Technische Universiteit
Eindhoven
University of Technology

2IMM20 - Foundations of datamining

Assignment 3

Students:
Joris van der Heijden (0937329)
Bram van der Pol (0780042)

Email addresseses:
j.j.m.v.d.heijden@student.tue.nl
a.f.v.d.pol@student.tue.nl

Supervisors:
Dr.ir. Joaquin Vanschoren

Eindhoven, March 29, 2018

Contents

1 PCA and Isomap

1.1 a

The first two principle components of PCA for all classes are:

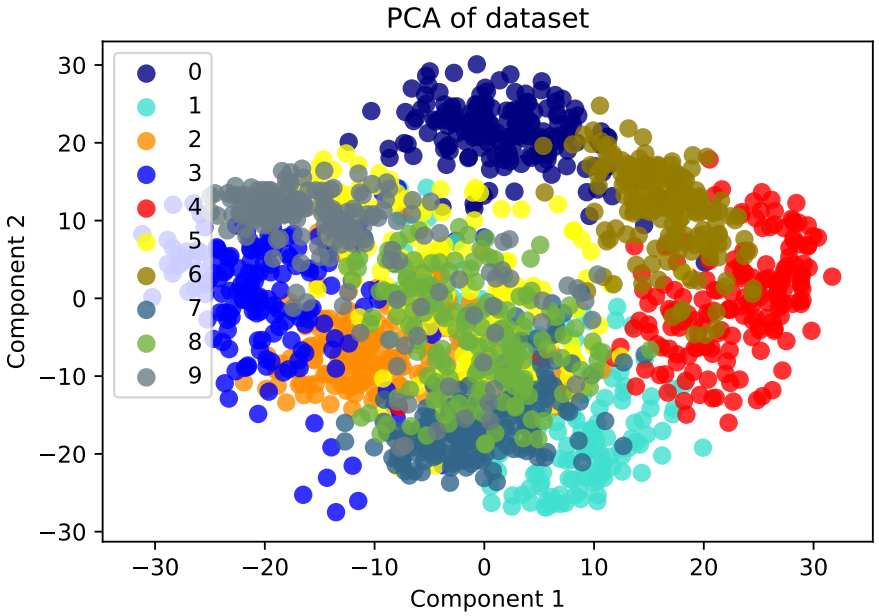

Figure 1: *Scatter plot for the first two principal components*

The classes are not very well separated.

1.2 b

The manifold is constructed using the isomap with the given conditions. The results are:

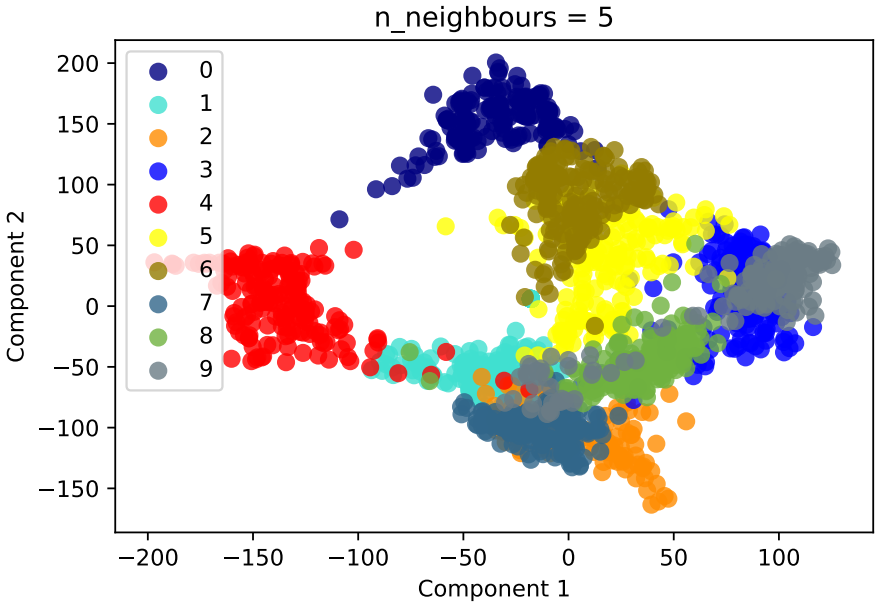

Figure 2: *Scatter plot for aligned data,* $n_{neighbours} = 5$*. Score: 102234.2*

