Return to Contents PDF Version

Structure of the Datasets

Brian T. Luke (lukeb@ncifcrf.gov)

The datasets distributed here have four sections that contain all of the information, though only the first and fourth sections are required.

The first section simply contains two integers, NFEAT and NSAMP. NFEAT is the number of features in the dataset and NSAMP is the number of samples.

The second (optional) section contains one or more labels for each of the features. In a mass spectroscopic dataset, for example, the label for each peak can be the *m/z* value of the approximate peak center as determined by a peak-picking algorithm. If this is an LC/MS dataset, each feature can have two labels, one for the LC retention time and one for the *m/z* value. Each label occupies a field that is 10 characters and eight labels are stored on each line in the dataset (8A10 in FORTRAN lingo). If there is more than one label for each feature, the first label must be given for all features followed by the second label and so on.

The third (optional) section contains the category or State index for each sample. These indices should be sequential starting with “1”. For example, in a dataset containing samples with and without a particular disease, all samples with the disease can be given an index of “1” and those without the disease must then have an index of “2”. If samples in different categories or States are stored in different datasets, then these indices are not required.

The fourth section contains the intensities of each feature for each individual. All feature intensities for the first sample are given and then for the second sample, and so on. The order of the samples must correspond to the order of the State indices if these indices are given.

In summary, the structure of the datasets is as follows.

1. NFEAT NSAMP (no required format)
2. Feature Labels (optional, 8A10 format)
3. Class Indices (optional, no required format)
4. Feature Intensities for Each Sample (no required format)

(Last updated 5/2/07)