Statistical methods for machine learning

Mauro Tellaroli

1 Introduzione

La *data inference* è lo studio dei metodi che utilizzano i dati per predirre il futuro. Il *Machine Learning* è uno strumento potente che può essere usato per risolvere una grossa parte dei problemi di *data inference*, inclusi i seguenti:

- **Clustering**: raggruppare i *data points* in base alle loro similarità;
- **Prediction**: assegnare delle etichette (*label*) ai *data points*;
- **Generation**: generare nuovi *data points*;
- **Control**: eseguire una sequenza di azioni in un ambiente con l'obiettivo di massimizzare una nozione di utilità.

Con *data point* si intende una serie di informazioni legate ad un unico elemento; un'analogia può essere un *record* in un database.

Gli algoritmi che risolvono una *learning task* in base a dei dati già semanticamente etichettati lavorano in modalità ***supervised learning***. A etichettare i dati saranno delle persone o la natura. Un esempio dell'ultimo caso sono le previsioni del meteo. D'altra parte, gli algoritmi che utilizzano i dati senza la presenza di etichette lavorano in modalità ***unsupervised learning***.

In questo corso ci si focalizzerà sul *supervised learning* e la progettazione di sistemi di *machine learning* il cui obiettivo è apprendere dei **predittori**, ovvero funzioni che mappano i *data points* alla loro etichetta.

Label set $\mathcal{Y}$

Verrà usata $\mathcal{Y}$ per indicare il label set, ovvero l'insieme di tutte le possibili etichette di un *data point*. Le etichette potranno essere di due tipi differenti:

1. **Categoriche** ($\mathcal{Y} = \{\text{sport}, \text{politica}, \text{economia}\}$): si parlerà di problemi di **classificazione**;
2. **Numeriche** ($\mathcal{Y} \subseteq \mathbb{R}$): si parlerà di problemi di **regressione**.

È importante sottolineare come la reale differenza tra le due tipologie di etichetta sia il significato e non la sua rappresentazione in quanto, si potrà sempre codificare un'etichetta categorica in un numero.

A sottolineare ciò è il fatto che nella regressione l'errore è tipicamente una funzione della differenza $|y - \hat{y}|$, dove $\hat{y}$ è la predizione di y. Nella classificazione, invece, l'errore è tipicamente binario: predizione corretta ($\hat{y} = y$) o errata ($\hat{y} \neq y$).

Quando ci sono solo due possibili etichette ($|\mathcal{Y}| = 2$), si ha un **problema di classificazione binario** e, convenzionalmente, verrà usata una codifica numerica $\mathcal{Y} = \{-1, 1\}$.

1.1 Loss function ℓ

Come già visto precedentemente, si vuole misurare l'errore che un predittore commette su una determinata predizione. Per farlo si userà una **funzione di loss** ℓ non negativa che misurerà la

discrepanza $\ell(y, \hat{y})$ tra l'etichetta predetta $\hat{y}$ e quella corretta y. Si assumerà sempre $\ell(y, \hat{y}) = 0$ quando $\hat{y} = y$.

La funzione di loss più semplice per la classificazione è la **zero-one loss**:

$$\ell(y, \hat{y}) = \begin{cases} 0 & y = \hat{y} \\ 1 & \text{altrimenti} \end{cases}$$

Nella regressione, le tipiche funzioni di loss sono:

- la ***absolute loss***: $\ell(y, \hat{y}) = |y - \hat{y}|$
- la ***quadratic loss***: $\ell(y, \hat{y}) = (y - \hat{y})^2$

In alcuni casi può essere conveniente scegliere l'etichetta predetta da un insieme $\mathcal{Z}$ diverso da $\mathcal{Y}$. Per esempio, si consideri il problema di assegnare una probabilità $\hat{y} \in (0, 1)$ all'evento $y =$ "pioverà domani". In questo caso, $\mathcal{Y} = \{$"piove", "non piove"$\}$ e $\mathcal{Z} = (0, 1)$. Indicando questi due eventi con 1 (piove) e 0 (non piove), si può usare una funzione di loss per la regressione, come la *absolute loss*:

$$\ell(y, \hat{y}) = |y - \hat{y}| = \begin{cases} 1 - \hat{y} & y = 1 \qquad \text{(piove)} \\ \hat{y} & y = 0 \qquad \text{(non piove)} \end{cases}$$