2IMW30 - Foundations of data mining

Quartile 3, 2015-2016
Assignment 1c (13 Points)

Deadline: Thursday 3 March (noon)

In this assignment you will experiment with random projections and show properties of projections and metric embeddings.

Exercise 1 (3+3 Points)

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 projection 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 1a (MNIST). Use the following values of $k = 50, 100, 500$ in your experiments.

(a) Evaluate how well the Euclidean distance is preserved for the first 20 points of the dataset (i.e., the first 20 instances) by plotting the distortion $\phi(p, q) = \frac{\|f(p)-f(q)\|}{\|p-q\|}$ for all $\binom{20}{2}$ pairs. Which distortion do you expect for different values of k? Is this confirmed by your experiment?

(b) Change your 1-NN implementation of Assignment 1a so that it uses the Euclidean distance instead of the cosine similarity. Run your implementation with and without random projection. Measure the performance of 1-NN as before and compare with and without random projection. Note: you need to rescale your vectors after projection by an appropriate factor (see also part (a) of this exercise).

Write a report to summarize your findings. Include in your report: for (a), plots of the distortion against the pairs of instances for each value of k; for (b) confusion matrix, precision and recall for each class with and without projection and for each value of k. Include any other interesting findings.

Exercise 2 (3 Points)

Let F be a k-dimensional linear subspace of $\mathbb{R}^d$, and let $f : \mathbb{R}^d \to F$ be the projection that maps every point $p \in \mathbb{R}^d$ to its nearest neighbor on F (where distances are measured using the Euclidean distance). Prove that for any $p, q \in \mathbb{R}^d$, it holds that

$$\|f(p) - f(q)\| \leq \|p - q\|.$$

(Hint: A linear mapping that maps each point to its nearest neighbor on F can be simulated by a rotation followed by an orthogonal projection.)

Exercise 3 (1+3 Points)

For a point $p = (p_1, p_2) \in R^2$ and $p \in [1, \infty)$, the ℓ_p-norm is defined as

$$\|p\|_p = \left(|p_1|^p + |p_2|^p\right)^{\frac{1}{p}},$$

while the ℓ_∞-norm is defined as

$$\|p\|_\infty = \max\left(|p_1|, |p_2|\right).$$

(a) Draw the unit circle for different values of $p = 1, 2, 10, \infty$.
(b) Prove that there exists a mapping function $f : \mathbb{R}^2 \to \mathbb{R}^2$ with

$$\|p - q\|_1 = \|f(p) - f(q)\|_\infty,$$

for any $p, q \in \mathbb{R}^2$. (Namely, there exists an isometric embedding of ℓ_1 into ℓ_∞.)
(Hint: Note the similarity of the unit disks under the two norms ℓ_1 and ℓ_∞. How can you exploit this to find an embedding?)