Teorema di Bayes

13 aprile 2017

Si introduce l'equazione di partizione o regola delle probabilitá totali. Successivamente si discute il teorema di Bayes.

1 *Regola delle probabilitá totali*

Uno dei prerequisiti per formulare il teorema di Bayes é la regola delle probabilitá totali, conosciuta anche come equazione di partizione.

Teorema 1.1 (Equazione di partizione o regola delle probabilitá totali) *Sia $(S, \mathcal{F}, P)$ il nostro spazio di probabilitá e sia $(A_1, A_2, \cdots, A_n)$ una successione finita o numerabile di eventi che partizionano S. Allora se $B \subseteq S$ con $B \in \mathcal{F}$:*

$$P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B \mid A_i). \tag{1}$$

Dimostrazione *Si ha che*

$$B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i).$$

Poiché per definizione di partizione $A_i \cap A_j = \emptyset, \quad i \neq j$, allora

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n) = \\ &= P(A_1) \cdot P(B \mid A_1) + P(A_2) \cdot P(B \mid A_2) + \cdots + P(A_n) \cdot P(B \mid A_n) = \\ &= \sum_{i=1}^{n} P(A_i) \cdot P(B \mid A_i). \end{aligned}$$

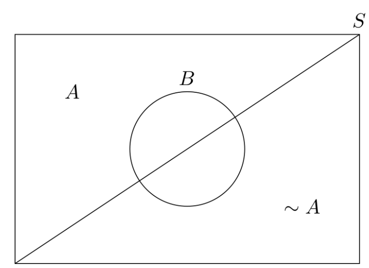

Figura 1: Spazio campionario partizionato in due eventi

Esempio 1.2 *Consideriamo il caso elementare in cui S risulti diviso in due partizioni A e $\sim A$, con B sottoinsieme di S, come mostrato in Figura 1, dunque valgono le seguenti:*

$$S = A \cup \sim A$$

$$B = (B \cap A) \cup (B \cap \sim A)$$

Applicando ora la misura di probabilitá ottieniamo,

$$P(B) = P(B \cap A) + P(B \cap \sim A).$$

E per la regola del prodotto:

$$P(B) = P(A) \cdot P(B \mid A) + P(\sim A) \cdot P(B \mid \sim A).$$

Mostriamo in Figura 2, la regola delle probabilitá totali (2) in forma di albero delle probabilitá (chance tree) per chiarire ancora meglio questo semplice esempio.

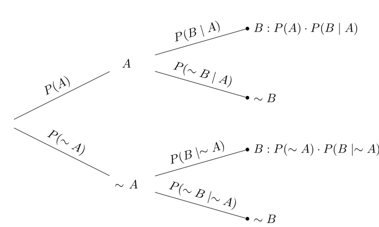

Figura 2: L'equazione di partizione esplicitata come albero di probabilitá sullo spazio degli stati di Figura 1

Vediamo ora alcuni esempi di applicazione della regola (2).

Esempio 1.3 *Abbiamo una moneta sbilanciata e due urne contenenti palline bianche e nere, lanciamo la moneta e sulla base dell'esito peschiamo da un'urna piuttosto che dall'altra, vogliamo predire la probabilitá di pescare una pallina nera sulla base del seguente modello generativo (Tabella 1):*

Tabella 1: Modello generativo

Moneta Sbilanciata	**Urna 1**	**Urna 2**
-	5 palline, 3 nere e 2 bianche	5 palline, 4 nere e 1 bianca
$P(T) = \frac{4}{10}$	$P(N) = \frac{3}{5}$	$P(N) = \frac{4}{5}$
$P(C) = \frac{6}{10}$	$P(B) = \frac{2}{5}$	$P(B) = \frac{1}{5}$

Dopo aver osservato che la probabilitá di pescare dall'urna 1 equivale alla probabilitá che esca testa sulla moneta e la probabilitá di pescare dall'urna 2 equivale alla probabilitá che esca croce ($P(U_1) = P(T)$ e $P(U_2) = P(C)$) costruiamo il seguente chance tree (Figura 3):

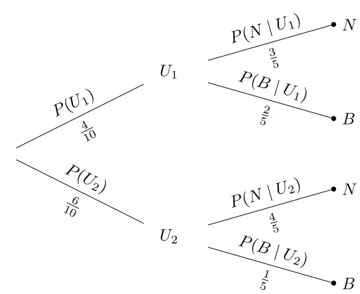

Figura 3: Chance tree per il problema delle monete nelle urne

L'evento di nostro interesse é N, applichiamo quindi la regola delle probabilitá totali (2) per trovare $P(N)$.

