Please complete all assignments in this notebook. You should submit this notebook, as well as a PDF version (See File > Download as).

Random Projections with 1-NN (6 points, 3+3)

Implement random projections for dimensionality reduction as follows. Randomly generate a $k \times d$ matrix $\mathbf{R}$ by choosing its coefficients

$$r_{i,j} = \begin{cases} +\frac{1}{\sqrt{d}} & \text{with probability} \quad \frac{1}{2} \\ -\frac{1}{\sqrt{d}} & \text{with probability} \quad \frac{1}{2} \end{cases}$$

Let $f : \mathbb{R}^d \to \mathbb{R}^k$ denote the linear mapping function that multiplies a d-dimensional vector with this matrix $f(p) = \mathbf{R}p$. For the following exercises use the same data set as was used for Assignment 1 (MNIST). Use the following values of $k = 45, 90, 400$ in your experiments.
You should not use sklearn.random_projection for this assignment, use numpy instead.

(a)

Evaluate how well the Euclidean distance is preserved by plotting a histogram of the values $\phi(p,q) = \frac{\|f(p)-f(q)\|}{\|p-q\|}$. These values should be concentrated around a certain value for fixed k. What is this value expressed in terms of k and d?

[47]:
```
# This is a temporary read-only OpenML key. Replace with your own key later.
oml.config.apikey = '11e82c8d91c5abece86f424369c71590'
```

[50]:
```
mnist_data = oml.datasets.get_dataset(554) # Download MNIST data
# Get the predictors X and the labels y
X, y = mnist_data.get_data(target=mnist_data.default_target_attribute);
```

[142]:
```
# Randomly sample with probability 1/2
 numpy.random.randint(0,2)
```

```
1
```

(b)

Compare the performance of a 1-NN classifier with and without random projection. Report multi-class confusion matrix, precision and recall for each class with and without projection and for each value of k.

[ ]:

PCA of a handwritten digit (7 points, 3+2+2)

Analyze the first two principal components of the class with label 4 of the MNIST data set (those are images that each depict a handwritten 4). Perfom the steps (a), (b), (c) described below. Note that these steps are similar to the analysis given in the lecture. Include all images and plots in your report. You may use sklearn.decomposition.PCA for this assignment. Do not scale the data.