Data Mining e Bioimmagini

**Classificatore di bayes, Network e reti neurali**

Immagine che contiene testo, lavagna, interno

Descrizione generata automaticamente

Prof. Pierangelo Veltri – 24/10/2023- Autori: Panarello, Saeedzadeh - Revisionatori: Panarello, Saeedzadeh

**C****LASSIFICATORE DI BAYES**

I classificatori di tipo bayesiano vengono utilizzati per la risoluzione di problemi di classificazione. In particolare, sono usati per il calcolo della probabilità condizionata: dati due eventi Y e X, se si conosce la probabilità di X dato Y (P(X|Y) probabilità condizionata di X su Y)utilizzo la parte contraria tramite il teorema di Bayes per tirar fuori la probabilità condizionata di Y dato X, e viceversa. Nell’immagine è riportata la formula.

**Esempio 1**

Dati:

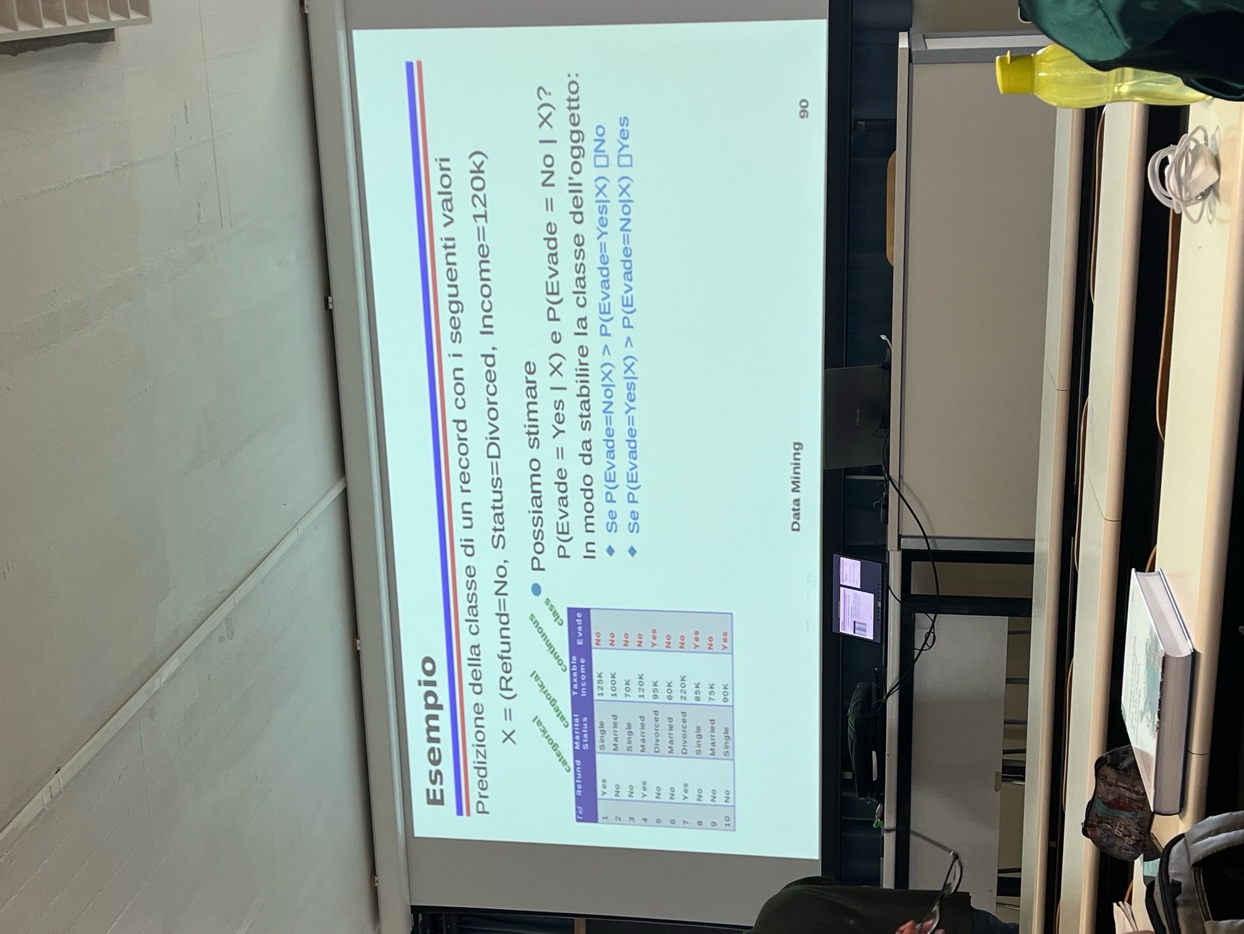
* ﻿﻿Un medico sa che la meningite provoca il torcicollo nel 50% dei casi (P(T|M)=0,5 ).
* ﻿﻿La probabilità che un paziente abbia la meningite è 1/50.000 (P(M)=1/50.000 )
* ﻿﻿La probabilità che un paziente abbia il torcicollo è 1/20 (P(T) = 1/20 )