$$
\begin{aligned}
P(N) &= P(U_1) \cdot P(N \mid U_1) + P(U_2) \cdot P(N \mid U_2) = \\
&= \frac{4}{10} \cdot \frac{3}{5} + \frac{6}{10} \cdot \frac{4}{5} = \frac{12}{50} + \frac{12}{25} = \frac{18}{25}
\end{aligned} \tag{2}
$$

Esempio 1.4 *Un esperimento prevede il lancio di una moneta, scelta fra un insieme di 3 monete con probabilitá uniforme: $P(A) = P(B) = P(C) = \frac{1}{3}$. Ciascuna moneta genera testa o croce secondo il seguente modello*

Tabella 2: Modello generativo delle tre monete

Moneta A	**Moneta B**	**Moneta C**
$P(T \mid A) = 1$	$P(T \mid B) = \frac{1}{2}$	$P(T \mid C) = 0$

Dunque solo la moneta B é bilanciata.

Vogliamo predire la probabilitá complessiva associata all'evento "esce testa".

Utilizziamo l'equazione di partizione (2):

$$
\begin{aligned}
P(T) &= P(A) \cdot P(T \mid A) + P(B) \cdot P(T \mid B) + P(C) \cdot P(T \mid C) = \\
&= \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot 0 = \frac{1}{3} + \frac{1}{6} = \frac{1}{2}
\end{aligned} \tag{3}
$$

Come ci aspettavamo, A e C si "annullano" nei loro esiti ed essendo sotto ipotesi di probabilitá uniforme, la nostra $P(T)$ equivale alla $P(T)$ di qualunque moneta bilanciata.

Esempio 1.5 *Vogliamo calcolare la probabilitá di vittoria in un torneo di scacchi considerando che esistono tre tipologie di giocatori di diversa bravura iscritti al torneo.*

Sulla base delle iscrizioni, la metá dei giocatori iscritti al torneo é di tipologia 1, $1/4$ di tipologia 2, e i restanti di tipologia 3.

La Tabella 3 ci fornisce probabilitá di vittoria contro ciascun tipo di giocatore

T_1	T_2	T_3
50%	25%	25%
$P(V\|T_1) = 0.3$	$P(V\|T_2) = 0.4$	$P(V\|T_3) = 0.5$

Tabella 3: Modello generativo per l'esempio 4

Possiamo effettuare una predizione sulla probabilitá complessiva di vincere usando l'equazione di partizione (4).

$$\begin{aligned} P(V) &= P(T_1) \cdot P(V \mid T_1) + P(T_2) \cdot P(V \mid T_2) + P(T_3) \cdot P(V \mid T_3) = \\ &= 0.5 \cdot 0.3 + 0.25 \cdot 0.4 + 0.25 \cdot 0.5 = 0.375 \end{aligned} \tag{4}$$

Vediamo ora un esempio piú complesso che illustra

- l'aspetto predittivo del teorema delle probabilitá totali;
- una formulazione iterativa, che consente di considerare predizioni successive come il risultato dell'evoluzione (temporale) di un sistema dinamico.

Esempio 1.6 *Consideriamo i due eventi complementari A_i e $\sim A_i = I_i$ che significano rispettivamente* "`sono avanti con la preparazione del corso di Statistica al tempo i`" *e* "`sono indietro con la preparazione del corso di Statistica al tempo i`".

Sia dato il seguente modello generativo (formulato, per esempio, dopo aver osservato nel tempo comportamenti e risultati di un gruppo di studenti frequentanti il corso) rappresentato con un chance tree in Figura 4):

Supponiamo ora di essere interessati a scoprire la probabilitá di essere avanti con la preparazione in un tempo k, applichiamo quindi la (2):

$$\begin{aligned} P(A_{i+1}) &= P(A_i) \cdot P(A_{i+1} \mid A_i) + P(I_i) \cdot P(A_{i+1} \mid I_i) = \\ &= 0.8 \cdot P(A_i) + 0.4 \cdot P(I_i) = 0.4 \cdot (2 \cdot P(A_i) + P(I_i)) = \\ &= 0.4 \cdot (2 \cdot P(A_i) + P(\sim A_i)) = 0.4 \cdot (2 \cdot P(A_i) + 1 - P(A_i)) = \\ &= 0.4P(A_i) + 0.4. \end{aligned} \tag{5}$$