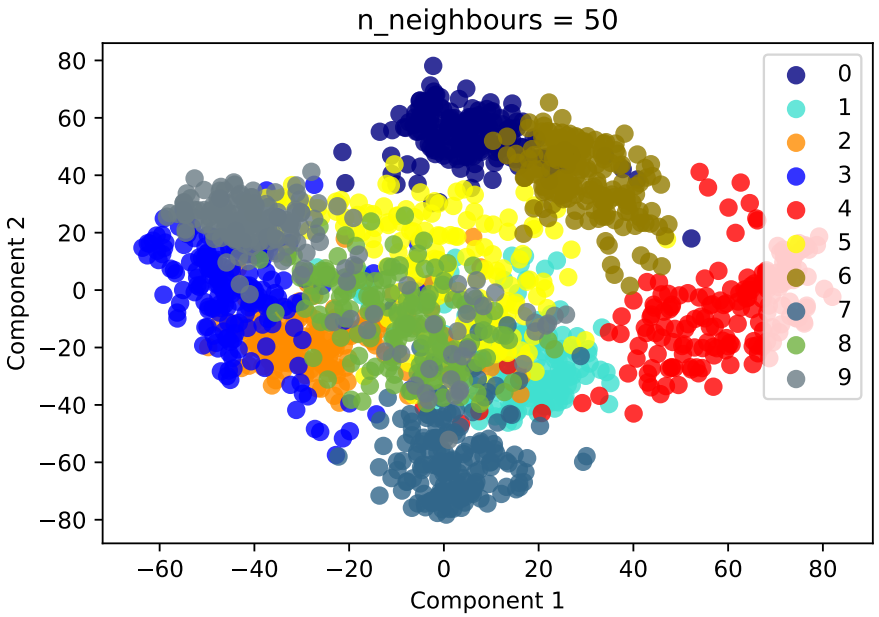

Figure 3: *Scatter plot for aligned data,* $n_{neighbours} = 50$*. Score: 31067.5*

