Università degli Studi dell'Insubria
Probabilità e Statistica per l'Informatica - Varese Fall-Winter 2023/24

02/02/2024 — Secondo appello

Nome e Cognome: _____ Matricola: _____

Indicazioni Importati: Siamo liberi di utilizzare appunti scritti o stampati. **Non possiamo però utilizzare dispositivi elettronici, comunicare tra di noi o con l'esterno, né passarci del materiale tra di noi**.

A meno che non venga esplicitamente detto il contrario, non è necessario calcolare il valore numerico delle soluzioni. A parte questo fatto, **è sempre importante mostrare il ragionamento e le formule usate, ed arrivare ad una risposta esatta**, anche se espressa in funzione di altri operatori (per esempio coefficienti binomiali, radici quadrate, esponenti, ecc). Un esempio: se arriviamo a una espressione del tipo $\binom{10}{5}$, possiamo lasciare questa come risposta senza ulteriori semplificazioni, oppure possiamo fare i calcoli ed arrivare al valore numerico 252. Invece scrivere solo "252" senza che sia chiaro da dove viene quel numero, **sarà considerata una risposta invalida.**

1 Rilevamento di intrusioni

La porta di uno switch è collegata a 3 host diversi, che possiamo chiamare **A**, **B** e **C**. Un sistema di *intrusion detection* cattura il traffico proveniente da ogni host e restituisce *True* o *False* a seconda che ritenga essere l'host in questione un attaccante o meno, rispettivamente. Per esempio, se il sistema ritenesse che l'host A sia un attaccante ma gli host B e C non lo siano, allora restituirebbe: (*True*, *False*, *False*)

1. Quanti sono le possibili combinazioni di output in cui due host sono considerati attacanti e uno è considerato legittimo? (1 pt)
2. Quanti sono invece tutti i possibili output che può dare il sistema? (1 pt)
3. Supponendo che tutti e tre gli host abbiano una probabilità di risultare *attaccanti* pari a **0.3**, e che gli output del sistema siano indipendenti per ogni host. Quale è la probabilità di avere la risposta (*False*, *False*, *True*) ? (1 pt)
4. Quale è la probabilità che il sistema ritenga che soltanto uno degli host è un attaccante? (3 pt)

   Suggerimento: un diagramma di Venn può aiutare a risolvere l'esercizio.
5. Ricalcolare la probabilità dell'evento precedente, ma si assuma ora che l'host **A** è indicato come attaccante con probabilità **0.9**, mentre gli host **B** e **C** continuano ad avere probabilità **0.3** di essere indicati come attaccanti. (2 pt)
6. Con le assunzioni del punto precedente, calcolate la probabilità che l'host ritenuto attaccante sia **B** *dato che* si sa che esattamente uno degli host è stato indicato come attaccante (2 pt)

1 Soluzione

1. In questo punto ci viene chiesto quante sono i possibili output in cui due host sono considerati attacanti e uno è considerato legittimo. É facile capire che sono tre:
   1. (**A** $\rightarrow$ *True*, **B** $\rightarrow$ *True*, **C** $\rightarrow$ *False*)
   2. (**A** $\rightarrow$ *True*, **B** $\rightarrow$ *False*, **C** $\rightarrow$ *True*)
   3. (**A** $\rightarrow$ *False*, **B** $\rightarrow$ *True*, **C** $\rightarrow$ *True*)

   Si può arrivare a questa risposta anche osservando che, in generale, le possibili risposte del sistema possono essere viste come la *scelta* di un elemento: Se il sistema restituisce *True*, per un determinato

host, allora lo *sceglie*. In quel caso viene naturale usare la formula delle combinazioni semplici di due elementi su un insieme di tre elementi (il numero degli host):

$$C_{3,2} = \binom{3}{2} = \frac{3!}{2! \cdot 1!} = 3$$

2. In generale, nelle risposte, è possibile che il sistema riporti zero, uno, due o tutti e tre gli host come attacanti. In questo caso dunque, dobbiamo contare il numero di possibili *combinazioni semplici* di uno, due e tre elementi presi da un insieme di tre elementi. Quindi:

$$\sum_{k \in \{0,1,2,3\}} C_{3,k} = \sum_{k \in \{0,1,2,3\}} \binom{3}{k} = \frac{3!}{0! \cdot 3!} + \frac{3!}{1! \cdot 2!} + \frac{3!}{2! \cdot 1!} + \frac{3!}{3! \cdot 0!} = 1 + 3 + 3 + 1 = 8$$

