

# Note del corso

## MACHINE LEARNING PER LA FISICA APPLICATA E FISICA DELLE ALTE ENERGIE

Raviola Alessio

18 ottobre 2022

### 1 Introduzione

A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ .

L'impostazione del corso è di tipo *probabilistico* (statistical learning). Le quantità non note sono trattate come **variabili aleatorie** (RANDOM VARIABLES) a cui viene associata una **distribuzione di probabilità** (PROBABILITY DISTRIBUTION) che descrive il set (pesato) di valori che la variabile può assumere.

Abbiamo tre tipi di machine learning:

- SUPERVISED LEARNING;
- UNSUPERVISED LEARNING;
- REINFORCEMENT LEARNING;

il corso si focalizza sui primi due tipi.

#### 1.1 Supervised learning

Il **compito**  $T$  consiste nell'imparare una mappa  $f$  dagli input  $x \in X$  agli output  $y \in Y$ . Gli **input**  $x$  sono chiamati FEATURES (o COVARIATES o PREDICTORS) e sono in genere costituiti da un vettore reale con dimensione fissata, ovvero abbiamo  $X \equiv \mathbb{R}^D$ . Gli **output** sono chiamati LABEL (o TARGET o RESPONSE).

L'**esperienza**  $E$  consiste in un TRAINING SET  $\mathcal{D}$  di  $N$  coppie input-output:

$$\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N, \quad (1)$$

dove  $N$  è detta SAMPLE SIZE.

La **performance** dipende dal compito  $T$ .

##### 1.1.1 Classificazione

Problemi comuni in machine learning sono quelli di **classificazione**. In un problema di questo tipo lo spazio degli output  $C$  è un set *non ordinato* di label  $y = \{1, 2, \dots, C\}$  dette CLASSES. Quello che chiede il problema è di predire una classe dato un input, problemi di questo tipo sono detti di PATTERN RECOGNITION<sup>1</sup>.

**Esempio 1.1** (Classificazione specie di iris). In generale in IMAGE CLASSIFICATION gli input  $X$  sono immagini, quindi:

$$X = \mathbb{R}^D, \quad D = C \times D_1 \times D_2, \quad (2)$$

---

<sup>1</sup>Se abbiamo solo due classi, i.e. solo due output, allora il problema si dice di CLASSIFICAZIONE BINARIA

Index	sl [cm]	sw [cm]	pl [cm]	pw [cm]	Label
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
50	7.0	3.2	4.7	1.4	Versicolor
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
150	5.9	3.0	5.1	1.8	Virginica

Tabella 1: Design matrix del training set per classificazione specie di iris.

ove  $C = 3$  sono i canali RGB. E cerchiamo una mappa

$$f : X \longrightarrow Y \quad (3)$$

che ci dica a quale delle classi appartenenti a  $Y$  l'immagine appartiene. Per le specie di iris però i botanisti hanno individuato 4 caratteristiche numeriche: lunghezza e larghezza del sepal e del petalo; dunque abbiamo  $X = \mathbb{R}^4$ . Supponiamo che il training set sia una collezione di 150 esempi delle 3 specie, 50 per ognuna. I dati possono essere raccolti in una matrice detta DESIGN MATRIX come TABULAR DATA - come in Tabella 1.

Se abbiamo  $N$  elementi nel training set, ognuno con dimensione  $D = \dim X + \dim Y$ , allora abbiamo:

- BIG DATA se  $N \gg D$ , ovvero se il numero di elementi è molto superiore alla loro dimensione;
- WIDE DATA se  $D \gg N$ , ovvero se la dimensione degli elementi è molto superiore al loro numero.

Una buona idea è fare una *esplorazione dei dati* (EXPLOATORY DATA ANALYSIS) per vedere se ci sono dei pattern ovvi, ad esempio tramite grafici. Per grandi basi dati (big data) possiamo procedere mediante DIMENSIONALITY REDUCTION:

$$f(\mathbf{x}, \boldsymbol{\theta}) = \left\{ \begin{array}{ll} p_l < 2.45 \text{ cm} & \text{Setosa} \\ \text{Altrimenti} & \left\{ \begin{array}{ll} p_w < 1.75 \text{ cm} & \text{Versicolor} \\ \text{Altrimenti} & \text{Virginica} \end{array} \right\} \end{array} \right\} \quad \text{DECISION TREE}, \quad (4)$$