Se un paziente ha il torcicollo, qual è la probabilità che abbia anche la meningite (P(M|T) )?

**Esempio 2**

Sia X la variabile casuale che rappresenta la squadra che ospita la partita e Y la variabile casuale che rappresenta il vincitore della partita. Sia X che Y possono assumere valori dall'insieme {0,1} (identificatori squadra).

﻿﻿Si assumano le seguenti probabilità:

* Probabilità Squadra 0 vinca: P(Y = 0): = 0,65
* Probabilità Squadra 1 vinca: P(Y = 1) = 1 - P(Y=0) = 0,35.
* Probabilità Squadra 1 ospiti la partita che ha vinto: P(×=1 | Y=1)=0,75.
* Probabilità Squadra 1 ospiti la partita vinta dalla squadra 0: P(X=1 I Y=0) = 0,3.

(Il problema non è completo e il prof non lo tratta, viene riportato per dare un’idea dell’esercizio)

**Esempio 3**

Il teorema di Bayes può essere utilizzato anche nel contesto della predizione; in particolare, permette di predire la classe di appartenenza di un dato elemento del dataset. Nell’esempio qui riportato si ha un dataset di possibili clienti di una banca di cui si vuole indicare l’affidabilità; tramite il calcolo della probabilità condizionata si può stimare l’appartenenza alla classe evasore o meno.

In ambito medico il prof suggerisce il caso del calcolo della probabilità condizionata che un individuo con la pressione alta abbia un infarto (partendo dal fatto che chi ha un infarto ha la pressione alta) oppure la probabilità che valori di colesterolemia alti siano legati ad infarto.

La classificazione si basa sempre sul record e sul dataset, così come sull’algoritmo e il suo addestramento. Importante è anche il tipo di dato: un valore discreto ha un’accuratezza minore rispetto ad un valore continuo.

Un altro approccio prevede il calcolo del *maximum a posteriori (MAP)*, che può essere usata per ottenere una stima puntuale di una quantità inosservata sulla base di dati empirici.

**NETWORK E RETI NEURALI**

Immagine che contiene testo, lavagna, interno, design

Descrizione generata automaticamenteData uno stimolo in ingresso (input) in un nodo sommatorio, in base all’intensità dello stimolo si avrà uno specifico output (stimolo in uscita). Per esempio, dati un mix di colori in ingresso restituisce bianco o nero in base a come si mescolano. Un altro caso in parte simile è il decoder in cui una serie di input vengono raccolti in un demultiplexer che traduce il segnale in uno stream di segnali della stessa ampiezza. La differenza è che mentre il decoder prende n segnali e ne restituisce n, la rete neurale elabora n segnali per ottenere un singolo valore binario (0/1, sano/malato).

I modelli matematici su cui si basa la rete neurale sono basati sulla combinazione degli input; per questo è necessario dare dei pesi adeguati ai segnali, così da poter elaborare correttamente l’output.