3. Denotiamo con $\mathcal{A}$ l'evento "l'host **A** è stato classificato come attaccante", e così, analogamente, usiamo $\mathcal{B}$ e $\mathcal{C}$ per le stesse inferenze riguardanti gli host **B** e **C**, rispettivamente. La traccia ci dice allora che:

$$P(\mathcal{A}) = P(\mathcal{B}) = P(\mathcal{C}) = 0.3$$

Poi, per ogni host, l'evento "non attaccante" avrà probabilità pari a **0.7**. Cioè:

$$P(\mathcal{A}^C) = P(\mathcal{B}^C) = P(\mathcal{C}^C) = 0.7$$

per la **complementarietà**. Infine, con l'ipotesi dell'indipendenza di questi eventi, avremmo che la risposta è:

$$P(\mathcal{A}^C \cap \mathcal{B}^C \cap \mathcal{C}) = P(\mathcal{A}^C) \cdot P(\mathcal{B}^C) \cdot P(\mathcal{C}) = 0.7 \cdot 0.7 \cdot 0.3 \approx 0.15$$

4. Denotiamo con $\mathcal{D}$ l'evento "esattamente uno degli host è un attaccante". Ci viene chiesto di trovare $P(\mathcal{D})$.

Osserviamo che questo evento si può verificare in tre modi diversi:

$$(\mathcal{A} \cap \mathcal{B}^C \cap \mathcal{C}^C) \implies \mathcal{D}$$
$$(\mathcal{A}^C \cap \mathcal{B} \cap \mathcal{C}^C) \implies \mathcal{D}$$
$$(\mathcal{A}^C \cap \mathcal{B}^C \cap \mathcal{C}) \implies \mathcal{D}$$

E quindi:

$$\mathcal{D} \equiv (\mathcal{A} \cap \mathcal{B}^C \cap \mathcal{C}^C) \cup (\mathcal{A}^C \cap \mathcal{B} \cap \mathcal{C}^C) \cup (\mathcal{A}^C \cap \mathcal{B}^C \cap \mathcal{C})$$

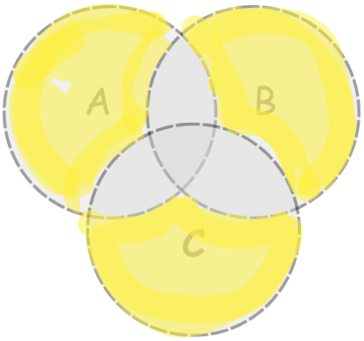
Ci viene quindi chiesto di trovare la probabilità dell'evento giallo in figura:

E cioè:

$$P(\mathcal{D}) = P((\mathcal{A} \cap \mathcal{B}^C \cap \mathcal{C}^C) \cup (\mathcal{A}^C \cap \mathcal{B} \cap \mathcal{C}^C) \cup (\mathcal{A}^C \cap \mathcal{B}^C \cap \mathcal{C}))$$

Essendo poi chiaramente eventi non compatibili, abbiamo che:

$$P(\mathcal{D}) = P(\mathcal{A} \cap \mathcal{B}^C \cap \mathcal{C}^C) + P(\mathcal{A}^C \cap \mathcal{B} \cap \mathcal{C}^C) + P(\mathcal{A}^C \cap \mathcal{B}^C \cap \mathcal{C})$$
$$P(\mathcal{D}) = 0.3 \cdot 0.7 \cdot 0.7 + 0.7 \cdot 0.3 \cdot 0.7 + 0.7 \cdot 0.7 \cdot 0.3 \approx 0.44$$



