GUIDA UTENTE

Introduzione:

Il software in questione è basato sull’algoritmo del KNN ( k-nearest neighbors ).   
L’algoritmo citato è un algoritmo di regressione, una delle aree di applicazione del data mining.   
Lo scopo degli algoritmi regressivi consiste nella predizione del valore di un attributo numerico associato a un esempio sulla base di valori osservati per altri attributi dell’esempio medesimo.  
Più nello specifico, l’algoritmo del KNN riceve come dati di input una collezione di esempi di apprendimento ( l’insieme di esempi è detto training set, dove ogni esempio è una tupla che contiene un insieme di variabili indipendenti e una sola variabile dipendente, detta variabile target ), un valore intero “k” maggiore di 0 ( che rappresenta il numero di distanze diverse da prendere in considerazione per il calcolo della predizione ) e un esempio di cui si vuole predire la variabile target ( l’esempio è composto da sole variabili indipendenti ed è detto query point ).

Il risultato generato dall’algoritmo del KNN è il valore numerico associato all’esempio ricevuto in ingresso corrispondente al valore che non era noto e si voleva predire.  
L’output dell’algoritmo dipenderà anche dal valore di k.  
La modalità di calcolo del risultato varierà in base al tipo di attributi presenti all’interno del training set: nel caso in cui gli attributi fossero discreti ( o letterali ) si procederà con il calcolo della distanza di Hamming ( nel caso in cui gli attributi presi in considerazione abbiano stesso valore la distanza varrà 0, altrimenti avrà valore 1 ), mentre nel caso in cui gli attributi fossero continui si procederà con l’algoritmo “minMax Scaler” ( il cui scopo è quello di adattare tutte le variabili indipendenti del training set e query point al range 0-1, mediante la formula:   
newValue = ( value – min ) / ( max – min ), dove min e max sono rispettivamente il minimo e il massimo valore della specifica colonna del training set ).

Modalità di utilizzo:

Il software prevede due versioni: una che utilizza l’interfaccia grafica e l’altra che viene eseguita sul terminale a linea di comando.

1. Interfaccia grafica ( GUI ):

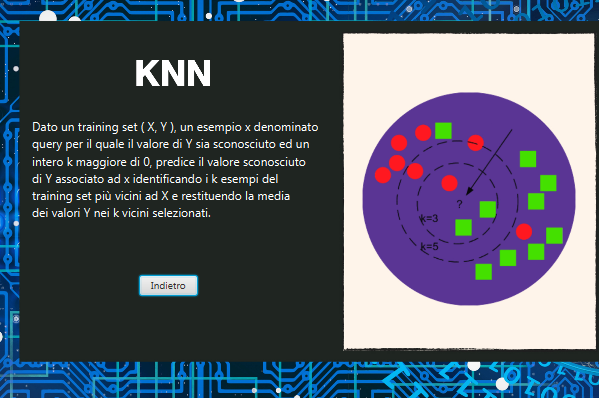
[ Pagina 1 ]  
Una volta aperto il software nella sua versione con l’interfaccia grafica ci ritroveremo sulla pagina iniziale. Su questa pagina l’utente può selezionare di connettersi al server mediante l’opportuno pulsante, oppure può accedere alla guida utente nella quale potrà visualizzare brevemente in cosa consiste il software.



( Pagina 1 - mainpage )

[ Pagina 2 ]

Nel caso in cui l’utente clicchi sul pulsante della guida utente ( presente sulla prima pagina ) si ritroverà sulla pagina nella quale viene spiegato brevemente l’algoritmo regressivo del KNN. All’interno della pagina sarà presente il pulsante per tornare indietro alla prima pagina.



( Pagina 2 – guida utente)

[ Pagina 3 ]  
Invece, nel caso in cui l’utente avesse cliccato sul pulsante per la connessione al server ( presente sulla prima pagina ) verrà portato nella pagina di connessione al server, nella quale dovrà inserire l’indirizzo IP e il Numero di porta giusti per potersi connettere correttamente. Per potersi connettere al server è necessario che il server sia già stato avviato. Nel campo indirizzo IP è possibile digitare, al posto dell’indirizzo della propria macchina sulla quale è in esecuzione anche il server, la stringa “localhost”. Il numero di porta per connettersi al server è 1010.