ove  $\boldsymbol{\theta}$  è detto THRESHOLD PARAMETER. Questo decision tree è visualizzato in Figura 1. La performance può essere quindi misurata con il MISCLASSIFICATION RATE:

$$\mathcal{L}(\boldsymbol{\theta}) \equiv \frac{1}{N} \sum_{n=1}^N \mathbb{I}(y_n \neq f(\mathbf{x}_n, \boldsymbol{\theta})), \quad (5)$$

dove  $\mathbb{I}(e)$  è l'**indicatore binario**

$$\mathbb{I}(e) = \begin{cases} 1 & \text{se } e \text{ è vero} \\ 0 & \text{se } e \text{ è falso} \end{cases} \quad (6)$$

Nel caso in cui alcuni errori di classificazione siano più dannosi di altri posso definire una loss function  $l(y, \hat{y})$  e ridefinire il misclassification rate come l'EMPIRICAL RISK:

$$\mathcal{L}(\boldsymbol{\theta}) \equiv \frac{1}{N} \sum_{n=1}^N l(y_n, f(\mathbf{x}_n, \boldsymbol{\theta})). \quad (7)$$

Un modo che abbiamo per definire il TRAINING (o MODEL FITTING) è modificare questo rischio empirico, ovvero trovare  $\hat{\boldsymbol{\theta}}$  tale che

$$\mathcal{L}(\hat{\boldsymbol{\theta}}) = \min[\mathcal{L}(\boldsymbol{\theta})]. \quad (8)$$

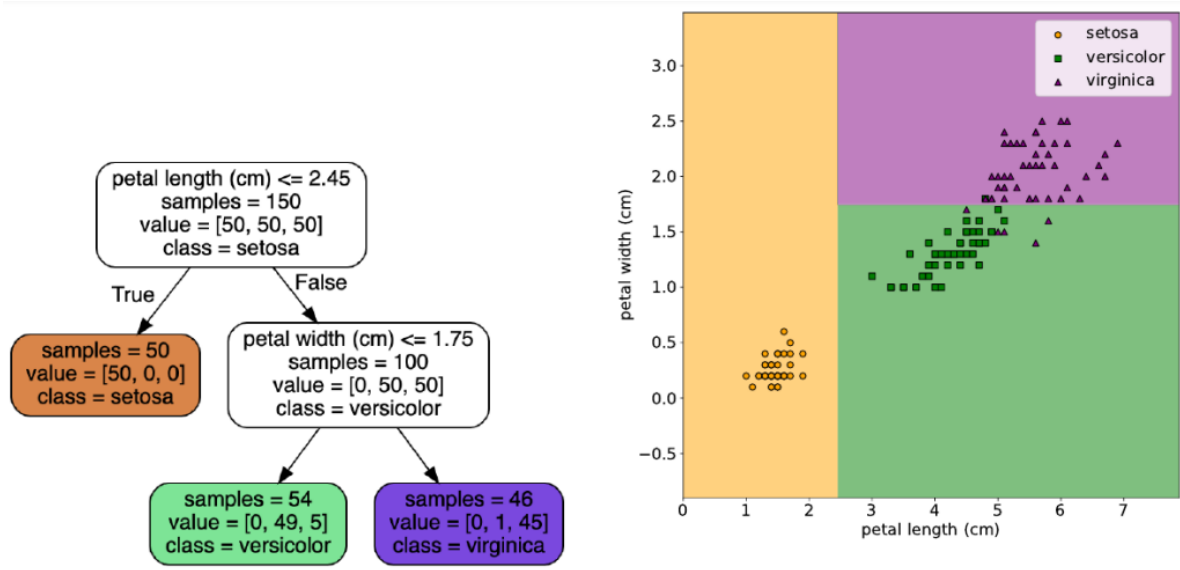


Figura 1: Decision tree per problema di classificazione specie di iris.

## 2 Richiami di probabilità

Abbiamo diverse definizioni di probabilità.

**Definizione Frequentistica** La probabilità di un evento è il rapporto tra il numero di casi favorevoli e il numero di casi possibili.

**Definizione Soggettiva** La probabilità di un evento è il prezzo che un individuo ritiene equo pagare per ricevere 1 se l'evento si verifica e 0 altrimenti.