Si potrebbe esser giunti allo stesso risultato considerando una v.a. Binomiale che indichi il numero di host considerati attaccanti su tre host analizzati:

$$\mathcal{X} \sim \text{Binomiale}(n = 3, p = 0.3)$$

e così:

$$P(\mathcal{X} = 1) = \binom{3}{1} \cdot 0.3 \cdot 0.7^2 \approx 0.44$$

5. Questa volta avremmo:

$$P(\mathcal{A}) = 0.9 \text{ e } P(\mathcal{A}^C) = 0.1$$

Quindi:

$$P(\mathcal{D}) = P(\mathcal{A} \cap \mathcal{B}^C \cap \mathcal{C}^C) + P(\mathcal{A}^C \cap \mathcal{B} \cap \mathcal{C}^C) + P(\mathcal{A}^C \cap \mathcal{B}^C \cap \mathcal{C})$$
$$P(\mathcal{D}) = 0.9 \cdot 0.7 \cdot 0.7 + 0.1 \cdot 0.3 \cdot 0.7 + 0.1 \cdot 0.7 \cdot 0.3 \approx 0.48$$

6. Le assunzioni del punto precedente implicano $P(\mathcal{D}) \approx 0,48$. Con queste assunzioni ci viene chiesto: $P(\mathcal{B}|\mathcal{D})$.

   Dalla definizione di prob. condizionata sappiamo che:

$$P(\mathcal{B}|\mathcal{D}) = \frac{P(\mathcal{B} \cap \mathcal{D})}{P(\mathcal{D})}$$

   E, in questo caso concreto, sappiamo che:

$$\frac{P(\mathcal{B} \cap \mathcal{D})}{P(\mathcal{D})} = \frac{P(\mathcal{A}^C \cap \mathcal{B} \cap \mathcal{C}^C)}{P(\mathcal{D})} \approx \frac{0.1 \cdot 0.3 \cdot 0.7}{P(\mathcal{D})} \approx 0.04$$

2 Classificazione di cyber-attacchi

I sistemi informativi di una certa banca hanno riportato l'arrivo medio di **240** tentativi di attacchi al giorno nel loro server di posta. Si può supporre che ogni attacco sia indipendente dagli altri e che non ci sia un limite massimo al numero di attacchi possibili.

1. Quale è il numero medio di attacchi ogni ora? (1pt)
2. Nelle ultime **2** ore si sono verificati **23** attacchi, ciascuno da un'associazione criminale diversa. Quante sono le possibili permutazioni che descrivono l'ordine in cui ogni gruppo aggressore ha cercato di attaccare l'azienda? (1pt)

3. Gli analisti della banca hanno rilevato **tre tipi** di attacchi: i *Denial-of-Service* (DoS), i *Malware*. e altre anomalie sconosciute. Negli ultimi **23** attacchi, **10** attacchi erano di tipo Dos, **5** di tipo Malware e **8** erano anomalie sconosciute. Quante sono le possibili permutazioni di questi **23** attacchi, considerando gli attacchi dello stesso tipo come indistinguibili? (2pt)

4. Si assuma che, in generale il **70%** degli attacchi sono DoS, il **20%** corrisponde al Malware, e il **10%** restante è di tipo sconosciuto. Quale è la probabilità che, nei prossimi **10** attacchi ce ne siano **7** di tipo sconosciuto? (Si assuma che il tipo di ogni attacco è indipendente dal tipo degli attacchi precedenti). (1pt)

5. Il **10%** delle richieste in entrata al server della banca è un attacco, mentre il resto è traffico legittimo. Con le assunzioni del punto precedente riguardo la distribuzione di attacchi per categoria, quale è la probabilità che, nelle prossime **20** richieste ce ne siano **7** attacchi DoS? (1pt)

   Suggerimento per questo punto e il punto precedente: Può aiutare l'utilizzo della distribuzione Binomiale.

6. Per ogni richiesta in entrata alla banca, un classificatore di attacchi cerca di classificarla in una delle quattro classi (i tre tipi di attacco oppure richiesta legittima). Tale strumento funziona con un'accuratezza del 80%. Cioè, per ognuna delle classi, il 20% delle volte dà risposte sbagliate. Quanti tentativi in media dovremmo aspettare per avere la prima classificazione corretta di un attacco di tipo Malware? (2pt)

7. Quale è la probabilità di dover aspettare esattamente 10 tentativi per avere la prima classificazione corretta di traffico legittimo? (2pt)

2 Soluzione

1. Scriviamo $\mathcal{X}$ per la variabile aleatoria che conta il numero di attacchi informatici in arrivo in un giorno. Avremmo che $\mathcal{X}$ segue una distribuzione di Poisson di parametro $\lambda = 240$, dove λ indica il numero medio di attacchi in 24 ore. Per capire quanti attacchi in media abbiamo ogni ora, dividiamo λ per 24 e troviamo 10.

