BIG DATA ANALYSIS
07/02/2017

Parte 0: Il Dataset

Il datataset new_weather.csv contiene dati che descrivono alcuni indicatori meteorologici di Seattle (temperatura, pressione atmosferica, punto di rugiada, …). I dati vengono codificati giornalmente e per ogni codifica viene inserito il valore minimo, massimo e la media del valore dell'indicatore. Il dataset presenta poi un campo ("Event") che evidenzia se in quella data c'è stato un fenomeno atmosferico.
Scopo dell'esercizio è predire se in una certa data si è avuto un evento atmosferico indipendentemente dalla sua natura.

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")? _____ Se esistono indicare quali sono e se l'assenza di questi valori pregiudica la possibilità di predire quanto richiesto (punti 1).

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

3. Creare un nuovo attributo "Range_Temperature" che misuri l'escursione termica giornaliera. Si osserva qualcosa di particolare nella distribuzione dei valori? (punti 3)

_____

_____

_____

_____

4. Rappresentare in un grafico la temperatura massima e la temperatura minima (punti 2).

5. Quanti eventi atmosferici accadono nel 2015? Come sono distribuiti per tipologie nei mesi? Rappresentare con il grafico più opportuno questa distribuzione (punti 2).

6. Realizzare una tabella pivot

http://pandas.pydata.org/pandas-docs/stable/generated/pandas.pivot_table.html

attraverso la quale mostrare per ogni anno, per ogni mese, e per ogni tipo di evento il valore medio della temperatura misurata. Mostrare il risultato in un diagramma (punti 2).

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 quindi un nuovo dataset dal precedente e trasformare i valori del campo "Events" in 1 / 0 a seconda del fatto che si registri o meno un evento.

2. Si vuole predire il il fatto che ci sia o meno un evento sulla base degli altri attributi presenti nel dataset. Dividere il dataset in modo che 2/3 degli elementi siano contenuti in un nuovo dataset “train” e 1/3 nel dataset “test” (punti 1).

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

3. Il valore di accuratezza ottenuto è pari a _____ (punti 1).

4. Cosa si scopre analizzando le confusion matrix? (punti 2)

_____

_____

_____

_____

5. Se si utilizza un modello basato su Decision Tree che valore di accuratezza si ottiene? Cambia qualcosa nella confusion matrix? (punti 2)

_____

_____

_____

_____

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

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).

_____

_____

_____

7. Provare e confrontare graficamente alcuni algoritmi di machine learning scelti dalle librerie di scikit (punti 2).

8. Creare un nuovo dataset “reduced” con esclusivamente i valori medi degli indicatori atmosferici. Per ogni feature, dividere il range dei valori 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 ottenuta con un qualsiasi algoritmo utilizzato in precedenza (punti 3).

_____

_____

_____

_____

9. Dividere nuovamente il dataset in modo che 3/4 degli elementi siano contenuti in un dataset "train" e 1/4 nel dataset "test" (punti 5).

Considerare il Train. Dividerlo in 3 nuovi dataset.
Month 1,2,3,4 --> dataset1
Month 5,6,7,8 --> dataset2
Month 9,10,11,12 --> dataset3

Eliminare l'attributo Month dai 3 dataset di training.
Eliminare l'attributo Month dal dataset di test

Allenare 3 classificatori di tipo decisionTree con i 3 dataset.

Effettuare la predizione del test con un sistema di votazione: la classe maggiormente predetta nei 3 dataset è quella che viene selezionata.
Calcolare l'accuratezza dell'approccio.

10. Effettuare (e giustificare) una variazione a piacere del dataset e calcolare l'accuratezza ottenuta dal nuovo dataset (punti 2).

_____

_____

_____

_____

Note:
Durata della prova:2 ore. Dove possibile rispondere nel file notebook
Creare una cartella esame sul desktop e scaricare in essa il file csv che si trova al link http://bit.ly/BDA07022017

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).