*Quello che abbiamo appena ottenuto in (5) applicando (2) prende il nome di **definizione iterativa o ricorsiva del modello probabilistico generativo**. Per procedere verso un calcolo effettivo di $P(A_k)$ attraverso l'equazione (5) occorre fissare la base della ricorsione, ovvero $P(A_1)$.*

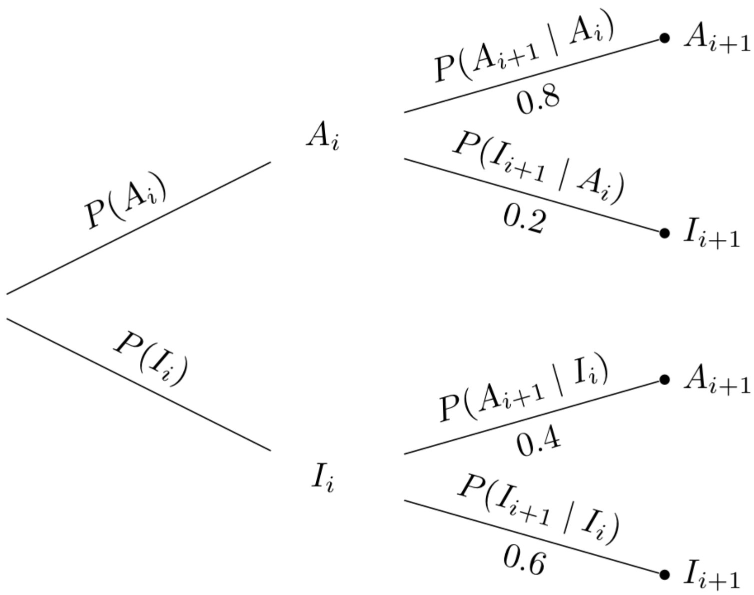

Figura 4: Chance tree che formalizza l'evoluzione del livello di preparazione dell'esame di Statistica dal tempo i al tempo $i+1$

Fissiamo, per esempio, $P(A_1) = 0.8$ *e* $k = 3$*, qui sotto riportiamo i calcoli (6) che ci permettono di calcolare* $P(A_3)$

$$\begin{aligned} P(A_3) &= 0.4 \cdot (1 + P(A_2)) = 0.4 \cdot (1 + 0.4 \cdot (1 + P(A_1))) = \\ &= 0.4 \cdot (1 + 0.4 \cdot (1 + 0.8)) = 0.688 \end{aligned} \tag{6}$$

Possiamo ora introdurre la regola di Bayes.

2 *Teorema o regola di Bayes*

La regola di Bayes ha trovato numerose applicazioni in campo diagnostico

Esempio 2.1 *Sia B l'evento di interesse: una "macchia" sulla radiografia (**effetto**) osservabile dal radiologo. Il problema del radiologo é effettuare la diagnosi, ovvero inferire la possibile **causa** della macchia.*

*Lo spazio campionario relativo al problema della diagnosi (Figura 5) é partizionato nelle seguenti **cause**:*

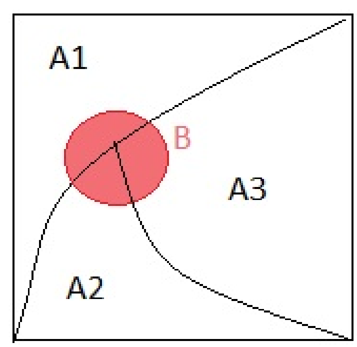

Figura 5: Spazio campionario relativo al problema della diagnosi

- *A1 = tumori maligni*
- *A2 = tumori benigni*
- *A3 = altre malattie*

Allora il modello generativo (**cause** $\rightarrow$ **effetto**) mi consente di calcolare la probabilitá di B ovvero di effettuare una **predizione**. Sostanzialmente, si considerano nel chance tree (Figura **??**) solo i cammini che portano a B e non si considerano quelli che portano a $\sim B$ e si usa la probabilitá totale:

$$P(B) = P(A1)P(B \mid A1) + P(A2)P(B \mid A2) + P(A3)P(B \mid A3)$$

Il problema della diagnosi é dunque quello di invertire il modello generativo per inferire la causa piú probabile (Figura 7): **cause** $\leftarrow$ **effetto**

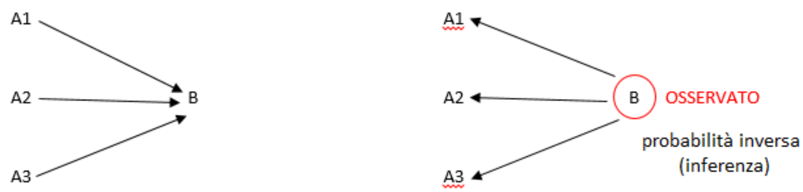

Figura 7: Inversione del modello generativo per inferire le cause

Consideriamo lo stesso problema (inferenza) nel caso semplice della selezione (cieca) di una moneta e della conoscenza dell'esito del risultato.

