Istruzioni per l'analisi dei data set di Data Analysis

1 Prima parte (per questa parte è obbligatorio utilizzare il linguaggio R)

Il data set da analizzare tramite modelli di regressione consta di $n = 70$ osservazioni di una variabile dipendente Y e di $p = 50$ regressori X_j $(j = 1, 2, \ldots, p)$ potenzialmente utili alla predizione di Y. L'obiettivo dell'analisi è, dopo aver confrontato diverse tecniche di regressione, l'individuazione del modello lineare che minimizza l'errore di predizione su un test set e la stima dei coefficienti delle variabili indipendenti significative per la predizione di Y. Nello specifico, si chiede di:

a. confrontare tra loro le tecniche per la costruzione di modelli empirici lineari presentate al corso, scartando quelle che non è opportuno utilizzare per questo tipo di data set;

b. indicare la strategia che permette di individuare il modello di regressione che minimizza l'errore di test sulla variabile Y;

c. individuare i regressori significativi per la predizione di Y e stimare i loro coefficienti β_j con la strategia determinata al punto **b**.

I valori stimati dei coefficienti β_j selezionati al punto **c** dovranno essere divisi per 100 e il risultato di tale operazione dovrà poi essere arrotondato all'intero più vicino. Gli interi così ottenuti rappresenteranno i codici ASCII decimali di caratteri alfanumerici che, ordinati per valori crescenti di j, formeranno una stringa che rappresenterà un indizio per la soluzione della seconda parte dell'esercizio. A tal fine, potrà essere utile la funzione di R `intToUtf8()`.

Si richiede che l'80% delle osservazioni del data set per la regressione venga utilizzato per il training dei modelli e la scelta dei loro parametri, mentre il test set sia costituito dal restante 20% dei dati forniti.

2 Seconda Parte (per questa parte si possono utilizzare Python o MATLAB, anche entrambi se desiderato)

Si consideri un training set ottenuto da un flusso di dati collezionati in successivi istanti di tempo $n = 1, 2, \ldots, N$. Il training set è costituito dalle coppie feature-label $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$, in cui $X_n \in \mathbb{R}^d$ e $Y_n \in \{-1, 1\}$. Le coppie (X_n, Y_n) sono contenute nel file "train.mat", in cui l'n-esima riga rappresenta la coppia (X_n, Y_n) (l'ultima colonna rappresenta Y_n).

a. Si effettui preliminarmente un'opportuna riduzione della dimensionalità del training set. Si denotino con X'_n e d' le nuove feature e la relativa dimensionalità ridotta.

b. Si addestri un classificatore logistico sul training set a ridotta dimensionalità. A tal fine, si applichi l'algoritmo del gradiente stocastico con step-size costante. Si valuti il comportamento dell'algoritmo per diverse scelte dello step-size. **Non è consentito utilizzare routine già disponibili per implementare l'algoritmo del gradiente stocastico.**

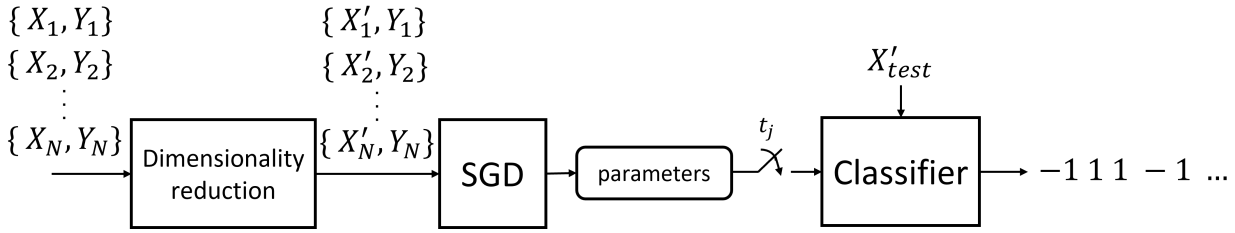
c. Il sistema viene chiamato ad effettuare predizione in specifici istanti di tempo $t_1, t_2, \ldots, t_K$, osservando le feature $X_{\text{test}}(1), X_{\text{test}}(2), \ldots, X_{\text{test}}(K)$. Il file "test.mat" contiene gli istanti di tempo e le relative feature. Nel dettaglio, per ogni riga j, la prima colonna rappresenta l'istante di tempo t_j, le successive

rappresentano la feature $X_{\text{test}}(j)$. Indicato con $\widehat{\beta}(t_j) \in \mathbb{R}^{d'}$ il parametro stimato dall'algoritmo del gradiente stocastico al tempo t_j, e con $X'_{\text{test}}(j)$ la feature a ridotta dimensionalità, il classificatore applica la seguente regola di decisione:

$$\begin{aligned} &+1 \text{ se } X'_{\text{test}}(j)\,\widehat{\beta}(t_j) > 0, \\ &-1 \text{ altimenti.} \end{aligned}$$

d. Si converta in caratteri ASCII codificati con 8 bit la stringa binaria ottenuta classificando le osservazioni, associando il bit 0 al valore -1. I caratteri ottenuti sono tratti da una celebre frase legata all'indizio ricavato dalla prima parte del compito assegnato.

Per completare l'esercizio, è necessario risalire alla frase originale (per quest'ultimo compito non servono metodi studiati al corso, ma un po' di intuito ed elementi di "cultura generale"...).



3 Istruzioni per la consegna

Entro il giorno precedente alla data dell'appello a cui si intende partecipare, è necessario inviare via e-mail ai docenti del corso l'indizio scoperto risolvendo la parte 1 e la frase scoperta risolvendo la parte 2.

Non bisogna inviare nessun altro materiale prima della prova in aula.