



UNIVERSITÀ DI PISA
Dipartimento di Informatica

MACHINE LEARNING

Appunti dalle lezioni del Prof. Alessio Micheli

Pasquale Miglionico

Guido Narduzzi

Enrico Negri

Filippo Quattrocchi

Francesco Zigliotto

ANNO ACCADEMICO 2019–2020

INDICE

1	INTRODUZIONE	5
1.1	Lezione di giovedì 26 settembre	5
1.1.1	Contestualizzazione	5
1.1.2	Terminologia	5
2	ESEMPI	7
2.1	Lezione di domenica 22 settembre	7
2.1.1	Alcune indicazioni	7

1 | INTRODUZIONE

1.1 LEZIONE DI GIOVEDÌ 26 SETTEMBRE

1.1.1 Contestualizzazione

L'obiettivo del Machine Learning è insegnare ad un sistema un compito preciso, costruendo un modello utilizzabile per predire il corretto output, dopo aver esaminato un gran numero di esempi. Tale metodo di apprendimento è detto *per generalizzazione*.

È utile per esempio quando l'approccio teorico a un determinato problema è difficilmente praticabile, o quando i dati in input sono poco accurati, affetti da errore o incompleti.

Definizione 1.1. (Machine Learning). Il *Machine Learning* studia e propone metodi per inferire funzioni o correlazioni che, a partire da dati osservati, producano il corretto output sui *samples* forniti e li generalizzino con ragionevole accuratezza.

1.1.2 Terminologia

Un *machine learning system* si compone di *dati*, *tasks*, *modelli*, *algoritmi di apprendimento* e *validazione*.

Definizione 1.2. (Dati). I *dati* rappresentano le *esperienze disponibili*. Possono essere organizzati in un certo numero l di istanze x_p (*samples*, *instances*), ciascuno contenente n attributi (*features*). Con $x_{p,j}$ indicheremo l'attributo j -esimo della p -esima istanza.

Osservazione 1.3. Se un attributo può assumere un numero finito di valori, risulta spesso conveniente rappresentarlo come un vettore di dimensione k (dove k è il numero dei valori possibili) con componenti tutte nulle a parte una.

Esempio 1.4. Se i valori possibili sono i tre colori *rosso* (R), *verde* (G), *blu* (B), si pone

$$R = (1, 0, 0), \quad G = (0, 1, 0), \quad B = (0, 0, 1). \quad (1.1)$$

Definizione 1.5. (Rumore, outliers). Il *rumore* è l'aggiunta di fattori esterni dovuta al processo di misura (e non alla legge soggiacente). Gli *outliers* sono dati che si collocano molto lontani rispetto agli altri.

I *tasks* sono in genere di due tipologie (ma ce ne sono altre):

Definizione 1.6. (Supervised learning). Nel *supervised learning* sono dati dei *samples* di una funzione f ignota, nella forma

$$\langle \text{input}, \text{output} \rangle$$

si tratta di trovare una buona approssimazione di f . Gli input sono detti anche *variabili indipendenti*, gli output *variabili dipendenti* o *risposte*. Se f è a valori discreti, il

problema si dice di *classificazione*. Se l'output è costituito da valori reali, allora si parla di *regressione*.

Definizione 1.7. (Unsupervised learning). Nell'*unsupervised learning* non si dispone di input e output nelle istanze, ma solo di dati non etichettati. Un problema tipico è quello di raggruppare tali dati secondo determinati criteri.

Noi ci occuperemo soprattutto di supervised learning.

Definizione 1.8. (Modello, ipotesi). Il modello cerca di descrivere la relazione tra i dati con un *linguaggio*, legato alla rappresentazione dei dati. Le *ipotesi* sono le funzioni h_w proposte dal modello per approssimare la “vera” funzione f . Le ipotesi sono indicizzate da parametri (w) e formano uno *spazio delle ipotesi* H .

In generale non esiste un modello *ottimo*: se un modello di apprendimento è il migliore in qualche problema, sarà peggiore di altri in altri problemi. Questo concetto è noto come *No Free Lunch Theorem*, ovvero “non c'è un pranzo gratis” (mah, sarà qualche detto inglese, ndr). In ogni caso, questo non significa che tutti i modelli siano equivalenti.

Definizione 1.9. (Algoritmo di apprendimento). Un *algoritmo* di apprendimento si occupa di cercare nello spazio delle ipotesi H (di un modello fissato) la migliore approssimazione della funzione f .

Come definiamo *buona approssimazione*? Si utilizza una *loss function* $L(h(x), d)$, che misura la distanza tra $h(x)$ e d , dove $d = f(x)$ è il valore osservato.

Definizione 1.10. (Errore). L'errore è definito da

$$E = \frac{1}{l} \sum_{i=1}^l L(h(x_i), d_i) \quad (1.2)$$

dove x_i sono le istanze e $d_i = f(x_i)$ i valori osservati.

Esempio 1.11. Nei problemi di regressione spesso come loss function si usa

$$L(h(x), d) = (d - h(x))^2 \quad (1.3)$$

e in tal caso l'errore si dice *errore quadratico medio* (MSE).

Se siamo di fronte a un problema di classificazione, allora è più conveniente usare la loss function che vale 1 se i suoi due argomenti sono uguali (e dunque la classificazione è corretta) e 0 altrimenti.

Osservazione 1.12. In Machine Learning, quando si parla di *performance*, si fa riferimento all'accuratezza predittiva, non all'efficienza computazionale.

2 | ESEMPI

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

2.1 LEZIONE DI DOMENICA 22 SETTEMBRE

2.1.1 Alcune indicazioni

Un paio di proposte, per uniformità:

- Ogni lezione inizia con `\lecture{\langle data \rangle}`, dove la data è del tipo 29/2.
- Per creare paragrafi numerati, non va usato `\section`, ma `\subsection`;
- Non usate il grassetto, ma il *corsivo*, soprattutto quando s'introduce un *nuovo termine*, e poi si può usare il nuovo termine anche senza corsivo;
- Usate `equation` per le equazioni, in modo che vengano tutte numerate; le equazioni a fine frase terminano con il punto, ma magari evitiamo altro tipo di punteggiatura alla fine di equazioni.
- usare molto gli ambienti `definition`, `theorem`, `example`, `remark`...
- negli elenchi, il “,” alla fine degli *item* e il punto alla fine dell'ultimo. Ma se sono frasi lunghe va bene anche il punto dappertutto.

Definizione 2.1. (Tensore). Un *tensore* è ciò che ruota come un tensore.

$$e^z = \sum_{n=0}^{+\infty} \frac{z^n}{n!}. \quad (2.1)$$

```
#include <stdio.h>
int main()
{
    // printf() displays the string inside quotation
    printf("Hello, World!");
    return 0;
}
```
