BIG DATA ANALYSIS
10/01/2017

Parte 0: Il Dataset

Il datataset contiene dati che descrivono i dipendenti di una impresa secondo le seguenti feature (l'ordine delle feature nel dataset potrebbe essere diverso):

1. Employee satisfaction level
2. Last evaluation
3. Number of projects
4. Average monthly hours
5. Time spent at the company
6. Whether they have had a work accident
7. Whether they have had a promotion in the last 5 years
8. Sales
9. Salary
10. Whether the employee has left

Scopo finale del dataset è predire se un dipendente abbandonerà o meno l'impresa.

Note:

Durata della prova:2 ore.
Creare una cartella esame sul desktop e scaricare in essa il file csv che si trova al link http://bit.ly/BDA10012017

Posizionarsi in quella cartella con la linea di comando e inizializzare ipython con il comando "ipython notebook"
Salvare frequentemente il file notebook creato attribuendogli il proprio nome-cognome.
Al termine della prova spedire a francesco.guerra@unimore.it il file html della prova (file / download as / HTML).

Parte 1: Analisi (10 punti)

1. Caricare il dataset e denominarlo con una variabile chiamata “dataset”

2. Quante sono le istanze contenute nel dataset? _____ Il dataset è completo (cioè per ogni istanza tutti i valori di attributo sono sempre specificati – non esistono “missing values”)? _____ Il dataset è bilanciato per quanto riguarda la classe da predire? _____ Osservando direttamente i dati, in che modo l’avere avuto una promozione negli ultimi 5 anni ha influito sulla scelta del dipendente di abbandonare l’impresa?

_____

_____

3. Guardando la media e la mediana si evince che i dipendenti che lavorano nell’impresa sono in generale soddisfatti? _____

_____

_____

Se si rappresenta un istogramma della soddisfazione dei dipendenti (usare il parametro bins=100 nella funzione che realizza l’istogramma) si osserva che c’è un blocco iniziale di dipendenti per nulla soddisfatti. Calcolare il valore di insoddisfazione di questo blocco di dipendenti _____

Che cosa hanno in comune questi dipendenti (a parte il livello di soddisfazione basso)?

_____

_____

4. Rappresentare in un grafico il salario dei dipendenti insoddisfatti descritti nel punto precedente.

5. Analizzare i settori (attributo “sales”) in cui i dipendenti insoddisfatti sono collocati. C’è un settore che è maggiormente fonte di disagio?

_____

_____

_____

_____

Parte 2: Trasformazione e Predizione (20 punti)

1. Scikit-learn utilizza un array numpy per effettuare le proprie predizioni. Gli elementi dell'array numpy devono essere delle stesso data type numerico. E' necessario pertanto trasformare i dati del dataset per renderli utilizzabili con scikit.

Creare un nuovo dataset dal precedente e chiamarlo reduced in si considerano unicamente le feature numeriche.

2. Nel dataset originale, trasformare i valori dei campi non numerici in numerici, utilizzando una opportuna funzione di trasformazione.

3. Si vuole predire il il fatto che un dipendente abbia lasciato o meno l'impresa (feature left)sulla base degli altri attributi presenti nel dataset. Dividere i due dataset (quello originale e quello ridotto) in modo che ¾ degli elementi siano contenuti in un nuovo dataset "train" e ¼ nel dataset "test".

Valutare l'accuracy ottenuta con il modello MultinomialNB su entrambi i dataset
(from sklearn.naive_bayes import MultinomialNB)

4. Il valore di accuratezza ottenuto è pari a _____

Cosa si scopre analizzando le confusion matrix?

_____

_____

_____

_____

5. Se si utilizza un modello basato su Decision Tree che valore di accuratezza si ottiene? Cambia qualcosa nella confusion matrix?

_____

_____

_____

_____

6. Che valore di accuratezza si ottiene con un 5 Fold cross validation e il modello basato su Decision Tree _____ e il modello basato su MultinomialNB _____

7. Creare un dataset bilanciato rispetto l'attributo "left", e verificare se cambia l'accuratezza con un 5 Fold cross validation e il modello basato su Decision Tree _____ e il modello basato su MultinomialNB _____

8. Creare une versione 1-of-V della matrice (indicator matrix) per quanto riguarda le colonne "sales" e "salary". Si tratta di una matrice in cui ogni colonna è trasformata in un numero di colonne pari alla cardinalità (i valori unici) del dominio della colonna originale. Ogni cella nella colonna assume un valore 0 o 1 a seconda del fatto che il valore inizialmente contenuto dall'attributo sia uguale a quello rappresentato dalla colonna
Calcolare l'accuratezza con i metodi precedenti. Migliora o peggiora? Questo tipo di trasformazione è in generale migliore o peggiore di quella effettuata al punto 2?

_____

_____

_____

_____

_____

_____

_____

_____

9. Raggruppare i valori dell'attributo "average_montly_hours" in 6 gruppi. Sostituire al valore originale dell'attributo un numero che va da 1 a 6 e che indica l'appartenenza allo specifico gruppo. Valutare l'accuratezza.

10. Si consideri un nuovo dataset ottenuto attraverso una variazione a piacere del precedente e si analizzi l'accuratezza ottenuta con un modello basato su Decision Tree. Che valore si ottiene? _____