**Definizione Bayesiana** La probabilità di un evento è l'*incertezza* con cui l'evento si verifica.

**Definizione Assiomatica** Kolmogorov nel 1933 costruisce la teoria della probabilità a partire da degli assiomi.

L'incertezza può essere di due tipi:

**aleatoria** ovvero è una DATA UNCERTANTY;

**epistemica** ovvero è una MODEL UNCERTANTY;

### 2.1 Proprietà della probabilità

Chiamiamo  $\Pr(A)$  la probabilità dell'evento  $A$ , allora abbiamo le seguenti proprietà.

**Proprietà 2.1.1** (Joint probability). Se  $A$  e  $B$  sono due eventi indipendenti, allora:

$$\Pr(A \wedge B) \equiv \Pr(A, B) = \Pr(A) \cdot \Pr(B). \quad (9)$$

**Proprietà 2.1.2** (Union probability). Anche detta regola di unione esclusione. Se  $A$  e  $B$  sono due eventi indipendenti, allora:

$$\Pr(A \vee B) = \Pr(A) + \Pr(B) - \Pr(A \wedge B). \quad (10)$$

**Proprietà 2.1.3** (Conditional probability).

$$\Pr(A | B) = \frac{\Pr(A, B)}{\Pr(B)}. \quad (11)$$

Se i due eventi sono indipendenti questa si riduce a  $\Pr(A | B) = \Pr(A)$ .

**Proprietà 2.1.4** (Conditional independence).

$$\Pr(A, B \mid C) = \Pr(A \mid C) \cdot \Pr(B \mid C) \quad (12)$$

## 2.2 Random variables

Rappresentiamo con  $X$  una variabile di cui non conosciamo il valore e la chiamiamo *variabile casuale* (RANDOM VARIABLE). Il set dei valori che  $X$  può assumere è detto *spazio di sampling* (SAMPLING SPACE). Un evento è dunque un set di risultati dato un sampling space definito.

Se la variabile è **discreta** abbiamo un sampling space numerabile e la PMF (PROBABILITY MASS FUNCTION):

$$p(x) \equiv \Pr(X = x). \quad (13)$$

Se invece la variabile è **continua** abbiamo un sampling space non numerabile e la CDF (CUMULATIVE DISTRIBUTION FUNCTION):

$$P(x) \equiv \Pr(X \leq x), \quad (14)$$

da cui possiamo definire la PDF (PROBABILITY DENSITY FUNCTION):

$$p(x) \equiv \frac{d}{dx} P(x), \quad (15)$$

da cui segue:

$$\Pr(a \leq X \leq b) = \int_a^b p(x) dx = P(b) - P(a) \quad (16)$$

$$\implies \Pr(x \leq X \leq x + dx) \approx p(x) dx. \quad (17)$$

Se la CDF  $P(x)$  è monotona crescente allora la sua inversa  $P^{-1}(q)$  è detta *quantile*. Il valore  $x_q = P^{-1}(q)$  è il valore per cui  $\Pr(X \leq x_q) = q$ , ovvero il quantile  $q$  della distribuzione  $P$ .

Se abbiamo due variabili casuali  $X$  e  $Y$  allora possiamo definire la JOINT DISTRIBUTION:

$$p(x, y) = p(X = x, Y = y) \equiv p(X = x \wedge Y = y). \quad (18)$$

Se le due variabili sono indipendenti e con cardinalità finita possiamo definire la *distribuzione marginale* (MARGINAL DISTRIBUTION) come:

$$p(X = x) = \sum_y p(X = x, Y = y), \quad (19)$$

altrimenti se sono dipendenti la *distribuzione condizionale* (CONDITIONAL DISTRIBUTION) come:

$$p(Y = y \mid X = x) = \frac{p(X = x, Y = y)}{p(X = x)} \quad (20)$$

$$\implies p(x, y) = p(y \mid x) \cdot p(x),$$

da cui segue la **chain rule**:

$$p(\mathbf{x}_1, D) = p(x_1) p(x_2 \mid x_1) p(x_3 \mid x_1, x_2) \cdots p(x_D \mid p_{1:D-1}). \quad (21)$$

Due variabili si dicono MARGINALMENTE INDIPENDENTI se

$$X \perp Y \iff p(X, Y) = p(X) p(Y), \quad (22)$$

mentre si dicono **condizionalmente indipendenti** se

$$X \perp Y \iff p(X, Y \mid Z) = p(X \mid Z) p(Y \mid Z). \quad (23)$$

## 2.3 Momenti di una distribuzione