  
( Pagina 3 – connessione al server)

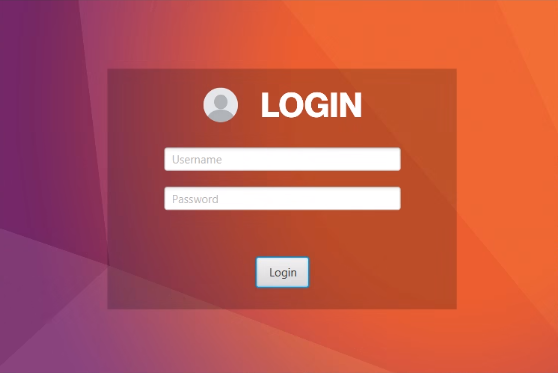
[ Pagina 4 ]  
Dopo essersi connessi al server, si potrà scegliere la modalità di acquisizione del training set tra: file di testo, file binario e database. Ciò avverrà mediante un click sull’opportuno pulsante presente sull’interfaccia. Una volta scelta la modalità, nel caso in cui l’utente selezionasse il file di testo o file binario, apparirà sulla stessa pagina una casella di testo nella quale si dovrà inserire il nome del file ( senza estensione ) al quale si fa riferimento per l’acquisizione del training set. In seguito bisognerà cliccare sul bottone “Conferma” in fondo alla pagina per poter confermare la propria scelta. In questo caso l’utente sarà reindirizzato alla pagina 5, mentre nel caso in cui l’utente selezionasse database, sarebbe reindirizzato alla pagina di login al database ( pagina 6 ).

  
( Pagina 4 –selezione training set )

[ Pagina 5 ]  
Nel caso in cui l’utente cliccasse il pulsante relativo al file di testo o al file binario nella pagina 4, verrà reindirizzato sulla pagina per l’acquisizione della variabile “k” e degli attributi discreti e/o continui ( in base agli attributi presenti nel training set selezionato ) del query point di cui si vuole predire la variabile dipendente ( o variabile target ).

  
( Pagina 5 –acquisizione query point )

[ Pagina 6 ]  
Invece, nel caso in cui l’utente avesse cliccato il pulsante relativo al database nella pagina 4, verrebbe reindirizzato sulla pagina di login. In questa pagina dovrà inserire le credenziali di un utente già registrato all’interno del database. Una volta compilati i due campi bisognerà cliccare sul pulsante di login presente sul fondo della pagina. Per poter registrare un nuovo utente sul database è necessario recarsi sulla barra di ricerca di Windows digitando *“cmd”,* cambiare la directory attraverso il comando “cd percorso”, spostandosi sulla cartella bin sulla quale è installato Mysql Server ( *C:\Program Files\MySQL Server 5.7\bin* ), autenticarsi con le credenziali di un utente già registrato nel database attraverso il comando *mysql –u ‘nomeutente’ root –p* e successivamente digitare la password dell’utente con il quale si vuole effettuare l’accesso.

  
( Pagina 6 –login database )

[ Pagina 7 ]  
All’interno della pagina corrente l’utente potrà selezionare la tabella nel database dalla quale vuole acquisire il training set. Gli utenti che in pagina 4 hanno selezionato file di testo o file binario hanno inserito all’interno della stessa pagina il nome della tabella dalla quale estrapolare il training set, mentre coloro che hanno scelto come modalità di acquisizione il database hanno dovuto prima loggarsi all’interno del database, per poi poter inserire il nome della tabella di riferimento per il calcolo della predizione. Una volta compilata la casella di testo con il nome della tabella, l’utente dovrà premere sul pulsante “Conferma” per essere reindirizzato alla pagina 5, nella quale potrà inserire il suo query point.

  
( Pagina 7 –acquisizione training set per database )

[ Pagina 8 ]  
Nell’ultima pagina del software verrà fornito il risultato finale dell’algoritmo del KNN ovvero la predizione della variabile target non nota.   
Una volta fatto ciò è possibile compiere una tra tre diverse operazioni:  
- cambiare il query point inserendone uno nuovo ma lavorando sempre sullo stesso training set selezionato in precedenza;  
- cambiare la modalità di acquisizione del training set ( per poi successivamente inserire un nuovo query point;  
- uscire dal software.   
Ciò sarà possibile mediante tre diversi pulsanti collocati nella parte bassa dell’ultima pagina.

