

Project / Homework Title

Student Name

17 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 RECOGNITION¹.

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

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

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

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

¹Se 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.

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

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

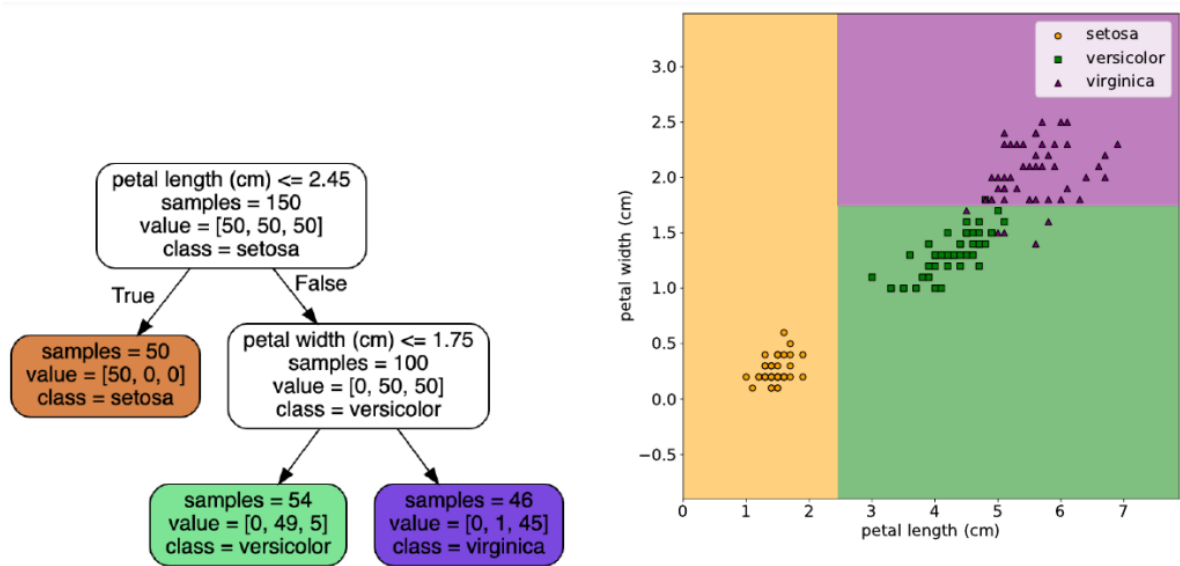


Figura 1: Decision tree per problema di classificazione specie di iris.