2. Usiamo in questo caso la formula delle permutazioni, cioè $n!$, e lasciamo scritta la risposta come $23!$

3. Questa volta usiamo la formula delle permutazioni con elementi indistiguibili. Abbiamo tre classi di elementi, con cardinalità dieci, cinque e otto, rispettivamente, quindi la risposta è:

$$\frac{23!}{10! \cdot 5! \cdot 8!}$$

4. Usiamo $\mathcal{Y}$ per la variabile che conta il numero di attacchi di tipo sconosciuto nei prossimi 10 attacchi. Avremmo che essa è distribuita come una binomiale:

$$\mathcal{Y} \sim \text{Binomiale}(n = 10, p = 0.1)$$

   In questo caso avremmo quindi che:

$$P(\mathcal{Y} = 7) = \binom{10}{7} \cdot 0.1^7 \cdot 0.9^3$$

5. Usiamo $\mathcal{X}$ per la variabile che conta il numero di attacchi di tipo DoS nelle prossime 20 richieste. Avremmo che essa è distribuita come una binomiale:

$$\mathcal{Y} \sim \text{Binomiale}(n = 20, p = 0.07)$$

Dove il parametro p corrisponde alla probabilità di avere un attacco di tipo DoS $P(\text{attacco} \cap \text{DoS})$ usando la definizione di probabilità condizionata:

$$P(\text{attacco} \cap \text{DoS}) = P(\text{DoS}|\text{attacco}) \cdot P(\text{attacco}) = 0.7 \cdot 0.1$$

Avremmo quindi che:

$$P(\mathcal{X} = 7) = \binom{20}{7} \cdot 0.07^7 \cdot 0.93^{13}$$

6. Per risolvere questo punto dobbiamo tenere in conto adesso l'arrivo di richieste legittime. Denotiamo con **DoS**, **Mw** e **Sc** gli eventi "arrivo di un attacco di tipo DoS, Malware, e Sconosciuto, rispettivamente. Denotiamo con **Ben** l'evento "arrivo di una richiesta legittima. Infine, denottiamo con **Corr** l'evento "classificazione corretta". Noi sappiamo che $P(\textbf{Corr}) = 0.8$.

   La traccia ci dice che il 10% delle richieste che arrivano è un'attacco. Quindi, in termini assoluti, le probabilità di attacco diventano:

$$P(\textbf{DoS}) = 0.1 \cdot 0.7 \qquad P(\textbf{Mw}) = 0.1 \cdot 0.2 \qquad P(\textbf{Sc}) = 0.1 \cdot 0.1$$

   Usiamo la v.a. Geometrica, sapendo che l'evento "classificazione corretta di un attacco di tipo Malware", $\hat{\textbf{Mw}}$, si verificherà con probabilità

$$p = P(\hat{\textbf{Mw}}) = P(\textbf{Corr} \cap \textbf{Mw}) = P(\textbf{Corr}) \cdot P(\textbf{Mw}) = 0.8 \cdot 0.02 = 0.016$$

   Usiamo poi la v.a. $\mathcal{W}_{\textbf{Mw}}$ per contare il numero di richieste in entrata alla banca in corrispondenza del primo successo di classificazione di un malware, avremmo:

$$\mathcal{W}_{\textbf{Mw}} \sim Geometrica(p = 0.016)$$

$$P(\mathcal{W}_{\textbf{Mw}} = k) = 0.016 \cdot (1 - 0.016)^{k-1}$$

   Dalle proprietà della distribuzione Geometrica, sappiamo quindi che la risposta è:

$$\mathbb{E}[\mathcal{W}_{\textbf{Mw}}] = \frac{1}{p} = \frac{1}{0.016} = 62.5$$

7. In modo simile al punto precedente, avremmo:

$$P(\hat{\textbf{Ben}}) = P(\textbf{Corr} \cap \textbf{Ben}) = P(\textbf{Corr}) \cdot P(\textbf{Ben}) = 0.8 \cdot 0.9 = 0.072$$

$$P(\mathcal{W}_{\textbf{Ben}} = k) = 0.072 \cdot (1 - 0.072)^{k-1}$$

$$P(\mathcal{W}_{\textbf{Ben}} = 10) = 0.072 \cdot (1 - 0.072)^{9}$$

3 Dado truccato

1. Un dado a sei facce truccato ha la seguente distribuzione di probabilità al risultato di ogni lancio:

$$P(X = x) = \begin{cases} 1/12 & \text{se} \quad x < 4 \\ 3/12 & \text{se} \quad x >= 4 \end{cases}$$

