BIG DATA ANALYSIS
26/01/2018

NOME	
COGNOME	
MATRICOLA	

Parte 0: Il Dataset

Il file bdastudents.csv (separatore ;) contiene una libera variazione del dataset "Student Perfomance" disponibile nell'UCI Machine Learning Repository
http://archive.ics.uci.edu/ml/datasets/Student+Performance
Il file rappresenta alcuni dati su studenti che frequentano 2 insegnamenti in 2 scuole diverse e il campo G3 la valutazione finale (da predire).
Lo schema del dataset è

School - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

sex - student's sex (binary: 'F' - female or 'M' - male)

age - student's age (numeric: from 15 to 22)

address - student's home address type (binary: 'U' - urban or 'R' - rural)

famsize - family size (binary: 0' - less or equal to 3 or 1 - greater than 3)

Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

guardian - student's guardian (nominal: 'mother', 'father' or 'other')

traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

failures - number of past class failures (numeric: n if 1<=n<3, else 4)

schoolsup - extra educational support (binary: yes or no)

famsup - family educational support (binary: yes or no)

paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

activities - extra-curricular activities (binary: yes or no)

higher - wants to take higher education (binary: yes or no)

internet - Internet access at home (binary: yes or no)

romantic - with a romantic relationship (binary: yes or no)

famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

freetime - free time after school (numeric: from 1 - very low to 5 - very high)

goout - going out with friends (numeric: from 1 - very low to 5 - very high)

Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

health - current health status (numeric: from 1 - very bad to 5 - very good)

absences - number of school absences (numeric: from 0 to 93)

G1 - first period grade (numeric: from 0 to 2)

G2 - second period grade (numeric: from 0 to 2)

G3 - final grade (numeric: from 0 to 2)

Note:

Durata della prova:2 ore. Dove possibile rispondere nel file notebook.
Creare una cartella esame e scaricare in essa il file csv che si trova al link
http://bit.ly/BDAgen2

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).

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 bilanciato rispetto alle scuole e ai generi degli studenti analizzati? (punti 1).

3. Creare un nuovo attributo “GRate” che misuri per ogni studente la differenza tra la valutazione ricevuta nel primo e nel secondo periodo (punti 2)

Realizzare un grafico che rappresenti per ogni età questa differenza.

_____

_____

4. Sono mediamente più bravi (attributo G3) i ragazzi o le ragazze? Esistono delle variazioni rilevanti nelle due scuole considerate? (punti 2)

_____

_____

_____

5. Tra i genitori degli studenti considerati, il livello di “educazione” maschile e femminile varia? Sono generalmente più scolarizzati i padri o le madri? Visualizzare poi un grafico che rappresenti il concetto (punti 2).

6. Indicare cosa visualizza l’istruzione

```
ds["G3"].groupby([ds["G3"],ds["address"]] ).count().plot()
```

Si tratta di una operazione significativa? (punti 3).

_____

_____

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 di tipo numerico. Creare un dataset chiamato “numeric” che contiene solo le features numeriche.

Creare poi un nuovo dataset “reduced” dall’originale con le colonne G1 e G2 e un dataset “lessReduced” togliendo da numeric unicamente le colonne G1 e G2(punti 1).

2. Si vuole predire G3 sulla base degli altri attributi presenti nel dataset. Dividere i dataset numeric, lessReduced e reduced in modo che 2/3 degli elementi siano contenuti in un nuovo dataset “train” e 1/3 nel dataset “test” (punti 2).

Valutare l’accuracy ottenuta con il modello LogisticRegression su tutti i dataset
(from sklearn.linear_model import LogisticRegression)

Il valore di accuratezza ottenuto è pari a _____ .La confusion matrix presenta qualche valore significativo (punti 1)?

3. Che valore di accuratezza si ottiene con un 10 Fold cross validation e il modello basato su Decision Tree _____

E' più affidabile la valutazione fatta con la cross validation o quella fatta con una suddivisione arbitraria del dataset in due parti, training set e test set? Per quale motivo? (punti 2).

_____

_____

4. Considerare il dataset numeric. Considerare l'intervallo di valori assunto dall'attributo age e dividerlo in tre parti. Associare a ogni istanza il valore 0,1,2 a seconda del fatto che l'età sia nel primo, nel secondo o nel terzo intervallo. Eliminare l'attributo age originale, non discretizzato e calcolare l'accuratezza con il metodo 10 cross fold validation.

_____

Trasformare la feature discretizzata in 3 feature booleane, una per ogni valore discretizzato. Il valore assegnato sarà 1 nella colonna che rappresenta il valore in esame. 0 nelle altre colonne. Calcolare l'accuratezza con il metodo 10 cross fold validation (punti 4).

_____

5. Aggiungere al dataset "numeric" gli attributi Mjob e Fjob il cui valore categorico deve essere mappato utilizzando una formula di conversione a scelta. Confrontare il risultato ottenuto con quelli ottenuti in precedenza (punti 4).

_____

_____

6. Partendo dal dataset originale, costruire due dataset contenenti solo le feature numeriche. Uno che rappresenti la scuola GP e l'altro la scuola MS. Costruire due modelli di predizione utilizzando il decision tree. Uno per gli studenti GP e l'altro per gli studenti MS. Allenare entrambi i modelli utilizzando 2/3 delle rispettive istanze come training.
Fondere i test in un unico file.
Verificare l'accuratezza ottenuta dal test in entrambi i modelli. Quale funziona meglio?
(punti 3)

_____

_____

7. Utilizzare un algoritmo di regressione da applicarsi al dataset del punto 1 per predire "G3". Arrotondare i valori ottenuti all'intero. Confrontare i risultati ottenuti con quelli ottenuti in precedenza (punti 3).

_____

_____