MIRI Data Mining & Business Intelligence

FIB UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

KMLMM course. ZIP Practical work

2016-2017 course
Prof. Tomàs Aluja

We have normalized handwritten digits, automatically scanned from envelopes by the U.S. Postal Service in 16 x 16 grayscale images (from -1 to 1). Each line consists of the id (0-9) followed by the 256 grayscale values. We dispose of a training set of 7291 digits and a test set of 2007 digits. (files “zip_train.dat” and “zip_test.dat” respectively).

The purpose is to use the training data set to build a classification function using the training data and evaluate its quality in the test data. First we will perform a Multivariate Regression and a Principal Components Regression.

Steps for conducting the practice

1. Read the “zip_train.dat” and “zip_test.dat” files provided. Select a 5% random sample (without replacement) of the train data. Use this sample as your training data, and the complete test data for testing.
2. Define the response matrix (Y) and the predictor matrix (X). Center the predictor matrix.
3. Perform a multivariate regression with the training data. Compute the average R2.
4. Compute the average of the R2 by Leave One Out.
5. Predict the responses in the test data, be aware of the appropriate centering. You can compute the prediction by a direct scalar product without using the predict function. Compute the average R2 in the test data.
6. Assign every test individual to the maximum response and compute the error rate.
7. Perform a PCR (using LOO). Decide how many components you retain for prediction.
8. Repeat steps 5 and 6 for the PCR model.