Appunti di machine learning

Daniele Besozzi

Anno accademico 2025/2026

Contents

Premesse

Questi sono appunti realizzati per riassumere e schematizzare tutti i concetti presentati durante il corso di machine learning tenuto presso il corso di laurea magistrale in informatica presso l'università degli studi di Milano Bicocca. Lo scopo di questo documento non è quello di sostituire le lezioni del corso o di essere l'unica fonte di studio, bensì integrare le altri fonti con un documento riassuntivo.
Mi scuso in anticipo per eventuali errori e prego i lettori di segnalarli contattandomi via mail all'indirizzo d.besozzi@campus.unimib.it.

Chapter 1

Introduzione

In questo capitolo presenterò gli aspetti matematici fondamentali per andare ad affrontare gli argomenti del corso.

1.1 Vettori e matrici

Denotiamo un vettore riga e colonna rispettivamente con (a, b, c) e $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ dove $a, b, c \in \mathbb{R}$ sono scalari.
In generale denotiamo con le lettere maiuscole le matrici, e.g. X e i suoi elementi con X_{ij}.
$x \in \mathbb{R}^n$ è un vettore di n elementi e $X \in \mathbb{R}^{m \times n}$ è una matrice di dimensione $m \times n$.

1.2 Norme di vettori e matrici

Dato un vettore $x \in \mathbb{R}^n$ vi sono diversi tipi di norme comunemente utilizzate.

- $||x||_2 = (\sum_{i=1}^n x_i^2)^{\frac{1}{2}}$ è il tipo più comune e viene chiamato **norma 2** di un vettore o **norma Euclidea**. Normalmente è denotato semplicemente con $||x||$.
- $||x||_1 = |x_1| + |x_2| + \cdots + |x_n|$, detta la **norma 1** o **distanza di Manhattan**
- $||x||_\infty = \max_{1 \leq i \leq n} |x_i|$, detta la **norma** ∞.

Analogamente, per una matrice $X \in \mathbb{R}^{m \times n}$, si possono definire diverse norme:

- **Norma di Frobenius**: $||X||_F = \left(\sum_{i=1}^m \sum_{j=1}^n |X_{ij}|^2 \right)^{\frac{1}{2}}$
- **Norma spettrale (o norma 2)**: $||X||_2$
- **Norma 1**: $||X||_1 = \sum_{i,j} |X_{ij}|$

1.3 Notazioni generiche

Siano:

- u: variabile indipendente (input), non necessariamente un vettore o uno scalare
- v: variabile dipendente (output), come sopra

allora abbiamo che:

- $x = \Phi(u)$, dove $x \in \mathbb{R}^d$ è il vettore di features e Φ è la funzione di mapping o embedding.
- $y = \Psi(v)$, dove $y \in \mathbb{R}^m$ è il vettore target (o di output) e Ψ è la funzione di mapping di feature in output.

Siano $x^1, \dots x^n$ e $y^1, \dots y^n$ due dataset di n esempi, dove x^i e y^i formano la i-esima coppia di dati. Dunque n è il numero di campioni, allora posso associarvi le due matrici dei dati

$$X = \begin{bmatrix} (x^1)^T \\ \vdots \\ (x^n)^T \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad Y = \begin{bmatrix} (y^1)^T \\ \vdots \\ (y^n)^T \end{bmatrix} \in \mathbb{R}^{n \times m}$$

Le cui righe sono i vettori feature e i vettori target rispettivamente, trasposti.
Definiamo allora:

- $g_\theta : \mathbb{R}^d \to \mathbb{R}^m$ è un predittore.
- $\hat{y} = g_\theta(x)$ è la predizione di y, dato x.
- $\Theta \in \mathbb{R}^p$ è il vettore di parametri del predittore.

La scelta dei parametri Θ a seconda dei dati viene chiamato *training* o *fitting* del predittore.