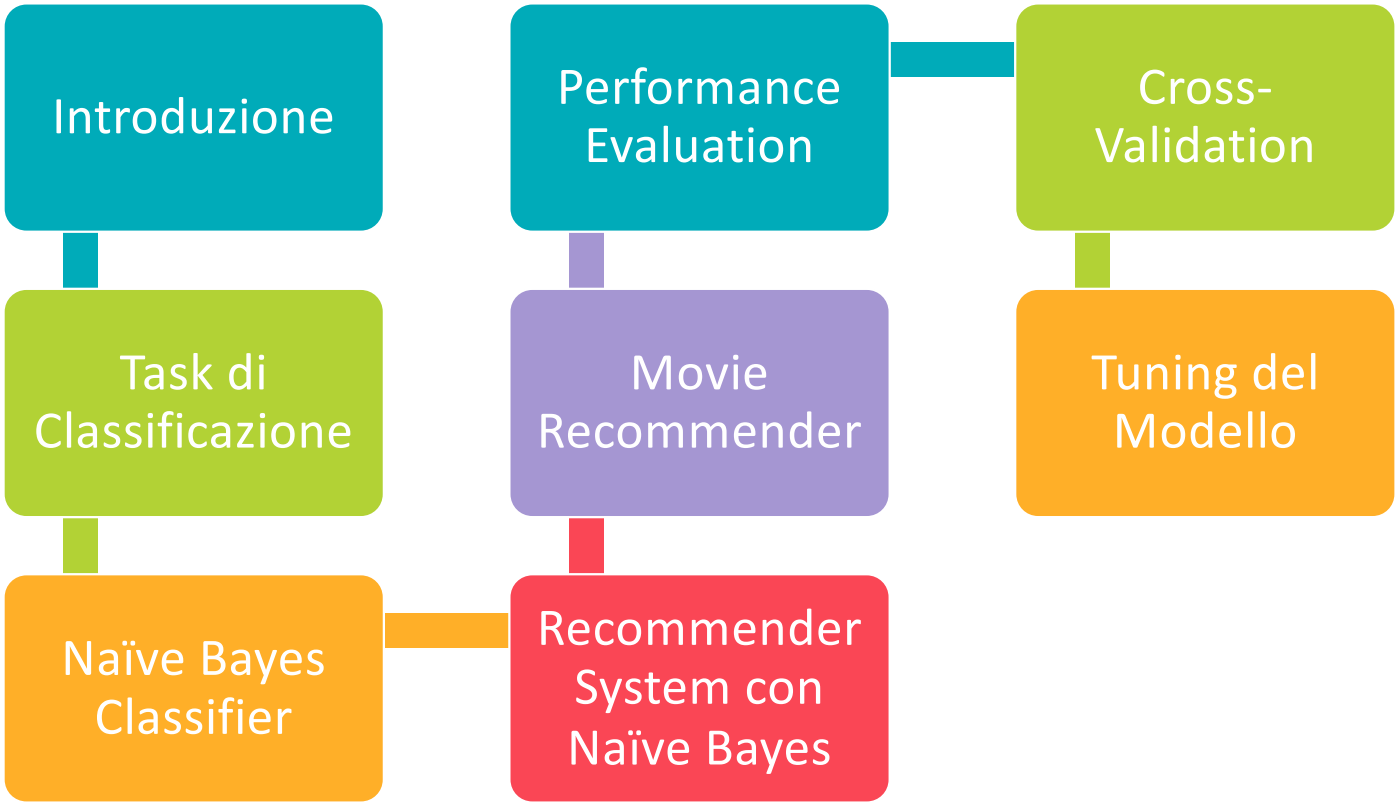
Machine Learning

Costruire un Movie Recommentation Engine con Naïve Bayes

Prof. Giuseppe Polese Prof.ssa Loredana Caruccio

Outline


Introduzione
Performance Evaluation
Cross-Validation
Task di Classificazione
Movie Recommender
Tuning del Modello
Naïve Bayes Classifier
Recommender System con Naïve Bayes

