13.4.1 USPS dataset

The dataset and this description is made available on http://www-stat.stanford.edu/~tibs/ElemStatLearn/data.html.
The dataset refers to numeric data obtained from the scanning of handwritten digits from envelopes by the U.S. Postal Service. The original scanned digits are binary and of different sizes and orientations; the images here have been deslanted and size normalized, resulting in 16 x 16 grayscale images (Le Cun et al., 1990).
There are 7291 training observations and 2007 test observations, distributed as follows:

	0	1	2	3	4	5	6	7	8	9	Total
Train	1194	1005	731	658	652	556	664	645	542	644	7291
Test	359	264	198	166	200	160	170	147	166	177	2007

or as proportions:

	0	1	2	3	4	5	6	7	8	9
Train	0.16	0.14	0.1	0.09	0.09	0.08	0.09	0.09	0.07	0.09
Test	0.18	0.13	0.1	0.08	0.10	0.08	0.08	0.07	0.08	0.09

The test set is notoriously "difficult", and a 2.5% error rate is excellent. This is a notorious example of multiclass classifiction task where $\mathbf{y} \in 0,1,\ldots,9$ and the inputs are real vectors.