Università degli studi di
Milano-Bicocca

Advanced Machine Learning

Assignment 1

Autore:
Federico Manenti - 790032 - f.manenti3@campus.unimib.t

19 ottobre 2019

Indice

1 Introduzione

Il dataset utilizzato è chiamato *train.csv* e contiene informazioni riguardanti i titolari di carta credito di una banca Taiwanese, è composto da 27000 righe e 25 variabili:

- $LIMIT\_BALL$ è il plafond che la banca concede al cliente espresso in NT (dollari taiwanesi)
- SEX è il sesso e assume valori 1 per uomo e 2 per donna
- $EDUCATION$ indica il grado di istruzione
- $MARRIAGE$ (1 = sposato; 2 = single; 3 = divoziato; 0 = altro).
- AGE è l'età del cliente
- PAY (1-6) indica lo storico dello stato dei pagamenti da Aprile (PAY_6) a Settembre (PAY_1) 2005. (-2 = nessun uso della carta; -1 = pagato per intero; 0 = ha scelto di usare il revolving credit; 1-9-... = numero di mesi scelti per posticipare il pagameto)
- $BILL\_AMT$ (1-6) è l'estratto conto (come sopra 1 settembre ... 6 aprile) che può assumere valori negativi nel caso in cui il cliente abbia pagato più del dovuto
- PAY_AMT (1-6) è l'ammontare dei pagamenti in riferimento al mese precedente (come sopra 1 settembre ... 6 aprile)

- *default.payment.next.month* indica il comportamento del cliente (1 = non ha pagato; 0 = ha pagato)

2 Preprocessing

Da una prima analisi esplorativa si nota l'assenza di valori mancati, l'asimmetria delle variabili continue *BILL_AMT 1-6* e *PAY_AMT 1-6* e una forte correlazione tra le variabili *BILL_AMT 1-6*. Quest'ultima è presumibilmente causata dal fatto che il totale dell'estratto conto varia di mese in mese in modo non significativo.

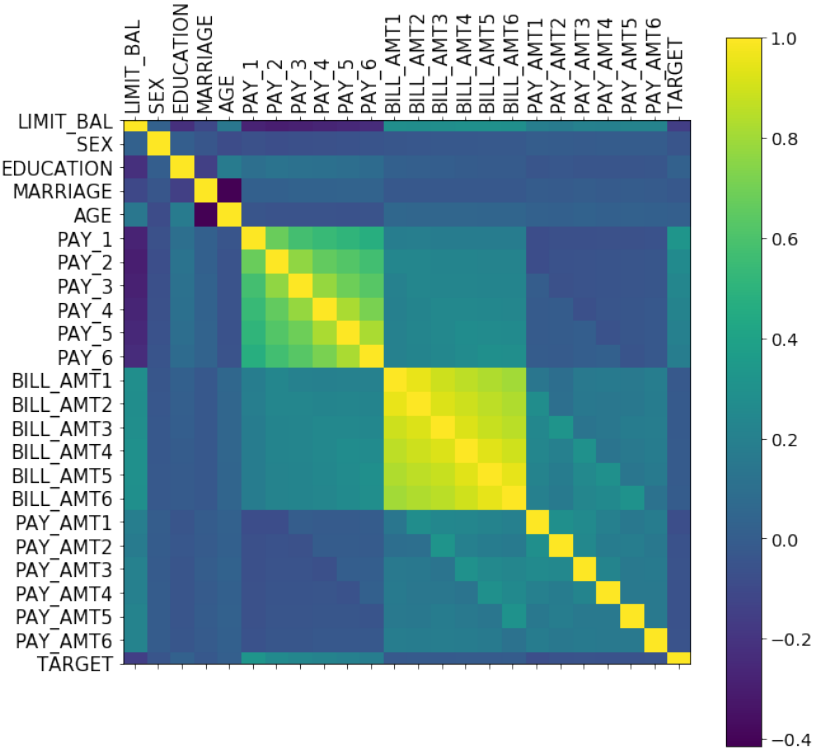

Figura 1: Matrice di correlazione

Per affrontare il problema dell'asimmetria sono stati sostituiti i valori negativi degli attributi con 0, non essendo interessati a considerare le posizioni in

credito verso la banca, e successivamente è stata applicata una trasformazione logaritmica del tipo $ln(x+1)$ dove x rappresenta la variabile in esame.
Per risolvere il problema della correlazione invece è stata calcolata la media tra le variabili *BILL_AMT 1-6*.
Infine è stata binnata e poi binarizzata la variabile AGE.

