USPS Dataset

Note: This dataset (scaled to [-1:1] instead of [0:2]) also appears in the book "The elements of statistical learning" by Hastie, Tibshirani and Friedman (Springer, 2001). It is available from the corresponding website here with the name "ZIP code".

The dataset consists of a training set (usps_train.jf, 1.4M) with 7291 images and a test set (usps_test.jf, 390k) with 2007 images.

The ".jf" format is an ASCII data file format we use because of easy portability (although the files are somewhat large) it contains:

line 1:
[number of classes [integer]] [number of features [integer]]
line 2...I+1:
[classnumber of pattern i [integer in [0;number of classes-1]]]
[features of pattern i [double]]
line I+2:
-1 (this is the end marker)

The features are floating point in [0,2] for "historical" reasons.

If you like, you can use the lines of C code included at the end of this message to read the files into memory (I hope they work).

```
int read_ascii_file(char *name, int* numclass, int* dim, int* num, int * classlabels, double ** content, int max)
{
FILE * file;
int class,i;
char buffer[128];
file = fopen (name,"r");
fscanf(file,"%d %d\n",numclass,dim);
fscanf(file,"%d ",&class);
*num=0;
while (class!=-1)
{ content[*num]=(double*)malloc((*dim)*sizeof(double));
classlabels[*num]=class;
for (i=0;i<*dim;i++)
{
if (fscanf(file, "%s", buffer) == EOF)
{
```

```
printf("%s: unexpected end of file\n","read_ascii_file");
exit(1);
}
content[*num][i]=(double)atof(buffer);
}
(*num)++;
if(*num>max)
{
printf("Only %d vectors are allowed here, \nchange input to use more.\n",max);
exit(1);
}
fscanf(file,"\n %d ",&class);
}
fclose (file);
return 0;
}
```

Daniel Keysers
Last modified: Wed Apr 14 17:15:29 CEST 2004