,

$$Pr[Coll] = Pr[\bigvee_{i \neq j} Coll_{i,j}] \leq \sum_{i \neq j} Pr[Coll_{i,j}] = \binom{q}{2} \cdot \frac{1}{N} \leq \frac{q^2}{2N}.$$

Lower bound

Lemma $A.15$. Sia N un intero positivo fissato, e siano $y_1, \ldots, y_q$ $q \leq \sqrt{2N}$ elementi scelti indipendentemente ed uniformemente da un insieme di taglia N. Allora la probabilità che esistano i e j distinti tali che $y_i = y_j$ è

$$coll(q, N) \geq 1 - e^{-\frac{q \cdot (q-1)}{2N}} \geq \frac{q \cdot (q-1)}{4N}.$$

Dim. Consultate l'Appendice A del libro di testo.

Conclusione: se $q = \Theta(\sqrt{N})$, la probabilità di avere una collisione è *costante*.

Variabili casuali

Variabile casuale: variabile che può assumere un insieme di differenti valori, ciascuno con una probabilità associata.

I valori vengono assunti in accordo al risultato dell'esperimento sottostante. Più formalmente

$$X : \Omega \to S$$

dove Ω è lo spazio degli eventi elementari con relative probabilità ed S un insieme di valori.

Solitamente S è un insieme finito di numeri reali.

Se X non assume valori negativi, è detta *non negativa*.

Se $S = \{0, 1\}$, X viene detta variabile casuale $0/1$ (o binaria).

Il concetto può essere esteso al caso più generale in cui S contiene altri elementi: vettori, sequenze, matrici ...

Valore medio

Diremo che le variabili casuali 0/1 $X_1, \ldots, X_k$ sono indipendenti se, per tutti i $b_1, \ldots, b_k$, vale che

$$Pr[X_1 = b_1 \wedge \ldots \wedge X_k = b_k] = \prod_{i=1}^{k} Pr[X_i = b_i].$$

Il valore medio $Exp(X)$ della variabile casuale X è definito come

$$Exp(X) = \sum_{s \in S} Pr[X = s] \cdot s.$$

Nota che può essere un valore $\notin S$.

Il valore medio soddisfa la proprietà di linearità. Date le variabili casuali $X_1, \ldots, X_k$ risulta:

$$Exp[\sum_{i=1}^{k} X_i] = \sum_{i=1}^{k} Exp[X_i].$$

Disuguaglianza di Markov

Se X_1 e X_2 sono indipendenti

$$Exp(X_1 X_2) = Exp(X_1) \cdot Exp(X_2).$$

Quando "si sa poco" di una variabile casuale, la disuguaglianza di Markov risulta utile.

Disuguaglianza di Markov. Sia X una variabile casuale non negativa, e sia $v > 0$. Allora

$$Pr[X \geq v] \leq \frac{Exp[X]}{v}.$$

Dim. Supponiamo X assuma valori in S. Risulta:

$$\begin{aligned}
Exp[X] &= \sum_{s \in S} Pr[X = s] \cdot s \\
&\geq \sum_{s \in S, s < v} Pr[X = s] \cdot 0 + \sum_{s \in S, s \geq v} Pr[X = s] \cdot \mathrm{v} \\
&\geq \mathrm{v} \cdot Pr[X \geq \mathrm{v}].
\end{aligned}$$

Varianza

La varianza di una variabile causale X, denotata con $Var[X]$, misura *quanto X devia dal valore medio.*

$$Var[X] \stackrel{def}{=} Exp[(X - Exp[X])^2] = Exp[X^2] - Exp[X]^2.$$

Si può facilmente mostrare che

$$Var[aX + b] = a^2 Var[X].$$

Inoltre, per variabili causali $0/1$ X_i risulta $Var[X_i] \leq 1/4$ perchè in questo caso $Exp[X_i] = Exp[X_i^2]$ e quindi

$$Exp[X_i^2] - Exp[X_i]^2 = Exp[X_i](1 - Exp[X_i])$$

che ha valore massimo per $Exp[X_i] = 1/2$ da cui $Var[X_i] \leq 1/4$.

Disuguaglianza di Chebychev

Disuguaglianza di Chebychev. Sia X una variabile casuale e sia $\delta > 0$. Allora

$$Pr[|X - Exp[X]| \geq \delta] \leq \frac{Var[X]}{\delta^2}.$$

Dim. Applicando la disuguaglianza di Markov.

Un po' di esempi semplici

Sia $\Omega = \{T, C\}$ (lancio di una moneta). Esempi di distribuzioni sono:

$$Pr[T] = 1/2 \quad Pr[C] = 1/2 \quad \Rightarrow Pr[\Omega] = 1 \text{ (distribuzione uniforme)}$$

$$Pr[T] = 1/4 \quad Pr[C] = 3/4 \quad \Rightarrow Pr[\Omega] = 1 \text{ (distribuzione non uniforme)}$$

In generale, se S è un insieme finito di valori e

$$X : \Omega \to S \quad \text{è tale che} \quad Pr[X = s] = \frac{1}{|S|} \quad \forall s \in S,$$

la distribuzione di probabilità si dice *uniforme* e la variabile aleatoria si dice *uniformemente distribuita*.

Algoritmi randomizzati e variabili casuali

Sia A un algoritmo probabilistico (o randomizzato, cioè che usa random bit).

Le variabili casuali sono utili per rappresentare l'output di A.

