

14. Introduzione (breve) a Scikit Learn

Corso di Python per il Calcolo Scientifico

Outline

- Stimatori e transformer
- Preprocessing dei dati

Stimatori e transformer

- Una volta installata la libreria (`pip install scikit-learn`) dovremo familiarizzare con il concetto di **stimatore (estimator)** e **transformer**.
- Gli stimatori sono oggetti che permettono di caratterizzare i diversi algoritmi di machine learning.
 - Offrono in generale i metodi `fit()` e `predict()`.
- I transformer sono oggetti che permettono di trasformare i dati.
 - Offrono in generale i metodi `fit()` e `transform()`.
- Gli oggetti di questo tipo ci permettono di definire la nostra **pipeline di processing**.

Preprocessing dei dati (1)

- Per effettuare il preprocessing dei dati dobbiamo usare alcune tecniche.
- Per assegnare i valori mancanti, dovremo usare degli **imputer**, come ad esempio gli oggetti di classe **SimpleImputer** o **IterativeImputer**.
- Per trasformare le feature numeriche, potremo usare degli **scaler**, come ad esempio gli oggetti di classe **MinMaxScaler** e **StandardScaler**.
- Per trasformare le feature categoriche, potremo usare degli **encoder**, come ad esempio gli oggetti di classe **OrdinalEncoder** e **OneHotEncoder**.

Preprocessing dei dati (2)

Classe	Applicazione
SimpleImputer	Assegna un valore ad una feature mancante sulla base degli altri valori della stessa, secondo una strategia ben definita (ad esempio, la media).
IterativeImputer	Assegna un valore ad una feature mancante come funzione dei valori assunti dalle altre feature. Tale valore è assegnato usando un approccio di regressione (anche multivariata) tra le altre feature e la feature target.
MinMaxScaler	Normalizza i valori assunti dalle feature nel range che va tra un certo minimo ed un certo massimo.
StandardScaler	Standardizza i valori assunti dalle feature in modo da farli distribuire secondo una gaussiana a media nulla e varianza unitaria.
OrdinalEncoder	Codifica una feature categoriche in un range che va 0 ad $n - 1$, con n numero di possibili valori assunti dalla feature.
OneHotEncoder	Codifica una feature categorica mediante la tecnica del one hot encoding.

Domande?

42