

Apprendimento in Probabilità ed Approssimazione

Apprendimento PAC

PAC Learning (Probably Approximately Correct Learning)

Assume che: input ed output sono binari, non c'è rumore, le istanze sono estratte da X concordemente ad una distribuzione di probabilità \mathcal{D} *arbitraria ma stazionaria*.

Il framework di apprendimento PAC cerca di rispondere alle seguenti domande:

- Sotto quali condizioni apprendere con successo è possibile o impossibile ?
- Sotto quali condizioni si può assicurare che un particolare algoritmo di apprendimento apprenda con successo ?

Apprendimento PAC

Consideriamo una classe C di possibili concetti target (funzioni che vogliamo apprendere) definita su uno Spazio delle Istanze X (con istanze di dimensione m), ed un algoritmo di apprendimento L che utilizza uno Spazio delle Ipotesi \mathcal{H} .

Def.: C è PAC-apprendibile da L usando \mathcal{H} se per ogni

- $c \in C$,
- distribuzione \mathcal{D} su X ,
- ϵ tale che $0 < \epsilon < 1/2$,
- δ tale che $0 < \delta < 1/2$,

l'algoritmo di apprendimento L con probabilità almeno $(1 - \delta)$ restituisce una ipotesi $h \in \mathcal{H}$ tale che $error_{\mathcal{D}}(h) \leq \epsilon$, in tempo che è **polinomiale** in $1/\epsilon$, $1/\delta$, m , e $size(c)$ (spazio di memoria necessario per rappresentare c).

Apprendimento PAC

Si può dimostrare che assumendo:

- $c \in \mathcal{H}$
- L consistente, cioè che restituisce una ipotesi h consistente con Tr , o equivalentemente $h \in VS_{\mathcal{H}, Tr}$ (ad esempio, **Find-S** è consistente)

allora, con probabilità almeno $(1 - \delta)$, L restituisce una ipotesi $h \in \mathcal{H}$ tale che $error_{\mathcal{D}}(h) < \epsilon$ se il numero di esempi di apprendimento n soddisfa la seguente disuguaglianza:

$$n \geq \frac{1}{\epsilon} (\ln(|\mathcal{H}|) + \ln(\frac{1}{\delta}))$$

dove $|\mathcal{H}|$ è la cardinalità dello Spazio delle Ipotesi e $\ln(\cdot)$ è il logaritmo naturale

- infine, poiché $(1 - \epsilon) \leq e^{-\epsilon}$ se $0 \leq \epsilon \leq 1$, abbiamo $|\mathcal{H}|(1 - \epsilon)^n \leq |\mathcal{H}|e^{-\epsilon n}$

Prova...

Idea base: bisogna dare un limite al numero di esempi necessario ad assicurare che il Version Space (ricordiamo che L è consistente) non contiene ipotesi non “accettabili”:

$$(\forall h \in VS_{\mathcal{H}, Tr}) \text{error}_{\mathcal{D}}(h) < \epsilon$$

Apprendimento PAC

Si può dimostrare che assumendo:

- $c \in \mathcal{H}$

Prova...

$$(\forall h \in \mathcal{H}) \text{error}_{\mathcal{D}}(h) < \epsilon$$

Primo risultato: se $|\mathcal{H}| < \infty$, $|Tr| = n > 0$ e Tr è costituito da esempi di un concetto target c estratti indipendentemente ed a caso, allora per ogni $0 \leq \epsilon \leq 1$, la probabilità che la disuguaglianza **NON** sia soddisfatta è minore od uguale a $|\mathcal{H}|e^{-\epsilon n}$:

- siano h_1, \dots, h_k tutte le ipotesi con $\text{error}_{\mathcal{D}}(h) \geq \epsilon$; la disuguaglianza NON è soddisfatta se e solo se **ALMENO** una di tali ipotesi è consistente con Tr
- la probabilità che una di tali ipotesi sia consistente con un singolo esempio è $(1 - \epsilon)$, e quindi per n esempi indipendenti la probabilità è $(1 - \epsilon)^n$
- poiché esistono k di tali ipotesi, la probabilità che ci interessa è a più
 $k(1 - \epsilon)^n \leq |\mathcal{H}|(1 - \epsilon)^n$ (poiché $k \leq |\mathcal{H}|$)
- infine, poiché $(1 - \epsilon) \leq e^{-\epsilon}$ se $0 \leq \epsilon \leq 1$, abbiamo $|\mathcal{H}|(1 - \epsilon)^n \leq |\mathcal{H}|e^{-\epsilon n}$

Prova...

Cosa abbiamo ottenuto: abbiamo un limite superiore alla probabilità che usando un Tr con n esempi, il Version Space contenga qualche ipotesi CATTIVA, cioè che L restituisca h , una delle ipotesi cattive, per cui $error_{\mathcal{D}}(h) \geq \epsilon$!