   Calcolare il valore atteso di questa distribuzione (1pt)

2. La varianza di tale distribuzione è 2.35. Scrivere l'equazione che ci porta a quel risultato. (1pt)
3. Quale è la media della var. aleatoria S_{20} che indica la somma di 20 lanci di questo dado? (1pt)
4. Quale è la deviazione standard di S_{20}? (1pt)
5. Si indichi la distribuzione che avrà la variabile aleatoria media campionaria: $\frac{S_{20}}{20}$. (1pt)
6. Si indichi una variabile aleatoria Gaussiana standard in funzione della media campionaria. (Cioè, come si può normalizzare la media campionaria?) (1pt)
7. (2pt) Con quanta probabilità si avrà che tale media campionaria sarà compresa nell'intervallo

$$[\mathbb{E}[X] - 0.5, \mathbb{E}[X] + 0.5]$$

   Suggerimenti: ci viene chiesto di calcolare la seguente probabilità:

$$P(\frac{S_{20}}{20} > \mathbb{E}[X] - 0.5, \frac{S_{20}}{20} < \mathbb{E}[X] + 0.5)$$

   Può essere utile sapere che: $\frac{0.5}{\sqrt{\frac{2.35}{20}}} \approx 1.46$.

8. Quanto dev'essere grande n affinchè $|\frac{S_n}{n} - \mathbb{E}[X]| < 0,5$ con probabilità del 95%? (3pt)

3 Soluzione

1. Se usiamo la definizione di valore atteso:

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x \in \{1,2,\ldots,6\}} x \cdot P(X = x) \\ &= 1 \cdot \frac{1}{12} + 2 \cdot \frac{1}{12} + 3 \cdot \frac{1}{12} + 4 \cdot \frac{3}{12} + 5 \cdot \frac{3}{12} + 6 \cdot \frac{3}{12} \\ &= \frac{6}{12} + \frac{12}{12} + \frac{15}{12} + \frac{18}{12} \\ &= \frac{51}{12} \end{aligned}$$

2. La formula della varianza invece è:

$$\begin{aligned} Var[X] &= \sum_{x \in \{1,2,\ldots,6\}} (x - \mathbb{E}[X])^2 \cdot P(X = x) \\ &= (1 - \mathbb{E}[X])^2 \cdot \frac{1}{12} + (2 - \mathbb{E}[X])^2 \cdot \frac{1}{12} + (3 - \mathbb{E}[X])^2 \cdot \frac{1}{12} + \\ &\qquad (4 - \mathbb{E}[X])^2 \cdot \frac{3}{12} + (5 - \mathbb{E}[X])^2 \cdot \frac{3}{12} + (6 - \mathbb{E}[X])^2 \cdot \frac{3}{12} \end{aligned}$$

Facendo i calcoli otteniamo $Var[X] = 2.354$

E' anche considerata giusta la risposta:

$$Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

sebbene questa non sia la defizione di varianza ma una conseguenza di essa.

3. Per il Teorema del Limite Centrale io so che la variabile $S_{20} = X_1 + X_2 + ... + X_20$, che somma il punteggio ottenuto lanciando il dado 20 volte, ha una media pari a $20 \cdot \mathbb{E}[X] = 20 \cdot \frac{51}{12} = 85$

4. Per il Teorema del Limite Centrale io so che la variabile $S_{20} = X_1 + X_2 + ... + X_{20}$, che somma il punteggio ottenuto lanciando il dado 20 volte, ha una deviazione standard pari a $\sqrt{20 \cdot Var[X]} = \sqrt{20 \cdot 2.354} = 6.86$

5. Sempre per il TLC, sappiamo che $\frac{S_{20}}{20} \sim \mathcal{N}(\mu = \mathbb{E}[X], \sigma = \sqrt{\frac{Var[X]}{n}})$.