Precisamente, la variabile casuale A rappresenta i possibili valori - con relative probabilità - che l'algoritmo A può dare in output, a seconda delle scelte casuali che compie.

Algoritmi randomizzati e variabili casuali

In uno schema di cifratura

- $Gen() \rightarrow k \in K$ (spazio delle chiavi)
  - la variabile casuale K può essere usata per rappresentare i possibili $k \in K$ che l'algoritmo di generazione delle chiavi può dare in output, a seconda delle scelte casuali che effettua
- $Enc_k(m) \rightarrow c \in C$ (spazio dei cifrati)
  - la variabile casuale $\mathrm{C}_{k,m}$ può essere usata per rappresentare i possibili $c \in C$ che, dati k ed m, l'algoritmo di cifratura può dare in output, a seconda delle scelte casuali che effettua

Algoritmi randomizzati e variabili casuali

Nota: se siamo interessati a valutare la probabilità dei cifrati in generale (non per una specifica chiave ed uno specifico messaggio) utilizziamo la variabile casuale C definita da

$$Pr[\mathsf{C} = c] = Pr[Enc_{\mathsf{K}}(\mathsf{M}) = c]$$

La distribuzione di C dipende dalle distribuzioni di K ed M e dalle scelte casuali che l'algoritmo di cifratura effettua.

$$Pr[\mathsf{C} = c] = \sum_{k \in K} \sum_{m \in M} \underbrace{Pr[Enc_k(m) = c | \mathsf{K} = k, \mathsf{M} = m]}_{\Downarrow \; \mathsf{C}_{k,m}} \cdot Pr[\mathsf{K} = k \wedge \mathsf{M} = m]$$

Nelle analisi nel testo non confondete oggetti diversi (e.g., $Enc_{\mathsf{K}}(\mathsf{M})$ con $Enc_k(m)$).

Algoritmi randomizzati e variabili casuali

Sia ancora $\Omega = \{T, C\}$. Un altro esempio di distribuzione è:

$$Pr[T] = 1 \quad Pr[C] = 0 \quad \Rightarrow Pr[\Omega] = 1 \text{ (distribuzione degenere)}$$

Un algoritmo deterministico può essere visto come un caso particolare degli algoritmi probabilistici, in cui la distribuzione di probabilità della variabile casuale che rappresenta l'output è degenere, i.e., ha valore 1 in un punto e 0 in tutti gli altri.

Famiglie di distribuzioni*

Una famiglia di distribuzioni (distribution ensemble) è una famiglia di distribuzioni di probabilità o variabili casuali

$$X = \{X_i\}_{i \in I}, \qquad \text{dove}$$

- X_i denota una distribuzione di probabilità o variabile casuale
- i è l'indice che denota la i-esima distribuzione
- I è l'insieme degli indici e può essere
  - un sottoinsieme degli interi
  - un insieme di stringhe
  - un generico insieme contabile

Famiglie di distribuzioni*: perchè...

Studio degli algoritmi: utilizziamo l'analisi asintotica per capire, al crescere della taglia dell'input, come si comportano gli algoritmi progettati.

- e.g., si pensi agli algoritmi di ordinamento e all'analisi della loro efficienza in funzione del numero di elementi da ordinare

Crittografia: valuteremo *asintoticamente* il comportamento degli schemi crittografici al crescere di un parametro, detto *parametro di sicurezza*, passato come input allo schema.

Pertanto, al variare del parametro di sicurezza, avremo una *famiglia di schemi* e, per modellarne il comportamento, avremo bisogno di una *famiglia di distribuzioni*.

Nota: in buona parte del testo tuttavia la presentazione viene semplificata e le famiglie di distribuzioni non vengono usate.

Qualche esempio di ensemble

Consideriamo l'insieme $\{0,1\}^n$ delle stringhe di n bit. L'ensemble

$$\{U_n\}_{n \in N}$$

rappresenta l'ensemble che contiene le distribuzioni di probabilità uniformi sugli insiemi di stringhe lunghe n bit

$$\begin{array}{c|cc} U_1 & 0 & 1 \\ prob & 1/2 & 1/2 \end{array}$$

$$\begin{array}{c|cccc} U_2 & 00 & 01 & 10 & 11 \\ prob & 1/4 & 1/4 & 1/4 & 1/4 \end{array}$$

$$\dots$$

$$\begin{array}{c|ccc} U_n & 00\dots0 & \dots & 11\dots1 \\ prob & 1/2^n & & 1/2^n \end{array}$$

Qualche esempio di ensemble

Un esempio invece diverso dall'ensemble uniforme è il seguente. Sia

$$\{P_n\}_{n \in N}$$

l'ensemble delle distribuzioni di probabilità su $\{0,1\}^n$ che associano probabilità uniforme alle prime 2^{n-1} stringhe e 0 alle restanti

$$\begin{array}{c|cc} P_1 & 0 & 1 \\ prob & 1 & 0 \end{array}$$

$$\begin{array}{c|cccc} P_2 & 00 & 01 & 10 & 11 \\ prob & 1/2 & 1/2 & 0 & 0 \end{array}$$

$$\dots$$

$$\begin{array}{c|cccccc} P_n & 00\dots0 & \dots & 01\dots1 & 10\dots0 & \dots & 11\dots1 \\ prob & 1/2^{n-1} & & 1/2^{n-1} & 0 & & 0 \end{array}$$