COVID19 PREDICTION

Alessandro Mastrorilli : 697586

INTRODUZIONE

Il sistema lavora su due task :

1) **Classificazione :** predire se un individuo in base alle feature di input possa essere positivo al covid19. A tal proposito vengono utilizzati diversi classificatori le cui prestazioni verranno confrontate attraverso la tecnica della k – fold cross validation(k= 10). Per ogni fold verranno salvate le seguenti metriche : precision , recall , average precision , f1 score. Al termine della k fold cross validation , il sistema restituisce le medie di ciascuna metrica.
2) **Task di inferenza probabilistica:** una volta trovata la migliore struttura della Belief Network , attraverso il metodo dell' eliminazione di variabili, il sistema risponde a query probabilistiche del tipo P(query|evidenza).

L' elemento che accomuna entrambi i task è l' utilizzo della Belief Network.

OSSERVAZIONI SUL DATASET

Il dataset contiene informazioni inerenti a 3128 individui in merito alla variante brasiliana del coronavirus.

Link sorgente : https://www.kaggle.com/saurabhshahane/brazilian-covid-symptomatic-patients-data

In particolare il dataset viene descritto da 11 feature booleane , ovvero :

- throat pain (mal di gola)
- dyspnea (dispnea)
- fever (febbre)
- cough (tosse)
- headache (mal di testa)
- taste disorders (disturbi del gusto)

- olfactory disorders (disturbi olfattivi)
- coryza (rinite)
- gender (sesso)
- health professional(medico)
- class (feature target)

STRUMENTI UTILIZZATI

E' stato utilizzato come linguaggio di programmazione python e in particolare l' utilizzo delle librerie scikit-learn(per la classificazione , la k fold cross validation e il calcolo delle metriche) e pgmpy(per la costruzione e la ricerca della migliore Belief Network).

CLASSIFICAZIONE

Sono stati utilizzati i seguenti classificatori per risolvere un task di apprendimento supervisionato :

- KNN
- Random Forest Classifier
- Decision Tree Classifier
- SVM
- Belief Network

Per quanto riguarda i primi 4 classificatori sono stati utilizzati i parametri di default in particolare :

- Per il KNN : si utilizzano 5 vicini , tutti i punti in ciascun vicino sono pesati equamente , come metrica di default(per misurare la vicinanza degli esempi) viene utilizzata la distanza euclidea.
- Per il Decision Tree Classifier : per scegliere le condizioni per il partizionamento viene scelta una misura che minimizza l' indice di Gini.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Pi : proporzione dei dati appartenenti a quella classe

Inoltre non viene indicata alcuna profondità , il minimo numero di esempi richiesto per partizionare un nodo interno è 2 e il numero minimo di esempi per essere in un nodo foglia è 1.

- Per il Random Forest Classifier : il numero di alberi nella foresta è 100 , per il resto gli altri parametri sono identici a quelli del Decision Tree Classifier.
- Per l' SVM : si utilizza come funzione kernel rbf

$$K(x, x') = e^{-\gamma||x-x'||^2}$$

Dove $\gamma = 1/(n_{feature} * X.var())$

Per modellare la rete utilizzo i seguenti passaggi :

- Attuo una ricerca basata su score volta a minimizzare

$$-\log_2 P(E \mid m) + |m| \cdot \log_2(|E|)$$ (indice di BIC).

- Trovo il modello con il migliore score attraverso la HillClimbSearch
- Stimo la migliore struttura trovata
- Addestro la rete sui dati di training in base al miglior modello trovato

MEDIA DELLE PRESTAZIONI DEI CLASSIFICATORI

Terminata la K fold Cross Validation , il sistema ha restituito le seguenti medie:

```
Media delle metriche del RandomForest
Media  Average Precision: 0.819462
Media Precision: 0.846533
Media Recall: 0.921442
Media f1: 0.882029

Media delle metriche del KNN
Media  Average Precision: 0.816321
Media Precision: 0.838209
Media Recall: 0.936719
Media f1: 0.883904

Media delle metriche del DecisionTree
Media  Average Precision: 0.840651
Media Precision: 0.866590
Media Recall: 0.929914
Media f1: 0.896832

Media delle metriche del SVM
Media Average Precision: 0.819462
Media Precision: 0.846533
Media Recall: 0.921442
Media f1: 0.882029

Media delle metriche della BN
Media Accuracy: 0.783692
Media Precision: 0.823091
Media Recall: 0.876704
Media f1: 0.848443
```

Se si prendono in considerazione le performance dell' Average Precision , il classificatore migliore è stato in questo caso il Decision Tree Classifier.

TASK DI INFERENZA

La rete viene modellata in base ai passaggi descritti precedentemente , dopodichè quest'ultima verrà utilizzata per risolvere delle query poste dall'utente data un' evidenza , mediante l' utilizzo di un algoritmo di eliminazione di variabili.
In particolare l' utente scriverà le variabili di query (come ipotesi) scegliendole dalle feature del data set , dopodichè scriverà le variabili di da inserire nel ' corpo' dell' evidenza assegnando un valore booleano.

La Belief Network , come anche visto precedentemente, può essere utilizzata come forma di apprendimento supervisionato , ovvero deve apprendere la distribuzione P(Y|Xi) dove :

- Y: rappresenta la feature target
- Xi : Rappresenta la feature di input

Esempio :

Data la seguente distribuzione di probabilità :
P(class| throat_pain:1 dyspnea:0 fever:1 cough:1 headache:0 taste_disorders:0 olfactory_disorders:0 coryza:0 gender:0)

La probabilità che un individuo con quelle caratteristiche possa essere positivo o meno al covid19 sono le seguenti :

class	phi(class)
class(0)	0.4587
class(1)	0.5413