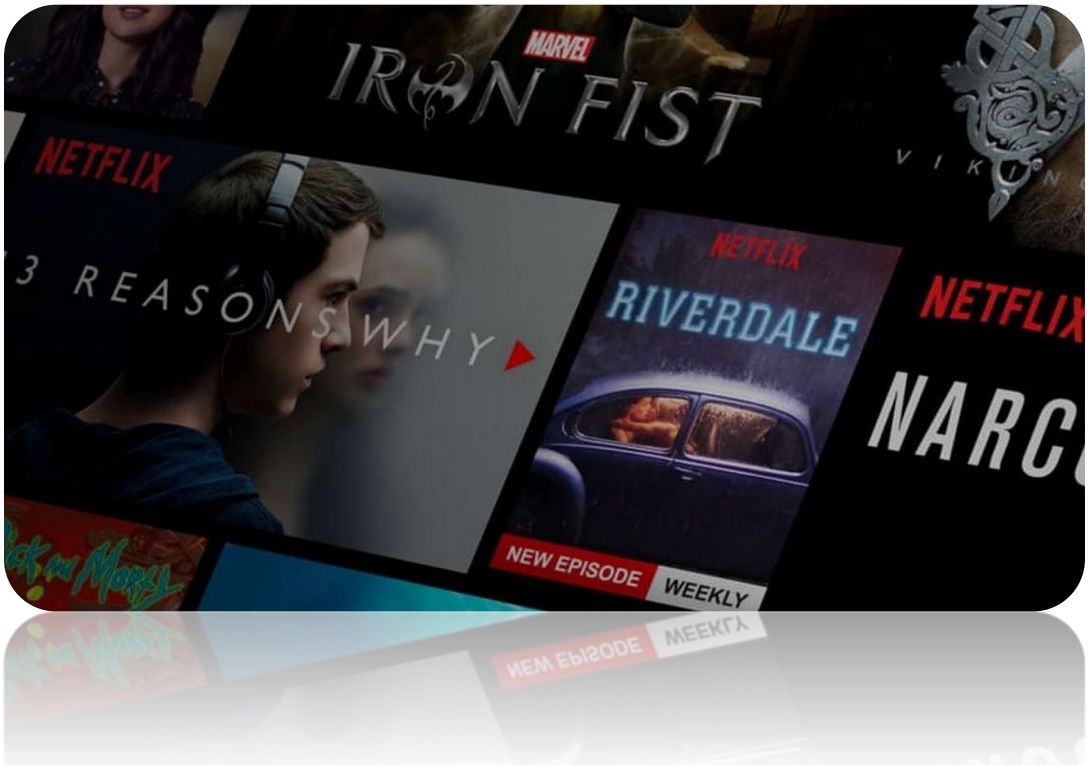
Introduzione

1

Introduzione

- In questa lezione affronteremo un task di **classificazione binaria**
  - Costruire un sistema di raccomandazione di film

Un buon punto di partenza imparare la classificazione da un esempio reale



Cos'è un Reccomender System

- Un **Recommender System** è un algoritmo che suggeriscono agli utenti elementi rilevanti per loro: film da guardare, per l'appunto, prodotti da acquistare, articoli da leggere.
- Il processo di raccomandazione, come nel caso della raccomandazione dei film, può essere inquadrato come **un problema di classificazione binaria**
  - Uno specifico film può piacere ad un utente?
  - Se si predice di si, allora sarà inserito nella sua lista di film consigliati
  - Altrimenti no!