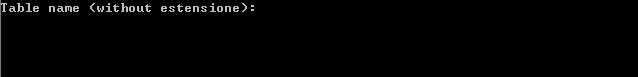
  
( Pagina 8 –risultato predizione )

2. Interfaccia a linea di comando ( CLI ):

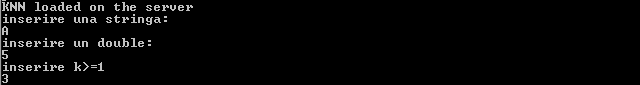
Una volta aperto il programma verrà chiesto all’utente di selezionare la modalità di acquisizione del training set tra file di testo, file binario e database. L’utente dovrà inserire il numero corrispondente alla scelta che vorrà effettuare.



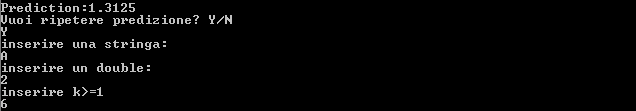
Successivamente verrà chiesto all’utente di inserire il nome del file ( senza estensione) dal quale si vuole acquisire il training set.



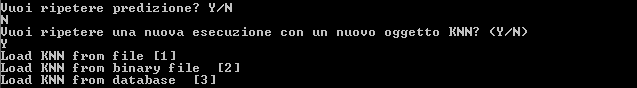
Una volta fatto ciò il training set selezionato verrà caricato e sarà disponibile per il calcolo finale del valore che si vuole predire. In seguito all’utente verrà chiesto di inserire i valori degli attributi discreti e/o continui ( in base ai valori degli attributi presenti nel training set ) e il valore di k ( corrispondente al numero di distanze che si vuole prendere in considerazione ).



Una volta inserito il valore di k il software genererà il risultato finale della predizione. In seguito verrà chiesto all’utente di ripetere la predizione. Nel caso in cui l’utente decidesse di ripetere la predizione digitando “Y”, verrà modificato il query point e, quindi, saranno richiesti nuovamente i valori degli attributi discreti e/o continui e un nuovo valore di k.



Invece, nel caso in cui l’utente dovessi digitare “N” alla richiesta di ripetere la predizione, verrà chiesto all’utente di cambiare la modalità di acquisizione del training set. Nel caso in cui dovesse rispondere nuovamente “N” il programma terminerà ( verrà chiuso solo il client, il server rimarrà in stato di esecuzione ), altrimenti ricomincerà da capo chiedendo la nuova modalità di acquisizione del training e, successivamente, il nuovo query point e valore di k.



Test:

A seguire sono stati eseguiti alcuni casi di test sui file presenti nel progetto.   
Nelle tabelle rappresentate sarà rappresentato il query point, il valore di k e il risultato della predizione.

File per acquisizione Training set: provac

|  |  |  |  |
| --- | --- | --- | --- |
| **ATTRIBUTO DISCRETO** | **ATTRIBUTO CONTINUO** | **K** | **PREDIZIONE** |
| A | 2 | 1 | 1.0 |
| A | 5 | 1 | 1.5 |
| B | 10 | 3 | 10 |
| A | 7 | 4 | 1.25 |
| B | 1 | 5 | 7.428571… |

Tabella per acquisizione Training set: provac ( del database )

|  |  |  |  |
| --- | --- | --- | --- |
| **ATTRIBUTO DISCRETO** | **ATTRIBUTO CONTINUO** | **K** | **PREDIZIONE** |
| A | 2 | 1 | 1.0 |
| A | 5 | 1 | 1.5 |
| B | 10 | 3 | 10 |
| A | 7 | 4 | 1.25 |
| B | 1 | 5 | 7.428571… |

Ovviamente i risultati generati con acquisizione tramite file di testo provac e il database provac sono identici, essendo identico il contenuto del file e della tabella.

File per acquisizione Training set: simple

|  |  |  |  |
| --- | --- | --- | --- |
| **ATTRIBUTO DISCRETO** | **ATTRIBUTO DISCRETO** | **K** | **PREDIZIONE** |
| A | A | 1 | 1.0 |
| A | B | 2 | 2.791666… |
| A | C | 1 | 3.346153… |
| B | C | 2 | 10 |
| A | B | 3 | 4.233333… |

File per acquisizione Training set: servo

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ATTRIBUTO DISCRETO** | **ATTRIBUTO DISCRETO** | **ATTRIBUTO CONTINUO** | **ATTRIBUTO CONTINUO** | **K** | **PREDIZIONE** |
| A | C | 3 | 4 | 2 | 2.846875… |
| E | E | 4 | 1 | 3 | 0.683336… |
| B | D | 2 | 8 | 1 | 4.099967… |
| C | C | 5 | 4 | 4 | 0.571877 |
| A | E | 6 | 6 | 3 | 0.514585… |