Nozioni minime di machine learning per il corso di bioinformatica

Giorgio Valentini
DSI – Università degli Studi di Milano

1

Metodi di machine learning

- I metodi di apprendimento automatico (Machine Learning - ML) consentono di effettuare inferenze e predizioni da un insieme limitato di dati disponibili
- Gli algoritmi di ML “apprendono” una funzione predittiva utilizzando una serie di esempi (dati) estratti secondo una determinata distribuzione di probabilità dall’ “Universo” dei dati.

2

Apprendimento induttivo

- Insieme finito di oggetti: $S = \{x_1, x_2, ..., x_n\}, S \subset U$
  - **Obiettivo**: apprendere una proprietà *P* di *U* dall'analisi di *S* tramite un algoritmo di apprendimento *A.*
    - **Es. 1**: Dato un insieme *S* sottoinsieme di *U* (universo delle proteine) predire per un *x* di *U* se *x* ha struttura secondaria α o β, utilizzando un algoritmo *A(S).*
    - **Es. 2**: Dato un insieme *S* sottoinsieme di *U* (universo dei pazienti) predire per un *x* di *U* se è sano o malato, utilizzando un algoritmo *A(S).*

3

Principali tipologie di problemi di ML

1. Problemi supervisionati:
   Dati XCU (insieme di oggetti) e *I(X)* (etichette), predire $P(x), x \in U$
   L'algoritmo *A* (eventualmente con parametri ω) utilizza *X* per costruire un predittore $\hat{P}$ che approssimi *P* (non noto a priori $\forall x \in U$).
2. Problemi non supervisionati:
   Dati XCU (insieme di oggetti) senza *I(X)* (etichette), predire *P(x)* (problema mal posto).

- Esistono anche altre tipologie di problemi.

4

Principali problemi supervisionati

1. Problemi di classificazione: P(x) è una funzione a valori discreti
2. Problemi di regressione: P(x) è una funzione a valori continui

Gli esempi delle slide precedenti erano relativi a problemi di classificazione.

5

Gli oggetti sono caratterizzati da feature

$\phi : U \rightarrow F$

$Es : \phi(x) = z$

$z \in R^d opp. \quad z \in N^d$

z può essere un vettore, una stringa, un albero, un grafo, …

Un algoritmo *A* per apprendere *P* utilizza $\phi(S), S \subset U$

6

Algoritmi di *di classificazione supervisionati*

Un algoritmo *A* con parametri ω utilizza un data set *S* con etichette per generare un predittore (*classificatore*):

$$A(\phi(S), l(S), \omega) \rightarrow \hat{P}$$

Obiettivo della classificazione supervisionata*:*

$$P^* = \arg\min_{\hat{P}} Prob\left(P(\phi(x)) \neq \hat{P}(\phi(x))\right)$$

(per un *x* estratto a caso da *U*)

7

Stima dell'errore

- In pratica non si dispone di *U* e delle rispettive etichette.
- Ad es:. Si dispone solo di *XCU* (insieme di oggetti) e *l(X)* (etichette):
  1. Si partiziona *X* in *S* e *T*
  2. Training: $A(\phi(S), l(S), \omega) \rightarrow \hat{P}$
  3. Testing: $\varepsilon = \frac{1}{|T|} \sum_{x \in T} I\left(P(\phi(x)) \neq \hat{P}(\phi(x))\right)$

I(z)=1 se z è vero, 0 altrimenti (funzione di perdita 0/1)

8

Tecniche di stima dell'errore

- La tecnica di stima dell' errore vista precedentemente è detta di *hold-out*
- Altre tecniche:
  - Hold-out multiplo
  - Cross-validation
  - Leave one out
  - Out-of-bag
  - …
- *Scelta del predittore*:
  Se $\varepsilon(\hat{P}_\omega)$ è l'errore sul test set ottenuto dal predittore $\hat{P}$ generato da *A* con parametri ω, allora:

$$\hat{P}_{\omega^*} = \arg\min_{\omega} \varepsilon(\hat{P}_\omega)$$

9

Algoritmi di *di classificazione non supervisionati*

Un algoritmo *A* con parametri ω utilizza un data set *S* senza etichette per generare un *clustering*:

$$A(\phi(S), \omega) \rightarrow C_\omega$$

$$C_\omega = \{C_1, C_2, ..., C_n\}, \;\; C_i \subseteq S, \;\; 1 \le i \le n, \;\; \cup C_i = S$$

Se $\forall i, j \;\; i \neq j, \;\; C_i \cap C_j = \varnothing$

allora C_ω è una partizione

Problema: come valutare la qualità di C_ω?

10