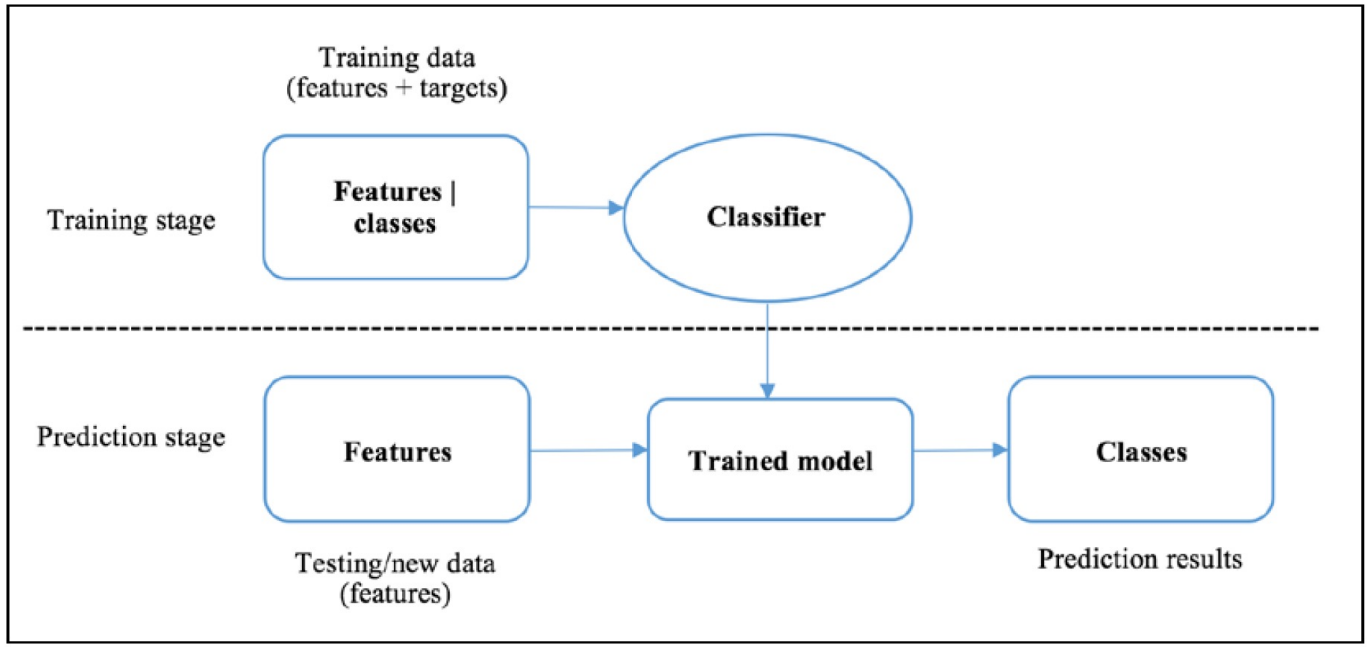
Task di Classificazione (1)

- La classificazione è uno dei principali esempi di apprendimento supervisionato.

*Dato un insieme di dati di addestramento contenente osservazioni e i relativi output, l'obiettivo della classificazione è imparare una regola generale che mappi correttamente le osservazioni (chiamate anche **feature** o **variabili predittive**) alle categorie di destinazione (chiamate anche **etichette** o **classi**).*

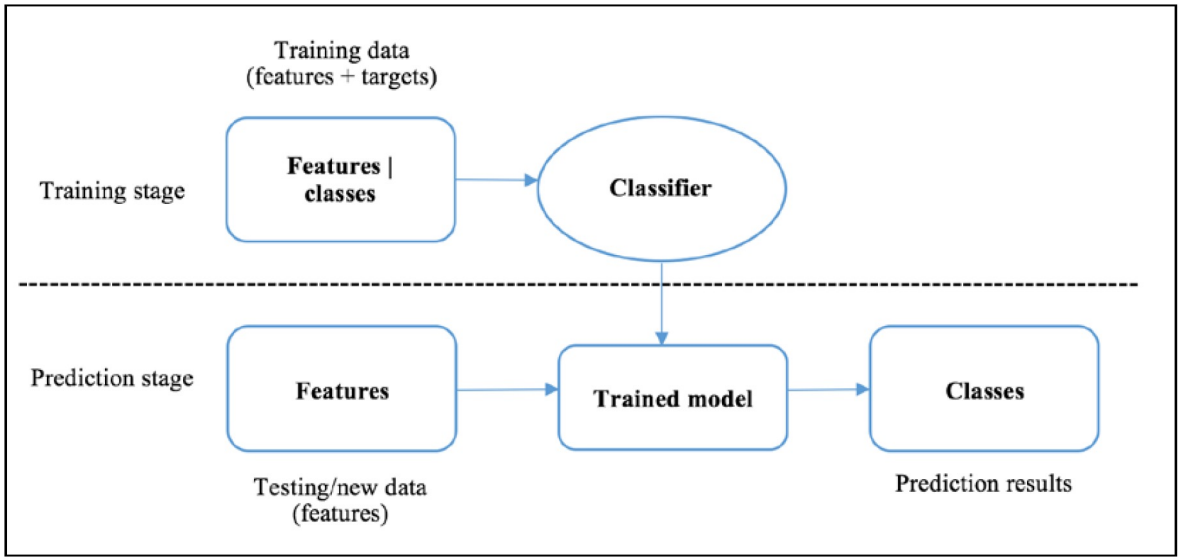
Task di Classificazione (2)

- Un modello di classificazione addestrato sarà generato dopo che il modello avrà imparato dalle **feature** e dalle etichette (**target**) dei dati di addestramento (**samples**)



