文字データについて

今回、文字認識の演習で用いる文字は、

「あ」、「か」、「ば」、「ぱ」

の 4 字種であり、これらは文字認識の研究分野で広く用いられている手書き文字データベース “ETL9B” から抜粋したものである。ETL9B は産業技術総合研究所において、電子情報技術産業協会、大学、民間の研究機関等の協力のもとに、パターン認識の 1 分野である文字認識研究用に収集された、手書きひらがな、漢字の約 60 万サンプルの文字パターンデータを納めたデータベースである。

これらを編集して、$16 \times 16 = 256$ 次元の、0,1 の値を持つ 2 値画像を作成した。
学習データ、テストデータのファイル名は、

- `moji_a.dat` – 「あ」の学習データ
- `moji_ka.dat` – 「か」の学習データ
- `moji_ba.dat` – 「ば」の学習データ
- `moji_pa.dat` – 「ぱ」の学習データ
- `moji_test.dat` – テスト (未知) データ

となっており、学習データが各字種 180 文字、テスト (未知データ) が $20 \times 4 = 80$ 個である。

ファイルの形式は、前回までの 2 値データと同じ形式になっている。

学習データファイル　「あ」、「か」、「ば」、「ぱ」の 4 種

データ 1 の 1 次元目	データ 1 の 2 次元目	$\cdots$	データ 1 の 256 次元目
データ 2 の 1 次元目	データ 2 の 2 次元目	$\cdots$	データ 2 の 256 次元目
データ 3 の 1 次元目	データ 3 の 2 次元目	$\cdots$	データ 3 の 256 次元目
$\vdots$	$\vdots$	$\cdots$	$\vdots$
データ 180 の 1 次元目	データ 180 の 2 次元目	$\cdots$	データ 180 の 256 次元目

テストデータファイル

データ 1 の 1 次元目	$\cdots$	データ 1 の 256 次元目	正解クラス番号
データ 2 の 1 次元目	$\cdots$	データ 2 の 256 次元目	正解クラス番号
データ 3 の 1 次元目	$\cdots$	データ 3 の 256 次元目	正解クラス番号
$\vdots$	$\cdots$	$\vdots$	$\vdots$
データ 80 の 1 次元目	$\cdots$	データ 180 の 256 次元目	正解クラス番号

2 次元データ用プログラムからの修正

文字認識の演習では、前回まで各自作成した 2 クラス 2 次元データ用プログラムを修正して用いる。以下に修正点のヒントを示すので参考にしていただきたい。

1. `TDATA,LDATA,DIM` の値を書き換える
2. 2 クラス用から 4 クラス用へ配列を増やす (例：`class3[][],class4[][],mean3[],mean4[]` など)。
3. その他、データ入力部 `input()`、結果出力部 `output()` なども適当に変更する。