BIG DATA ANALYSIS
08/01/2020

Parte 0: Il Dataset

Il datataset trainMobile.csv (preso da kaggle -- https://www.kaggle.com/iabhishekofficial/mobile-price-classification) contiene dati relativi a telefoni cellulare, utilizzando le seguenti feature:

battery_power: Total energy a battery can store in one time measured in mAh
blue: Has bluetooth or not
clock_speed: speed at which microprocessor executes instructions
dual_sim: Has dual sim support or not
fc: Front Camera mega pixels
four_g: Has 4G or not
int_memory: Internal Memory in Gigabytes
m_dep: Mobile Depth in cm
mobile_wt: Weight of mobile phone
n_cores: Number of cores of processor
pc: Primary Camera mega pixels
px_height: Pixel Resolution Height
px_widthPixel Resolution Width
ram: Random Access Memory in Mega Bytes
sc_h: Screen Height of mobile in cm
sc_w: Screen Width of mobile in cm
talk_time: longest time that a single battery charge will last
three_g: Has 3G or not
touch_screen: Has touch screen or not
wifi: Has wifi or not
price_range: This is the target variable with value of 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost).

Il dataset è costituito da attributi con valori numerici. La variabile da predire è price_range.

Parte 1: Analisi (10 punti)

1. Quante sono le istanze contenute nel dataset? _____ Il dataset è completo (cioè per ogni istanza tutti i valori di attributo sono sempre correttamente specificati – non esistono "missing values")? _____ Il dataset è bilanciato per quanto riguarda la classe da predire? _____ (punti 1).

2. La variabile sc_w assume valori discreti o continui? Analizzare la distribuzione dei valori e verificare se i telefoni costosi hanno mediamente una dimensione superiore di schermo. Verificare se eliminando gli elementi con sc_w uguale a 0 il risultato cambia. (punti 2)

3. E' vero che mediamente i telefoni meno costosi hanno anche una batteria meno potente? Realizzare 4 istogrammi (uno per ogni valore di price_range) che rappresentino la distribuzione dei valori di battery power per ogni categoria. (punti 3)

4. Verificare se tutti i telefoni che hanno il 4G hanno anche il 3G (punti 2)

5. Quanti sono i telefoni 4G che non hanno wifi e bluetooth? (punti 2)

Parte 2: Trasformazione e Predizione (20 punti)

1. Si vuole predire il valore di price_range sulla base degli attributi presenti nel dataset. Dividere il dataset in modo che 3/4 degli elementi siano contenuti in un nuovo dataset "train" e 1/4 nel dataset "test".

Allenare il train con il modello Decision Tree e valutare l'accuracy ottenuta calcolata sia sul dataset train sia sul dataset test. Confrontare i risultati ottenuti con quelli ottenuti con una predizione basata sul modello Logistic Regression. Effettuare alcune considerazioni sui risultati ottenuti, tenendo in considerazione anche l'analisi della confusion matrix. (punti 4)

2. Confrontare l'accuratezza ottenuta nel punto precedente con l'accuratezza si ottiene con un una 10 Fold cross validation. (punti 1)

3. Utilizzare la funzione di gridSearchCV per trovare i parametri migliori del classificatore decision tree. Agire sui parametri criterion, max_features e min_samples_split. Vericare se l'accuratezza che si ottiene con la nuova configurazione supera quella standard ottenuta al punto 1 (punti 4)

4. Utilizzare la funzione MaxAbsScaler per scalare i valori del dataset tra 0 e 1 e confrontare se l'accuratezza ottenuta con il Decision Tree Classifier migliora (punti 3).

5. Discretizzare il valore di ram in 4 intervalli e verificare se l'accuratezza ottenuta con il Decision Tree Classifier migliora (punti 2).

6. Creare una pipeline in cui il valore di ram sia discretizzato in 4 intervalli, il valore di battery_power sia discretizzato in 10 intervalli e poi il dataset venga ricondotto a valori nell'intervallo (0,1) e normalizzato con la funzione Normalizer. Si applichi poi un modello DecisionTree. (punti 4) [Alternativa (punti 2): non applicare la discretizzazione]

7. Si verifichi l'accuratezza ottenuta con il file test.csv. Controllare le colonne del file. I risultati corretti sono nel file class.csv. (punti 2).

Note:

Durata della prova: 2 ore. Creare una cartella esame e scaricare in essa il file csv che si trova al link
http://bit.ly/MB2020BDA

Creare nella cartella un jupyter notebook e rispondere nel file notebook alle domande. Indicare CHIARAMENTE nel notebook a quale domanda si sta dando una risposta.
Salvare frequentemente il file notebook creato attribuendogli il proprio nome-cognome.
Al termine della prova spedire a francesco.guerra@unimore.it il file della prova o il notebook direttamente o la versione html (file / download as / HTML).