UNIVERSITÀ
DEGLI STUDI
FIRENZE

Scuola di Ingegneria
Corso di Laurea Triennale in
Ingegneria Informatica
Dipartimento di
Ingegneria dell'Informazione

Analisi e Implementazione Voted Perceptron

De Simone Mario
Matricola 70035480

Anno Accademico 2020/2021

Relazione Voted Perceptron

Il progetto si propone l'obbiettivo di implementare e testare il funzionamento del Voted-Perceptron, una variante dell'algoritmo Perceptron descritta in Freund & Schapire 1999.

L'implementazione effettiva in linguaggio Python segue alquanto fedelmente la versione introdotta durante il corso, per questo non verrà analizzata in questa relazione.
Ha maggiore interesse, analizzare il comportamento dell'algoritmo al variare del Dataset sul quale lavora.

I tre Dataset implementati differiscono per volume di dati:

- Dataset 1 (QSAR ANDROGEN RECEPTOR) presenta 1024 attributi binari per 1687 istanze

- Dataset 2 (QSAR ORAL TOXICITY) presenta 1024 attributi binari per 8992 istanze

- Dataset 3 (QSAR BIODEGRADATION), presenta 41 attributi non binari per 1055 istanze

Come si può notare, sono stati scelti Dataset seguendo una logica che permetta una varia sperimentazione, nel dettaglio ne utilizziamo: uno con un basso numero di istanze e di attributi (3), un altro con un basso numero di istanze ma un alto numero di attributi (1) e l'ultimo avente un alto numero di istanze e di attributi (2).

Come si può osservare, i Dataset implementati si differenziano per cardinalità di attributi e istanze.
Tale distinzione è fondamentale per comprendere come effettivamente i tempi di esecuzione scalino a dismisura con il volume di dati, fino addirittura a diventare intrattabile.

Ecco alcuni tempi:

- Dataset 3, 80% train, 10 ripetizioni, 1 epoca: 10.075206995010376s
- Dataset 3, 80% train, 10 ripetizioni, 2 epoche: 19.486937761306763s
- Dataset 3, 80% train, 10 ripetizioni, 3 epoche: 27.325947761535645s

- Dataset 1, 80% train, 10 ripetizioni, 1 epoca: 366.2578971385956s
- Dataset 1, 80% train, 10 ripetizioni, 2 epoche: 634.2080867290497s
- Dataset 1, 80% train, 10 ripetizioni, 3 epoche: 1021.2671251296997s

- Dataset 2, 80% train, 10 ripetizioni, 1 epoca: 8702.229641199112s
- Dataset 2, 80% train, 10 ripetizioni, 2 epoche: 14857.505167007446s
- Dataset 2, 80% train, 10 ripetizioni, 3 epoche: non provato

Queste misure giustificano 2 fatti:
- Non è stato provato Dataset 2, 80% train, 10 ripetizioni, 3 epoche a causa del tempo computazionale che avrebbe richiesto
- Nel seguito della trattazione, per motivi di convenienza temporale, i dati mostrati sono ricavati da sperimentazione su Dataset 3, ma al netto di differenze dovute alla randomizzazione ed alla conseguente distribuzione di positivi e negativi, si riflettono allo stesso modo anche negli altri due dataset, senza perdita di generalità.

Le seguenti tabella sono costruite per 10 ripetizioni di train/predict, con percentuale di addestramento sul dataset (3) dell' 80% e numero epoche massime da 1 a 3.

	Accuracy		Precision	
	Mean	**Std. Deviation**	**Mean**	**Std. Deviation**
1	0.6587677725118484	0.03071441089292826	0.20714285714285713	0.23112822629805635
2	0.6862559241706162	0.044052891141695714	0.57375	0.41125208731505264
3	0.7450236966824646	0.0569071332823731	0.7155984809997967	0.12597312182519468

	F1		Recall	
	Mean	**Std. Deviation**	**Mean**	**Std. Deviation**
1	0.025379576999353736	0.03451194462326357	0.013721323017097664	0.01861861676914514
2	0.1545319211680768	0.2310074040524611	0.12058945389440745	0.21254937185042752
3	0.462886554470329	0.2660207640615485	0.4077474339225528	0.2830597793529659

NOTA: In alcuni casi (soprattutto nel caso di basso numero di addestramento/predizione) è possibile che F1, Recall e Precision abbiano media e deviazione standard pari a 0, questo si verifica nei casi dove la rispettiva formula presenta una forma di indecisione del tipo $\frac{0}{0}$.
In tali casi, mediante il parametro zero_division del metodo, viene imposto "return 0", di default il programma avrebbe anche lanciato un warning.

OSSERVAZIONI FINALI: Dal punto di vista teorico è lecito aspettarsi che all'aumentare del numero massimo di epoche, l'accuratezza migliori (misura di cui era stata chiesta l'analisi).
La sperimentazione conferma quanto appena supposto: la media dell'accuratezza tende ad aumentare.
Inaspettatamente anche la deviazione standard tende ad aumentare, invece di diminuire, questo fatto assume un senso e non si pone in contrapposizione a quanto supposto in apertura, se consideriamo che a causa della randomizzazione dei dataset di training e predicting, si possano avere casi più o meno "fortunati" in termini di costruzione di iperpiani.