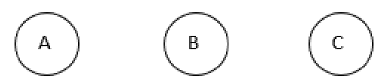

Figura 8: Tre monete con diverso bilanciamento

Esempio 2.2 (Lancio delle monete) *Date tre monete A, B, C in un'urna tali che:*

- $P(T \mid A) = 1$ *(sbilanciata),*
- $P(T \mid B) = \frac{1}{2}$ *(bilanciata),*
- $P(T \mid C) = 0$ *(sbilanciata).*

Una moneta viene estratta a caso con probabilitá uniforme:

$$P(A) = P(B) = P(C) = \frac{1}{3}$$

Prima di lanciare la moneta e osservare l'esito del lancio, sulla base delle informazioni il modello generativo rappresentato in Figura 9 ci consente di predire che $P(T) = \frac{1}{2}$

L'esperimento viene effettuato: chi ha estratto la moneta la lancia e fornisce la seguente informazione: "É uscita testa".

Problema*: osservato l'esito dell'esperimento, su quale moneta sareste disposti a scommettere?*

*Il problema é inferenziale (**moneta** $\leftarrow$ **esito del lancio**: scommetterete sulla moneta che massimizza la seguente probabilitá:*

$$P(\texttt{moneta} \mid \texttt{esito lancio} = T)$$

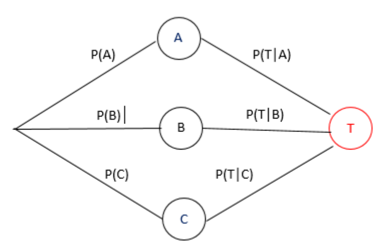

Figura 9: Chance tree ridotto dopo l'osservazione dell'esito "testa"

Intuitivamente, per scommettere sulla moneta A, con riferimento all'albero di possibilitá in Figura 9 pesiamo la probabilitá di percorrere il ramo $A \rightarrow T$ rispetto a tutti i cammini possibili:

$$P(A \mid \texttt{esito lancio} = T) = \frac{Prob(A \rightarrow T)}{Prob(A \rightarrow T) + Prob(B \rightarrow T)Prob(C \rightarrow T)} = \frac{Prob(A \rightarrow T)}{P(T)}$$

*Chiameremo la $P(A \mid T)$, **probabilitá a posteriori** di aver lanciato la moneta A dopo aver osservato l'esito T, ovvero:*

$$P(A \mid T) = \frac{P(A) \cdot P(T \mid A)}{P(A) \cdot P(T \mid A) + P(B) \cdot P(T \mid B) + P(C) \cdot P(T \mid C)} = \frac{\frac{1}{3} \cdot 1}{P(T)} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

La stessa probabilitá puó essere calcolata per B, C:

$$P(B \mid T) = \frac{P(B) \cdot P(T \mid B)}{P(A) \cdot P(T \mid A) + P(B) \cdot P(T \mid B) + P(C) \cdot P(T \mid C)} = \frac{\frac{1}{3} \cdot \frac{1}{2}}{P(T)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

$$P(C \mid T) = \frac{P(C) \cdot P(T \mid C)}{P(A) \cdot P(T \mid A) + P(B) \cdot P(T \mid B) + P(C) \cdot P(T \mid C)} = \frac{\frac{1}{3} \cdot 0}{P(T)} = \frac{0}{\frac{1}{2}} = 0$$

Notiamo il seguente fatto (cfr., Figura 10):

