Descrizione codice tesina Machine Learning

Luca Biggio

29 gennaio 2018

Il codice è costituito da 4 componenti:

- *pyt2.py* carica i dati di training (dividendoli in due parti di egual numero per produrre un validation set) e di test, normalizza le features dei datasets e trasforma i dati nel formato richiesto dal programma che implementa la rete neurale.
- *network2.py* implementa la rete neurale: permette di impostare il numero di layers e il numero di neuroni per layer; sviluppa l'algoritmo *stochastic gradient descent* e l'algoritmo di *backpropagation*; permette di monitorare l'andamento della fase di training in funzione del numero di epoche trascorse.
- *pyt2.py* avvia l'algoritmo *stochastic gradient descent* permettendo di impostare i vari iper-parametri. Produce diversi grafici utili per monitorare l'andamento della fase di training e per valutare la bontà della classificazione.
- *run.py* permette di avviare l'intero programma (mediante l'istruzione *execfile('run.py')*).

Il dataset a mia disposizione è costituito da:

- 1000 dati di *Segnale*
- 1000 dati di *Background*
- 100000 dati unlabeled, in cui sono presenti molti dati di background e pochi dati di segnale

Per comporre il validation set ho selezionato 500 dati di segnale e 500 dati di background dai file *Signal.txt* e *Background.txt* e ho creato così il file *validation_set.txt*. I restanti 1000 dati (di cui 500 di segnale e 500 di background) sono utilizzati per l'addestramento della rete e sono stati inseriti nel file *Training_set2.txt*. I dati del test set sono contenuti nel file *data.txt*.

A ogni dato in ciascun dataset sono associate 7 features, ognuna delle quali rappresenta il risultato di una misura di una quantità fisica di interesse. Siccome l'obiettivo è classificare un segnale che è caratterizzato da un valore molto preciso di una delle sette features (la massa invariante: l'ultima delle 7), in fisica delle particelle un approccio ricorrente nella ricerca di segnali rari coperti da rumore è di non utilizzare tale features nella fase di training. Essa verrà infatti utilizzata come ulteriore criterio per stabilire se la rete neurale ha effettivamente selezionato eventi che corrispondono a segnale. La massa invariante associata al segnale cercato è calcolata teoricamente e vale 5366.77 MeV.