

# Statistical methods for machine learning

Mauro Tellaroli

## 1 Introduzione

La *data inference* è lo studio dei metodi che utilizzano i dati per predire 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 ***predictors***, 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** ( $\mathcal{Y} = \{\text{sport, politica, economia}\}$ ): 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, banalmente, 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 ( $\hat{y} \neq y$ ).