CSE 4309 - Assignments - Description of UCI Datasets

The files in the UCI datasets directory contain training files and test files for three datasets. Both the training file and the test file are text files, containing data in tabular format. Each value is a number, and values are separated by white space. The i-th row and j-th column contain the value for the j-th dimension of the i-th object. The only exception is the LAST column, that stores the class label for each object. **Make sure you do not use data from the last column (i.e., the class labels) as parts of the input vector.**

The datasets are copied from the UCI repository of machine learning datasets. Here are some details on each dataset:

- The `pendigits` dataset. This dataset contains data for pen-based recognition of handwritten digits.
    - 7494 training objects.
    - 3498 test objets.
    - 16 dimensions.
    - 10 classes.

- The `satellite` dataset. The full name of this dataset is Statlog (Landsat Satellite) Data Set, and it contains data for classification of pixels in satellite images.
    - 4435 training objects.
    - 2000 test objets.
    - 36 dimensions.
    - 6 classes.

- The `yeast` dataset. This dataset contains some biological data whose purpose I do not understand myself.
    - 1000 training objects.
    - 484 test objets.
    - 8 dimensions.
    - 10 classes.

For each dataset, a training file and a test file are provided. The name of each file indicates what dataset the file belongs to, and whether the file contains training or test data.

Note that, for the purposes of your assignments, it does not matter at all where the data come from. The methods that you are asked to implement should work on all three datasets, as well as ANY other datasets following the same format.