- *Prima di osservare l'esito dell'esperimento (**probabilitá a priori**):*

$$P(A) = \frac{1}{3}, P(B) = \frac{1}{3}, P(C) = \frac{1}{3}$$

- *Dopo aver lanciato (**probabilitá a posteriori**):*

$$P(A \mid T) = \frac{2}{3}, P(B \mid T) = \frac{1}{3}, P(C \mid T) = 0$$

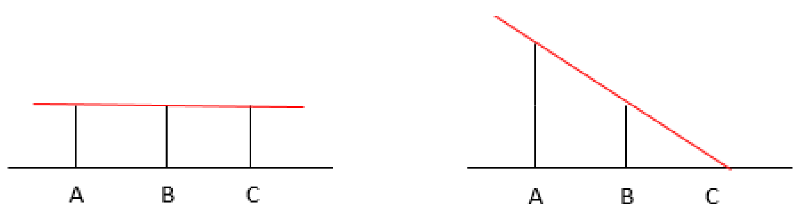

Figura 10: Probabilitá di aver utilizzato la moneta A, B, C *prima* e *dopo* l'osservazione dell'esito dell'esperimento. La linea rossa serve solo a "guidare lo sguardo".

In sintesi, abbiamo applicato per inferire dall'osservazione la moneta piú probabile, il seguente

Teorema 2.3 (Regola di Bayes) *Siano $E_1, E_2, ..., E_n$ eventi che partizionano S. Considerato l'evento $B \subseteq S$, e supposto $P(B) > 0$, allora:*

$$P(E_i \mid B) = \frac{P(B \mid E_i)P(E_i)}{\sum_{i=1}^{n} P(B \mid E_i)P(E_i)} \tag{7}$$

Dimostrazione

Per la definizione di probabilitá condizionata:

$$P(E_i \mid B) = \frac{P(E_i \cap B)}{P(B)}$$

Per la regola del prodotto:

$$= \frac{P(E_i) \cdot P(B \mid E_i)}{P(B)}$$

Per la regola della probabilitá totale:

$$= \frac{P(E_i)P(B \mid E_i)}{\sum_{i=1}^{n} P(B \mid E_i)P(E_i)}$$

2.1 *Un'interpretazione piú generale della Regola di Bayes*

La regola di Bayes ha una valenza piú generale se si assumono le seguenti:

- $\mathcal{H}_1, \mathcal{H}_2, \cdots, \mathcal{H}_n$: ipotesi o teorie (in competizione fra di loro)
- $\mathcal{D}$: dati/osservazioni

La (7) si puó pertanto riscrivere

$$P(\mathcal{H}_i \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{H}_i)P(\mathcal{H}_i)}{\sum_{i=1}^{n} P(\mathcal{D} \mid \mathcal{H}_i)P(\mathcal{H}_i)} \tag{8}$$

dove si possono fattorizzare le seguenti componenti

- PROBABILITÀ A POSTERIORI (dell'ipotesi): $P(\mathcal{H}_i \mid \mathcal{D})$
- FUNZIONE DI VEROSIMIGLIANZA (dei dati, likelihood): $P(\mathcal{D} \mid \mathcal{H}_i)$, a volte anche scritta come $f(\mathcal{D}; \mathcal{H}_i)$, per evidenziare il fatto che puó assumere valori al di fuori dell'intervallo $[0, 1]$
- PROBABILITÁ A PRIORI (sulle ipotesi): $P(\mathcal{H}_i)$;

- EVIDENZA (dei dati): $P(\mathcal{D})$, a volte detta funzione di partizione.

L'equazione (8) sintetizza in modo generale il processo di inferenza scientifica: considerato un'insieme di ipotesi o teorie $\{\mathcal{H}_i\}_{i=1}^n$ in competizione fra loro, con plausibilitá $P(\mathcal{H}_i)$ sulla base di conoscenza a priori, dopo aver effettuato un esperimento che misura la verosimiglianza $P(\mathcal{D} \mid \mathcal{H}_i)$ dell'ipotesi $\mathcal{H}_i$ nello spiegare i dati $\mathcal{D}$, ciascuna ipotesi assume, a posteriori, plausibilitá $P(\mathcal{H}_i \mid \mathcal{D})$.

Osservazione 2.4 *E' importante osservare la doppia valenza di $P(\mathcal{D} \mid \mathcal{H}_i)$:*

- *fissata l'ipotesi $\mathcal{H}_i$ puó essere letta come la probabilitá condizionata di osservare $\mathcal{D}$ sotto l'ipotesi $\mathcal{H}_i$;*
- *fissati i dati $\mathcal{D}$, ovvero avendo osservato gli esiti di un esperimento, é una funzione $f(\mathcal{D};\mathcal{H}_i)$ che, al variare di $\mathcal{H}_i$, descrive la verosimiglianza (likelihood) delle ipotesi sulla base dei dati a disposizione.*

Vediamo alcuni esempi di applicazione.