3 Split train e test

Un'ulteriore problema apparso è lo sbilanciamento della variabile obiettivo (28% *vs* 72%) per risolverlo quindi è stato deciso di applicare un *downsampling* dopo lo split tra train e test set. Infine i dati sono stati trasformati in tensori in modo da poter essere utilizzati con $Keras$ (X_train e X_test sono le feature, y_train e yy_test le etichette).

4 Rete Neurale

L'architettura della rete neurale utilizzata è composta da tre coppie di layer *Dense* e *Dropout* con un layer *Dense* finale per la divisione in classi desiderata. I layer di *Dropout* sono stati utilizzati per evitare l'overfitting. I layer *Dense*, ad eccezione di quello finale, sono composti tutti da 32 neuroni e utilizzano come funzione di attivazione la *Relu*. Lo strato finale ha come funzione di attivazione la *Sigmoide* perchè designato alla previsione dei dati ed è composto di un unico neurone poiché la variabile obiettivo possiede una shape di $(n^o di\ dati, 1)$. La *loss function* utilizzata è la *binary_crossentropy* che da letteratura risulta la migliore per un task di classificazione binaria. L'ottimizzatore scelto invece è $ADAM$ con parametri di default. Come metrica oltre alla classica *Accuracy* è stato scelto di osservare anche l'$F - measure$ per la classe obiettivo minoritaria. La scelta è stata dettata sia dal fatto che, essendo dati finanziari, la banca sia interessata maggiormente a prevedere correttamente i clienti che non pagheranno rispetto ai paganti, sia dallo sbilanciamento della classe obiettivo. Durante il fit del modello è stato deciso di utilizzare il 30% dei dati di train come validazione. Le scelte della loss, ottimizzatore e funzioni di attivazione sono state dettate dalla letteratura specifica. Per quanto riguarda l'architettura della rete e il numero di epoche (10) invece si sono effettuate diverse prove

per trovare la dimensione che restituisse il miglior trade of tra risultati e peso computazionale.

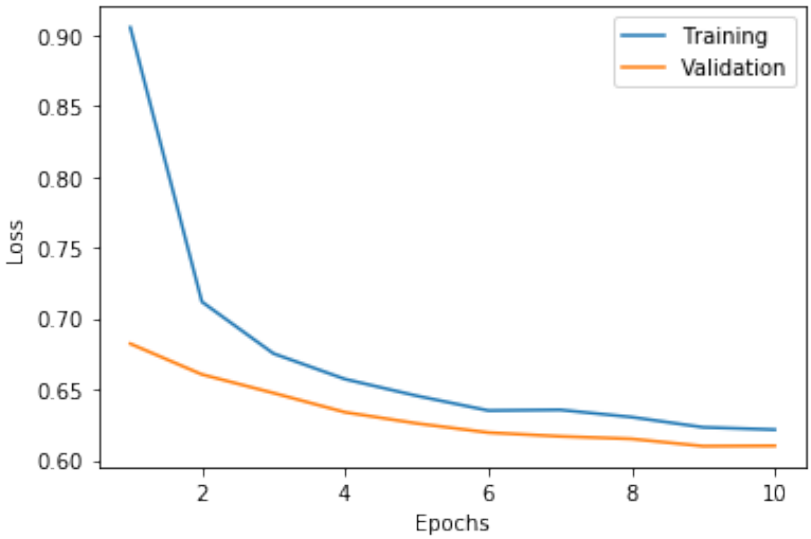

Figura 2: Valore del loss lungo le epoche

Come si vede dal grafico non è presente overfit, potrebbe sembrare strano che il valore della loss nella validation sia sempre più basso, ma ciò è un possibile risultato quando vengono utilizzati layer di dropout.

5 Risultati

La rete è stata testata sul l'ultima porzione del dataset iniziale. Raggiunge un'accuracy di circa 75% e un loss score di circa 0.58 valori paragonabili a quelli del train e validation. La misura più importante però è la $F-measure$ della classe positiva che sul test set raggiunge un valore di circa 54%. In fine si riporta anche la $F-measure$ pesata per le due classi pari a circa 78%. I risultati raggiunti quindi sono soddisfacenti.

Come ultima cosa è stata usata la rete per prevedere il label dei dati contenuti nel file *test.csv*, i risultati sono riportati nel file di testo:
Federico_Manenti_790032_score1.txt