Note del corso
Machine Learning per la fisica applicata e fisica delle alte energie

Raviola Alessio

19 ottobre 2022

1 Introduzione

> A computer program is said to learn from experience E with respecto 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:

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

dove N è detta SAMPLE SIZE.

La **performance** dipenda 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 $y = \{1, 2, \cdots, C\}$ dette CLASSES. Quello che chiede il problema è di predire una classe dato un input, problemi di questo tipo sono detti di PATTERN RECOGNITION[1].

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

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

[1] 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
⋮	⋮	⋮	⋮	⋮	⋮
50	7.0	3.2	4.7	1.4	Versicolor
⋮	⋮	⋮	⋮	⋮	⋮
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

$$f : X \longrightarrow Y \tag{3}$$

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 sepalo e del petalo; dunque abbiamo $X = \mathbb{R}^4$. Supponiamo che il traning 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:

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

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

$$\mathcal{L}(\theta) \equiv \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\left(y_n \neq f(\boldsymbol{x_n}, \boldsymbol{\theta})\right), \tag{5}$$

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

$$\mathbb{I}(e) = \left\{ \begin{array}{l} 1 \text{ se } e \text{ è vero} \\ 0 \text{ se } e \text{ è falso} \end{array} \right. . \tag{6}$$

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

$$\mathcal{L}(\theta) \equiv \frac{1}{N} \sum_{n=1}^{N} l\left(y_n, f(\boldsymbol{x_n}, \boldsymbol{\theta})\right). \tag{7}$$

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

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

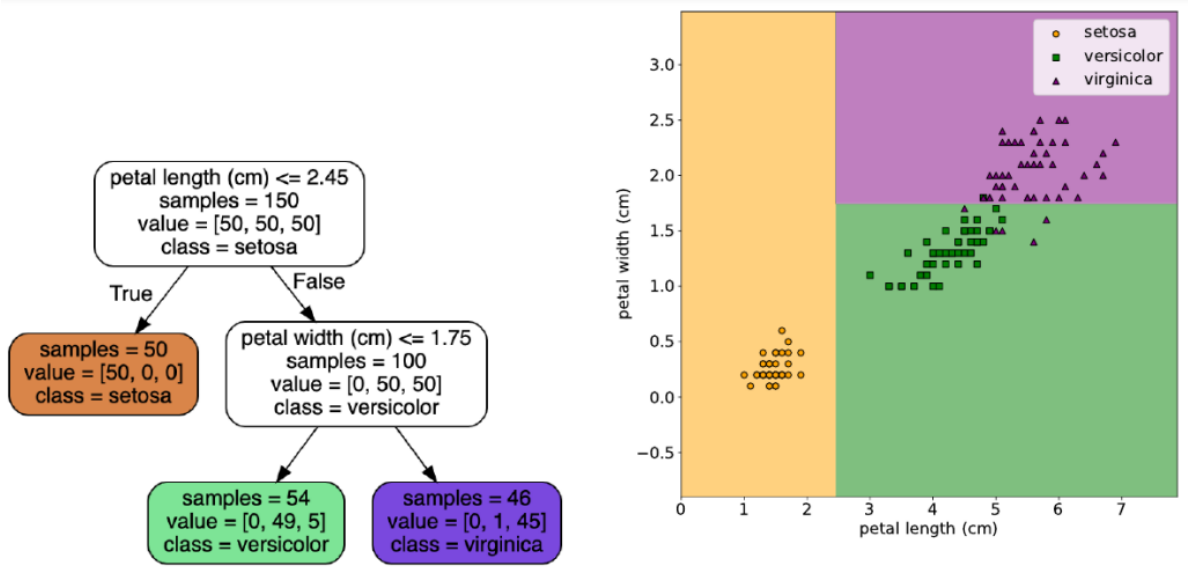

Figura 1: Decision tree per problema di classificazione specie di iris.

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:

$$\Pr(A \wedge B) \equiv \Pr(A, B) = \Pr(A) \cdot \Pr(B). \tag{9}$$

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

$$\Pr(A \vee B) = \Pr(A) + \Pr(B) - \Pr(A \wedge B). \tag{10}$$

Proprietà 2.1.3 (Conditional probability)**.**

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

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

Proprietà 2.1.4 (Conditional independece)**.**

$$\Pr(A, B|C) = \Pr(A|C) \cdot \Pr(B|C) \tag{12}$$

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):

$$p(x) \equiv \Pr(X = x). \tag{13}$$

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

$$P(x) \equiv \Pr(X \leq x), \tag{14}$$

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

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

da cui segue:

$$\Pr(a \leq X \leq b) = \int_a^b p(x) = P(b) - P(a) \tag{16}$$

$$\implies \Pr(x \leq X \leq x + dx) \approx p(x). \tag{17}$$

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:

$$p(x, y) = p(X = x, Y = y) \equiv p(X = x \wedge Y = y). \tag{18}$$

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

$$p(X = x) = \sum_y p(X = x, Y = y), \tag{19}$$

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

$$\begin{aligned} p(Y = y|X = x) &= \frac{p(X = x, Y = y)}{p(X = x)} \\ \implies \quad p(x, y) &= p(y|x) \cdot p(x), \end{aligned} \tag{20}$$

da cui segue la **chain rule**:

$$p(\boldsymbol{x}_1, D) = p(x_1)\, p(x_2|x_1)\, p(x_3|x_1, x_2) \cdots p(x_D|p_{1:D-1}). \tag{21}$$

Due variabili si dicono MARGINALMENTE INDIPENDENTI se

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

mentre si dicono **condizionalmente indipendenti** se

$$X \perp Y \quad \iff \quad p(X, Y|Z) = p(X|Z)\, p(Y|Z). \tag{23}$$

2.3 Momenti di una distribuzione