Corso di Laurea in Matematica, Università di Roma "La Sapienza"

Corso di Data Mining

M. Falcone

Esercitazioni in Laboratorio

Foglio 4: Riconoscimento dei caratteri

Per il riconoscimento dei caratteri useremo una base di dati piuttosto larga, basata sia su caratteri tipografici che su caratteri scritti a mano. I dati vanno prima letti dai file (`EnglishFnt.tgz, EnglishHnd.tgz, EnglishImg.tgz`). I files sono scaricabili su elearning e i caratteri sono delle immagini standardizzate in bianco e nero di 28×28 pixels. Il procedimento del riconoscimento utilizza un sottinsieme dei caratteri (il training set) per istruire il sistema, successivamente sarà possibile utilizzare le informazioni contenute nel training set per classificare un altro carattere. La informazione contenuta nel training set viene ottenuta tramite la SVD che avete usato nelle precedenti esercitazioni.

Il procedimento iniziale prevede la trasformazione di ogni immagine $\mathcal{I}$, di 28×28 pixel, del training set in un vettore a 784 componenti. Questo vettore viene memorizzato nella colonna di una matrice, quindi se usiamo n immagini, ad esempio dello 0, creeremo una matrice $A(0) = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_n\}$. In modo analogo procediamo per gli altri caratteri, ottenendo le matrici $A(1)$, $A(2)$, $A(3)$,....

A questo punto si calcolano tramite la SVD le fattorizzazioni delle matrici $A(X)$, con $X = 0, 1, 2, \ldots$. In questo modo potremo confrontare la matrice corrispondente al carattere C che ci interessa (e non contenuto nel training set) con tutte le fattorizzazioni che abbiamo ottenuto e potremo allora calcolare l'errore corrispondente (in norma di Frobenius).

Il meccanismo è analogo a quello già utilizzato nel programma sulla compressione delle immagini (Foglio 1). Il carattere C verrà riconosciuto come X se l'errore corrispondente ad $A(X)$ è il più piccolo degli errori. Questa valutazione si può fare sia in termini numerici che in termini grafici.

ALGORITMO DI CLASSIFICAZIONE

1. Leggere un insieme di caratteri estratto dal data base scelto (ad esempio 10 immagini per ognuna delle lettere)
2. Creare le matrici $A(X)$, per $X = 0, 1, 2, 3 \ldots$
3. Calcolare la SVD delle matrici $A(X)$ e costruire un rappresentante per ogni classe di caratteri (ad esempio usando i primi k valori singolari)
4. Scegliere un carattere C dal data base (possibilmente diverso da quelli già usati per il training set) e confrontarlo con i rappresentanti delle varie classi
5. Per il confronto, creare una tabella a k fissato che calcoli il residuo rispetto alle varie classi

6. Rappresentare in un grafico i residui anche al variare del numero dei valori singolari utilizzati (ad esempio cambiando colore al variare di k).

7. Mandare a video le prime due classi individuate come più probabili indicando anche il residuo corrispondente.

Funzioni Matlab(R) utili: `rgb2gray`, `squeeze`, `reshape`, `svds`, `colormap`, `plot`