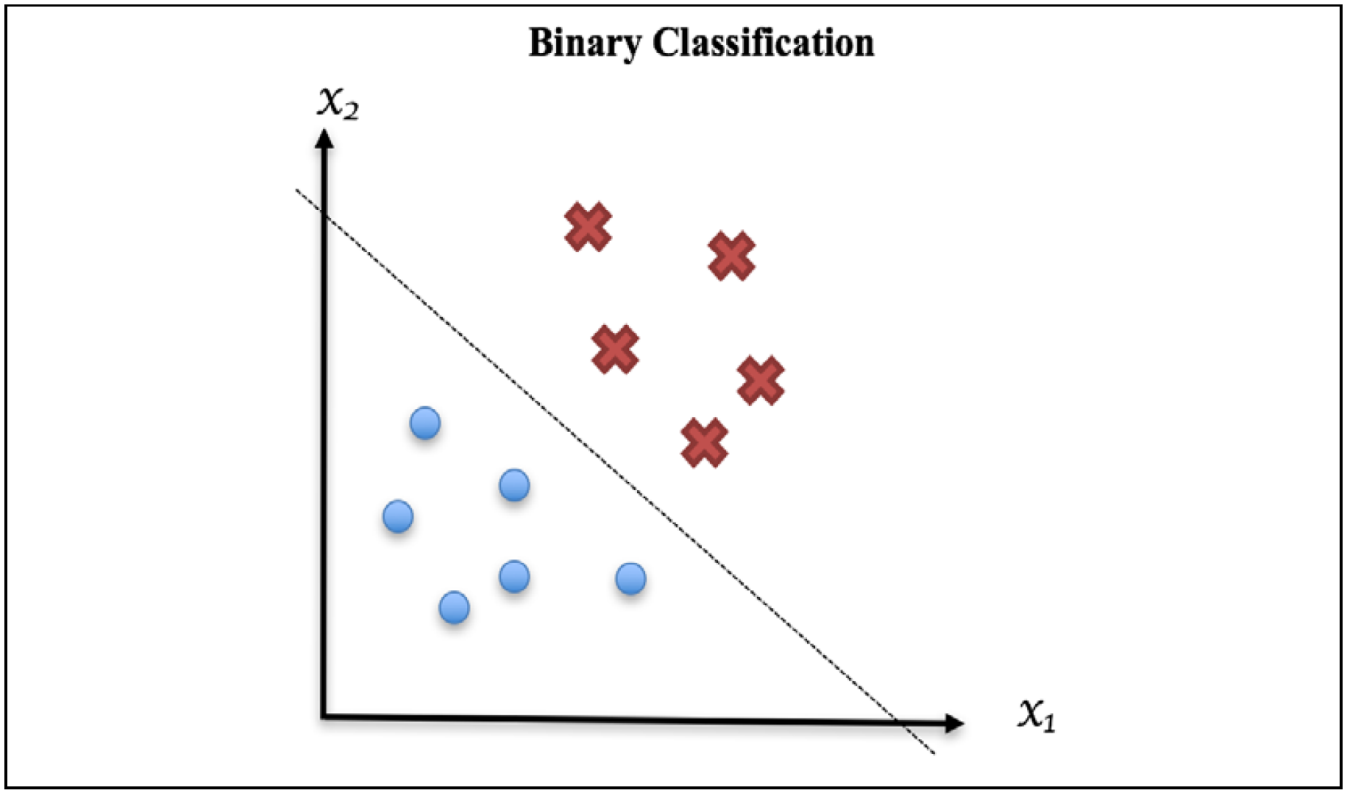
Task di Classificazione (2)

- Quando arrivano dati nuovi o non visti, il modello addestrato sarà in grado di determinare l'appartenenza alla classe desiderata.
  - Le informazioni sulla classe verranno predette in base alle caratteristiche di ingresso note, utilizzando il modello addestrato



Classificazione Binaria

La classificazione binaria cerca di trovare un modo per separare i dati in due classi.



Algortimi di Classificazione

- Per risolvere i problemi di **classificazione binaria** o altri tipi di classificazione, tra cui **classificazione multiclasse** o **classificazione multilabel**, i ricercatori hanno sviluppato molti potenti algoritmi di classificazione.
- Esempi di algoritmi che vengono spesso utilizzati sono:
  - Naïve Bayes,
  - support vector machine (SVM),
  - decision tree, e
  - regressione logistica.