Ad esempio, una persona che si trova in prima fila manderà uno stimolo diverso rispetto a chi si trova in ultima fila, nonostante di base (idealmente) entrambi emettono lo stesso segnale. In base a questa differenza di stimolo è possibile distinguere i due suoni e stabilire la rilevanza dei segnali in ingresso. Questa differenza di peso viene trasformata in una formula matematica che permette di normalizzare i valori ottenuti rispetto ad una sogna ed ottenere così il risultato binario.

Immagine che contiene testo, lavagna, interno, innovazione

Descrizione generata automaticamenteIl modello di base della rete neurale prende ispirazione dal funzionamento dei neuroni biologici. I neuroni, infatti, possiedono strutture arboree chiamate **dendriti** che ricevono segnali da altri neuroni mediante giunzioni dette **sinapsi**. Alcuni neuroni comunicano mediante poche sinapsi, altri ne posseggono migliaia.

Generalmente le reti neurali funzionano come delle black box: ricevono in input dei dati e restituiscono l’output senza spiegare il procedimento e/o la ragione di tale risultato.

Ad esempio, nell’immagine vediamo un tipo di ANN (rete neurale artificale) cui regola è la seguente: l’output Y è 1 se almeno due dei tre valori in input sono uguali a 1. Il risultato viene restituito senza alcuna spiegazione.

Immagine che contiene testo, presentazione, interno, Schermo di proiezione

Descrizione generata automaticamenteAd ogni segnale di input viene associato un peso cosicchè si abbia una omogenizzazione dei dati e li si possa processare. Ad esempio, potremmo combinare (esempio sommatoria) tali dati per poi ricombinare i risultati parziali prima di restituire l’output. Questo livello nascosto di elaborazione, detto ***hidden layer***, permette di raffinare ulteriormente i risultati; tale elaborazione a strati costituisce il deep learning.

L’assenza di una motivazione del risultato è un limite; ad esempio, il medico necessità di sapere il perché il suo paziente viene classificato come malato. Per questa ragione, nell’ambito della classificazione basata sul deep learning, oggi si parla di explainable AI. In un contesto medico, questo tipo di IA permette non solo di effettuare una classificazione di tipo sano/malato, ma spiega il motivo di tale scelta; in tal modo, il medico riesce a validare il dato in modo molto più rapido ed efficace (la diagnosi ha maggior probabilità di essere corretta). Il training di tali AI può prevedere l’accesso ai dati raccolti nei presidi ospedalieri e nei centri di ricerca.

Immagine che contiene testo, presentazione, Visualizzatore, interno

Descrizione generata automaticamente

**SUPPORT VECTOR MACHINES**

La support vector machine si basa sulla scelta di uno o più elementi di divisione dello spazio; infatti, è necessario individuare un indicatore (IPERPIANO) che massimizzi la distinzione tra classi. Alla base della ricerca vi è il calcolo della distanza tra elementi mediante un metodo geometrico-matematico apposito.

Immagine che contiene testo, interno, presentazione, Schermo di proiezione

Descrizione generata automaticamenteI divisori sono lineari in un ambiente bidimensionale, un piano in uno spazio 3D, iperpiani in uno spazio n-dimensionale.

Più il training set è “voluminoso” e si avvicina al test set, più il margine risulta netto. È una tecnica relativamente semplice.

Immagine che contiene testo, interno, presentazione, Viso umano

Descrizione generata automaticamenteNon consente di gestire dati molto eterogenei in quanto gli attribuiti sono considerati tutti alla stessa stregua con un riferimento di dimensione. Per esempio, non fa distinzione tra un attributo con dato discreto (come affidabile/non affidabile, 0/1) con un attributo con dati in continuo (ricordiamo che i dati continui danno risultati più accurati). Non viene, quindi, catturata la normalità di un dominio rispetto alla continuità dell’altro.

Nella prossima lezione affronteremo meglio l’argomento.