BIG DATA ANALYSIS
16/01/2018

Parte 0: Il Dataset

Il datataset BankMarketingDataSey.csv (formato csv con separatore ";") contiene dati di campagne di marketing attivate da una banca portoghese. Scopo dell'esercizio è predire se il cliente sottoscriverà un deposito (variabile y). Questi sono i campi del dataset (estratti da UCI http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#)

0 - id (numeric)
1 - age (numeric)
2 - job : type of job (categorical: 'admin.', 'blue- collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
3 - marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y', 'basic.6y', 'basic.9y',' high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')
8 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
9 - cons.price.idx: consumer price index - monthly indicator (numeric)
10 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
11 - euribor3m: euribor 3 month rate - daily indicator (numeric)
12 - nr.employed: number of employees - quarterly indicator (numeric)

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")? _____ (punti 1).

Il dataset è bilanciato per quanto riguarda la classe da predire? _____

3. Visualizzare la distribuzione delle età in uno specifico diagramma (punti 2)

4. Attraverso l'analisi del dataset è possibile ipotizzare se l'attributo "marital status" influisce nella predizione? In che modo influisce? Giustificare la risposta. (punti 3)

_____

_____

_____

_____

5. Calcolare una serie che rappresenti per ogni età la percentuale delle persone che hanno sottoscritto un deposito. Calcolare poi una serie che rappresenti per ogni età la percentuale delle persone che non hanno sottoscritto un deposito. Rappresentare graficamente le due serie (anche in diagrammi distinti e effettuare considerazioni sul risultato ottenuto – se ne nascono) (punti 4).

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.

Trasformare i valori categorici in valori numerici assegnando un valore specifico a ogni categoria di valori. In particolare, assegnare il valore 0 al valore "no" dell'attributo "y" e 1 al valore "yes".

Eliminare eventuali attributi che per qualche ragione (Specificarla) si ritiene essere inutili per la classificazione. Eliminare tutte le istanze per le quali c'è un attributo che assume valore unknown. (punti 3)

2. Si vuole predire la sottoscrizione di un deposito 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" (punti 1).

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

_____

_____

_____

_____

3. Utilizzare un altro modello di predizione e confrontare i risultati ottenuti. (punti 2)

_____

_____

_____

4. Confrontare l'accuratezza ottenuta nei punti 2 e 3 con accuratezza si ottiene con un 10 Fold cross validation

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 1).

_____

_____

_____

5. Creare un nuovo dataset “numeric” che contenga esclusivamente i valori numerici del dataset di partenza. Valutare l’accuratezza che si ottiene con questo dataset utilizzando i modelli proposti nei punti 2 e 3. (punti 2).

_____

_____

_____

_____

6. Aggiungere al dataset numeric gli attributi categorici in questo modo: per ogni attributo categorico costruire un numero di attributi booleani pari ai possibili valori della categoria. Ogni attributo rappresenta una categoria L’attributo booleano assume il valore 1 per le istanze del dataset originale che hanno il valore della categoria che l’attributo rappresenta. Valutare l’accuratezza che si ottiene con questo dataset utilizzando i modelli proposti nei punti 2 e 3. (punti 3).

_____

_____

_____

_____

7. Utilizzare un algoritmo di regressione da applicarsi al dataset del punto 1 per predire “y”. Arrotondare i valori ottenuti a 0 e a 1. Confrontare i risultati ottenuti con quelli ottenuti in precedenza (punti 4).

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/BDA2018GEN

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).