Naïve Bayes

2

Un Classificatore Probabilistico (1)

- Il classificatore **Naïve Bayes** appartiene alla famiglia dei classificatori probabilistici.
  - Calcola la probabilità di ciascuna feature predittiva (detta anche attributo o segnale) dei dati che appartengono a ciascuna classe, per fare una predizione della distribuzione di probabilità su tutte le classi.
- Naturalmente, dalla distribuzione di probabilità risultante, possiamo concludere la classe più probabile a cui è associato ogni nuovo sample di dati.

Un Classificatore Probabilistico (2)

- Ciò che **Naïve Bayes** fa nello specifico, come indica il suo nome, può essere così descritto:
  - **Bayes**: mappa la probabilità delle feature di input osservate di una possibile classe alla probabilità della classe sugli input osservati in accordo al teorema di Bayes.
  - **Naïve**: semplifica il calcolo delle probabilità assumendo che le feature predittive siano reciprocamente indipendenti.

Teorema di Bayes (1)

- È importante comprendere il **Teorema di Bayes** prima di immergersi nel classificatore.
- Siano A e B due eventi.
  - Gli eventi possono essere che domani piova, l'estrazione di due re da un mazzo di carte o che una persona abbia il cancro o meno.
- Nel teorema di Bayes, $P(A|B)$ è la probabilità che A si verifichi dato che B è vero. Può essere calcolata come segue:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Teorema di Bayes (2)

- La probabilità che A si verifichi dato che B è vero può essere calcolata come segue:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

  - $P(B|A)$ è la probabilità di osservare B se si verifica A,
  - $P(A)$ e $P(B)$ sono rispettivamente la probabilità che A e B si verifichino.

Teorema di Bayes: Esempio 1

- Date due monete,
  - una è ingiusta, con il 90% dei lanci che ottiene testa e il 10% croce,
  - mentre l'altra è equa.
- Scegliete a caso una moneta e lanciatela.

Qual è la probabilità che questa moneta sia quella ingiusta, se si ottiene una testa?

Teorema di Bayes: Esempio 1

- La probabilità che sia stata scelta la moneta ingiusta quando si ottiene una testa, $P(U|H)$, può essere calcolata con la seguente formula:

$$P(U|H) = \frac{P(H|U)P(U)}{P(H)}$$

- Come sappiamo:
  - $P(H|U)$ è il 90%.
  - $P(U)$ è 0,5 perché scegliamo a caso una moneta su due.
- Come possiamo ricavare la probabilità di ottenere una testa, $P(H)$?

Teorema di Bayes: Esempio 1

- Ricavare la probabilità di ottenere una testa, $P(H)$, non è così semplice, poiché la probabilità che si ottenga una testa dipende da due eventi indipendenti,

$$P(H) = P(H|U)P(U) + P(H|F)P(F)$$

  - dove U è quando viene scelta la moneta ingiusta e F è quando viene scelta la moneta giusta
- Quindi $P(U|H)$ può essere calcolata come segue:

$$P(U|H) = \frac{P(H|U)P(U)}{P(H)} = \frac{P(H|U)P(U)}{P(H|U)P(U) + P(H|F)P(F)} = \frac{0{,}9 * 0{,}5}{0{,}9 * 0{,}5 + 0{,}5 * 0{,}5} = 0{,}64$$

Teorema di Bayes: Esempio 2

- Supponiamo che un medico abbia riportato il seguente test di screening del cancro su 10.000 persone

	Cancer	No Cancer	Total
Test Positive	80	900	980
Test Negative	20	9000	9020
Total	100	9900	10000

- Ciò indica che:
  - 80 pazienti oncologici su 100 sono diagnosticati correttamente, mentre gli altri 20 non lo sono.
  - Il cancro viene erroneamente rilevato in 900 persone sane su 9.900.

Teorema di Bayes: Esempio 2

- Se il risultato di questo test di screening su una persona è positivo, *qual è la probabilità che abbia effettivamente il cancro?*
- Assegniamo:
  - l'evento di avere un cancro come C
  - l'evento di avere risultati positivi Pos
- Quindi abbiamo:

$$\mathrm{P}(\mathrm{C}|Pos) = \frac{\mathrm{P}(Pos|\mathrm{C})\mathrm{P}(\mathrm{C})}{\mathrm{P}(Pos)}$$