6. Per normalizzare $\frac{S_{20}}{20}$ sottraiamo la sua media e dividiamo per la sua deviazione standard, che abbiamo trovato prima:

$$\frac{\frac{S_{20}}{20} - \mathbb{E}[X]}{\sqrt{\frac{Var[X]}{n}}} \sim \mathcal{N}(0, 1)$$

7. Ci viene chiesta la seguente probabilità:

$$P(\frac{S_{20}}{20} > \mathbb{E}[X] - 0.5, \frac{S_{20}}{20} < \mathbb{E}[X] + 0.5)$$

che è equivalente a:

$$P(-0.5 < \frac{S_{20}}{20} - \mathbb{E}[X] < 0.5)$$

L'espressione in mezzo alla disequazione può essere ricondotta ad una v.a. Gaussiana Standard dividendo per la deviazione standard di $\frac{S_{20}}{20}$:

$$P(\frac{-0.5}{\sqrt{\frac{Var[X]}{n}}} < \frac{\frac{S_{20}}{20} - \mathbb{E}[X]}{\sqrt{\frac{Var[X]}{n}}} < \frac{0.5}{\sqrt{\frac{Var[X]}{n}}}) = \mathbf{\Phi}(\frac{0.5}{\sqrt{\frac{Var[X]}{n}}}) - \mathbf{\Phi}(\frac{-0.5}{\sqrt{\frac{Var[X]}{n}}})$$

Dal suggerimento abbiamo poi che:

$$\mathbf{\Phi}(\frac{0.5}{\sqrt{\frac{Var[X]}{n}}}) - \mathbf{\Phi}(\frac{-0.5}{\sqrt{\frac{Var[X]}{n}}}) = \mathbf{\Phi}(1.46) - \mathbf{\Phi}(-1.46)$$

E ciò equivale a:

$$2\mathbf{\Phi}(1.46) - 1 = 2 \cdot 0.93 - 1 \approx 0.86$$

8. La domanda ci chiede il valore di n tale che:

$$P(-0.5 < \frac{S_n}{n} - \mathbb{E}[X] < 0.5) = 0.95$$

Con lo stesso ragionamento del punto precedente abbiamo:

$$P(\frac{-0.5}{\sqrt{\frac{Var[X]}{n}}} < \frac{\frac{S_n}{n} - \mathbb{E}[X]}{\sqrt{\frac{Var[X]}{n}}} < \frac{0.5}{\sqrt{\frac{Var[X]}{n}}}) = \mathbf{\Phi}(\frac{0.5}{\sqrt{\frac{Var[X]}{n}}}) - \mathbf{\Phi}(\frac{-0.5}{\sqrt{\frac{Var[X]}{n}}})$$

Dobbiamo risolvere l'equazione:

$$\mathbf{\Phi}(\frac{0.5}{\sqrt{\frac{Var[X]}{n}}}) - \mathbf{\Phi}(\frac{-0.5}{\sqrt{\frac{Var[X]}{n}}}) = 0.95$$

Che è equivalente a:

$$2\mathbf{\Phi}(\frac{0.5}{\sqrt{\frac{Var[X]}{n}}}) - 1 = 0.95$$

E quindi:

$$\mathbf{\Phi}(\frac{0.5}{\sqrt{\frac{Var[X]}{n}}}) = 0.975$$

Da cui, facendo un lookup inverso sulla tabella $\mathbf{\Phi}$ avrò:

$$\frac{0.5}{\sqrt{\frac{Var[X]}{n}}} = 1.96 \quad \rightarrow \quad \frac{0.5}{\sqrt{\frac{2.35}{n}}} = 1.96 \quad \rightarrow \quad \frac{0.5}{\sqrt{\frac{2.35}{n}}} = 1.96$$

e cioè:

$$\frac{0.5}{1.96} = \sqrt{\frac{2.35}{n}} \quad \rightarrow \quad (\frac{0.5}{1.96})^2 = \frac{2.35}{n} \quad \rightarrow \quad n = \frac{2.35}{(\frac{0.5}{1.96})^2}$$