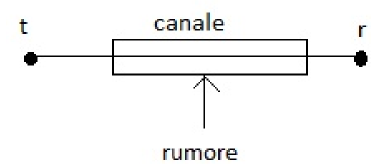

Figura 11: Un canale binario dove il trasmettitore o sorgente t puó trasmettere 1 ($t \to 1$) o 0 ($t \to 0$) e il ricevitore r riceve 1 ($r \leftarrow 1$) o 0 ($r \leftarrow 0$). Il canale di trasmissione é soggetto a rumore che rende piú o meno incerto l'esito della trasmissione, con possibilitá del "flip" del simbolo trasmesso

.

Esempio 2.5 (Canale binario) *Si consideri il canale binario rappresentato in Figura 11*

Definiamo la probabilitá di emissione alla sorgente (trasmettitore t)

$$P(t \to 0) = p$$

$$P(t \to 1) = q = 1 - p$$

La probabilitá di trasmissione corretta di un bit pari a 0 *é formalizzata come:*

$$P(r \leftarrow 0 \mid t \to 0)$$

e la probabilitá di trasmissione corretta di un bit pari a 1

$$P(r \leftarrow 1 \mid t \to 1)$$

Ovviamente:

$$P(r \leftarrow 0 \mid t \to 1) = 1 - P(r \leftarrow 0 \mid t \to 0)$$

$$P(r \leftarrow 0 \mid t \to 1) = 1 - P(r \leftarrow 1 \mid t \to 1)$$

Per effettuare la predizione $P(r \leftarrow 1)$ si usa l'equazione di partizione:

$$P(r \leftarrow 1) = P(t \to 1) \cdot P(r \leftarrow 1 \mid t \to 1) + P(t \to 0) \cdot P(r \leftarrow 1 \mid t \to 0)$$

Supponiamo ora di osservare al ricevitore $r \leftarrow 1$

Sappiamo statisticamente che 60% dei bit trasmessi dalla sorgente sono 1*: questo é il modello del segnale.*

Inoltre, l'errore statistico sul canale (commutazone di bit) é pari al 5% (probabilitá di sbagliare 0.05). Questo rappresenta il modello del canale.

Dopo aver ricevuto il bit 1*, qual é la probabilitá che sia stato effettivamente emesso* 1 *alla sorgente?*

Applicando Bayes:

$$P(t \to 1 \mid r \leftarrow 1) = \frac{P(r \leftarrow 1 \mid t \to 1) \cdot P(t \to 1)}{P(r \leftarrow 1)} = 0.97$$

Esempio 2.6 (Vasca dei pesci a.k.a. Sushi Delight) *C'é una vasca contenente un pesce che puo' essere un piranha (p) o un pesce bandiera (b).*

Viene aggiunto un pesce piranha e poi subito estratto uno dei due in modo casuale. Il risultato dell'estrazione é un piranha.

Qual é la probabilitá che il primo pesce presente nella vasca fosse effettivamente un piranha?

Intuitivamente (considerando l'albero delle possibilitá di Figura 12:

$$\frac{path1}{path2 + path1} = \frac{\frac{1}{2}}{\frac{1}{4} + \frac{1}{2}} = \frac{2}{3}$$

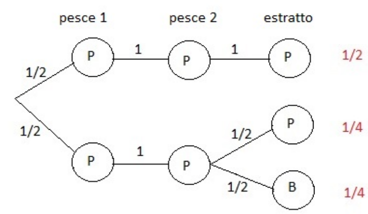

Figura 12: Chance tree per il problema del Sushi Delight

Applicando formalmente Bayes:

$$P(P1 = p \mid E = p, P2 = p) =$$

$$P(E = p, P2 = p \mid P1 = p) \cdot P(P1 = p) =$$

$$P(E = p \mid P2 = p, P1 = p) \cdot P(P2 = p \mid P1 = p) \cdot P(P1 = p) =$$

$$\frac{P(E = p \mid P2 = p, P1 = p) \cdot P(P2 = p) \cdot P(P1 = p)}{\sum_{i=[''p'',''b'']} P(E = p \mid P2 = p, P1 = i) \cdot P(P2 = P) \cdot P(P1 = i)} =$$

$$\frac{P(E = p \mid P2 = p, P1 = p) \cdot P(P1 = p)}{\sum_{i=[''p'',''b'']} P(E = p \mid P2 = p, P1 = i) \cdot P(P1 = i)}$$

Per i = "p" : 1
Per i = "b" : $\frac{1}{2}$