Quindi, se vogliamo che tale probabilità sia inferiore a livello desiderato δ

$$|\mathcal{H}|e^{-\epsilon n} \leq \delta$$

bisogna usare un valore per n per cui

$$n \geq \frac{1}{\epsilon} (\ln(|\mathcal{H}|) + \ln(\frac{1}{\delta}))$$

o, alternativamente, se si usa un tale valore per n , con probabilità $1 - \delta$ l'ipotesi h restituita da L avrà $error_{\mathcal{D}}(h) < \epsilon$

Apprendimento PAC: esempio \mathcal{H}_4

Congiunzione di m letterali

- Spazio delle Istanze \rightarrow stringhe di m bit: $X = \{s | s \in \{0, 1\}^m\}$
- Spazio delle Ipotesi \rightarrow tutte le sentenze logiche che riguardano i letterali l_1, \dots, l_m (anche in forma negata, $\neg l_i$) e che contengono solo l'operatore \wedge (**and**):

$$\mathcal{H} = \{f_{\{i_1, \dots, i_j\}}(s) | f_{\{i_1, \dots, i_j\}}(s) \equiv L_{i_1} \wedge L_{i_2} \wedge \dots \wedge L_{i_j}, \\ \text{dove } L_{i_k} = l_{i_k} \text{ oppure } \neg l_{i_k}, \{i_1, \dots, i_j\} \subseteq \{1, \dots, 2m\}\}$$

Notare che se in una formula un letterale compare sia affermato che negato, allora la formula ha sempre valore di verità *false* (formula non soddisfacibile)

Quindi, tutte le formule che contengono almeno un letterale sia affermato che negato sono equivalenti alla funzione che vale sempre *false*

Apprendimento PAC: esempio \mathcal{H}_4

E' la congiunzione di m letterali PAC-apprendibile usando **Find-S** con \mathcal{H}_4 ? **Si !**

Infatti:

- ogni congiunzione di m letterali è inclusa in \mathcal{H}_4
- **Find-S** è consistente
- $|\mathcal{H}_4| = 3^m + 1$ e poiché $\frac{1}{\epsilon}(\ln(3^{m+1}) + \ln(\frac{1}{\delta})) > \frac{1}{\epsilon}(\ln(3^m + 1) + \ln(\frac{1}{\delta}))$

$$n \geq \frac{1}{\epsilon}((m + 1)\ln(3) + \ln(\frac{1}{\delta}))$$

quindi n è polinomiale in $1/\epsilon$, $1/\delta$, m , e $size(c)$ (che non compare)

- per ogni esempio di apprendimento **Find-S** impiega tempo lineare nella dimensione della ipotesi corrente (e tale dimensione è $\geq size(c)$) e nella dimensione dell'input (m), quindi di nuovo polinomiale, e globalmente è polinomiale per Tr

Quindi tutte le condizioni per la PAC-apprendibilità sono soddisfatte !

Apprendimento PAC

Usando la disuguaglianza precedente ed altre considerazioni è possibile mostrare che alcune classi di concetti non sono PAC-apprendibili dato uno specifico algoritmo di apprendimento L e \mathcal{H} .

In particolare è possibile mostrare che:

- se \mathcal{H} contiene tutte le funzioni booleane definite su X allora $C = \mathcal{H}$ non è PAC-apprendibile da algoritmi consistenti
- esistono classi di concetti C che non sono PAC-apprendibili da algoritmi consistenti che usano C come Spazio delle Ipotesi, tuttavia diventano PAC-apprendibili se uno Spazio delle Ipotesi “più grande” è usato, cioè $C \subset \mathcal{H}$

Il problema con la disuguaglianza data è che questa non può essere usata se $|\mathcal{H}| = \infty$

Tuttavia, il fattore chiave non è quante funzioni diverse sono contenute in \mathcal{H} , ma quante funzioni “utili” sono in \mathcal{H} : **VC-dimension** dà una risposta a questa domanda

Apprendimento PAC

Si può mostrare che, se assumiamo:

- $c \in \mathcal{H}$
- L consistente

allora con probabilità almeno $(1 - \delta)$, l'algoritmo di apprendimento L restituisce una ipotesi $h \in \mathcal{H}$ tale che $error_{\mathcal{D}}(h) \leq \epsilon$ se il numero di esempi di apprendimento n soddisfa la seguente disuguaglianza:

$$n \geq \frac{1}{\epsilon} \left(4 \log_2 \left(\frac{1}{\delta} \right) + 8VC(\mathcal{H}) \log_2 \left(\frac{13}{\epsilon} \right) \right)$$

Notare che $VC(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$

$$VC(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$$

Mostriamo che $VC(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$

- per ogni S tale che \mathcal{H} frammenta S abbiamo $|\mathcal{H}| \geq 2^{|S|}$, infatti \mathcal{H} può implementare tutte le possibili dicotomie di S , che sono esattamente $2^{|S|}$.
- scegliendo un S tale che $|S| = VC(\mathcal{H})$, otteniamo $|\mathcal{H}| \geq 2^{VC(\mathcal{H})}$

Quindi, applicando \log_2 ad entrambi i lati della disuguaglianza, possiamo concludere che $\log_2(|\mathcal{H}|) \geq VC(\mathcal{H})$