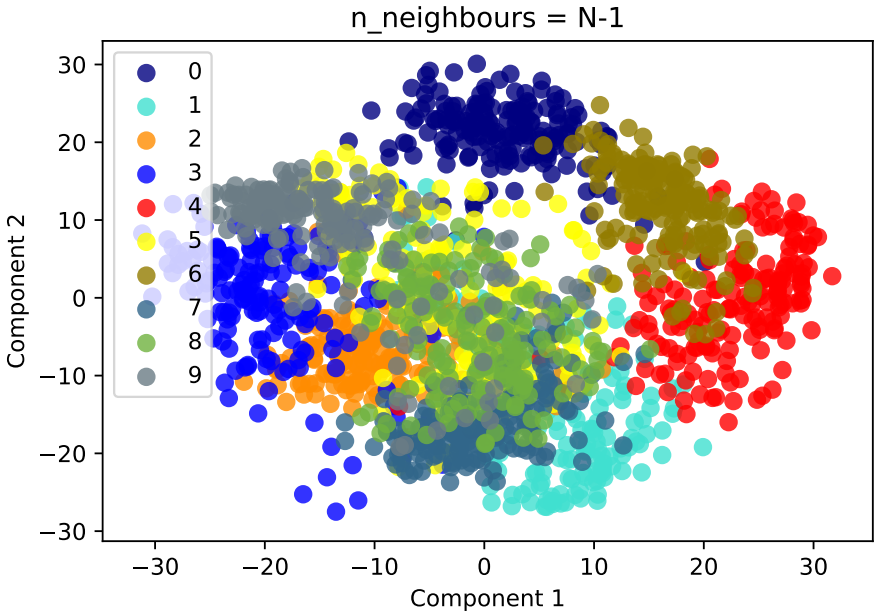

Figure 4: *Scatter plot for aligned data, $n_{neighbours}$ = N-1. Score: 0.0*

1.3 c

The 1-norm score is shown in the caption of the figures of question b. More neighbours decreases the computed score. As the number of neighbours increases, the separation of the different classes becomes more vague. For $n_{neighbours} = 5$ the classes are decently separated, while for $n_{neighbours} = 50$ and $n_{neighbours} = N - 1$ there is increasingly more overlap. If the 1-norm (score) is zero, it means that the of the transformed dataset is identical to the the reference data (PCA) for one of the 8 transformations. This is the case when the isomap embedding does not have any effect on the PCA data, as can been seen by comparing figures 1 and 4. The reason for this is that all the the neighbours are used the isomap cannot make a manifold.

2 Classical multidimensional scaling

2.1 a

$$d_{ij} = ||\mathbf{p_i} - \mathbf{p_j}||^2 = ||\mathbf{p_i} - \mathbf{p_j}||\ ||\mathbf{p_i} - \mathbf{p_j}|| = 2\langle \mathbf{p_i}, \mathbf{p_j} \rangle + ||\mathbf{p_i}||^2 + ||\mathbf{p_j}||^2 \quad (1)$$

2.2 b

The explanation is that because the point set is mean-centered the scaling factor vector ($\mathbf{q}$) does not matter.If statement in question b can be proven in 1 direction all the other dimensions are proven too. The proof can be given mathematically by first expanding the sum sign:

$$\sum_{1\leq i\leq n} \langle \mathbf{p_i}, \mathbf{q} \rangle = p_1 q + \ldots + p_n q = q(p_1 + \ldots + p_n) = 0 \tag{2}$$

The mean-centered definition in respectively all dimensions and one dimension is given by:

$$\sum_{i=1}^{n} \mathbf{p_i} = \mathbf{0} \rightarrow \sum_{1\leq i\leq n} (p_1 + \ldots + p_n) = 0 \tag{3}$$

Therefore equation 2 is zero and equation b in the question holds.

2.3 c

Prove that:

$$||\mathbf{p_i}||^2 = \sum_{l=1}^{n} \frac{d_{il}}{n} - \sum_{k=1}^{n}\sum_{l=1}^{n} \frac{d_{kl}}{2n^2} \tag{4}$$

This equation can be expanded to:

$$||\mathbf{p_i}||^2 = \sum_{l=1}^{n} \frac{||\mathbf{p_i} - \mathbf{p_l}||^2}{n} - \sum_{k=1}^{n}\sum_{l=1}^{n} \frac{||\mathbf{p_k} - \mathbf{p_l}||^2}{2n^2} \tag{5}$$

Which can be written as:

$$||\mathbf{p_i}||^2 = \sum_{l=1}^{n} \frac{2\langle \mathbf{p_i}, \mathbf{p_l} \rangle + ||\mathbf{p_i}||^2 + ||\mathbf{p_l}||^2}{n} - \sum_{k=1}^{n}\sum_{l=1}^{n} \frac{2\langle \mathbf{p_k}, \mathbf{p_l} \rangle + ||\mathbf{p_k}||^2 + ||\mathbf{p_l}||^2}{2n^2} \tag{6}$$

Since equation 2 holds this can be rewritten as:

$$||\mathbf{p_i}||^2 = \sum_{l=1}^{n} \frac{||\mathbf{p_i}||^2 + ||\mathbf{p_l}||^2}{n} - \sum_{k=1}^{n}\sum_{l=1}^{n} \frac{||\mathbf{p_k}||^2 + ||\mathbf{p_l}||^2}{2n^2} = \tag{7}$$

Because $\mathbf{p_i}$ and $\mathbf{p_k}$ are added n times, this can be rewritten to:

$$\frac{1}{n}\left(n||\mathbf{p_i}||^2 + \sum_{l=1}^{n}||\mathbf{p_l}||^2\right) - \frac{1}{2n^2}\sum_{k=1}^{n}\left(n||\mathbf{p_k}||^2 + \sum_{l=1}^{n}||\mathbf{p_l}||^2\right) \tag{8}$$

Vectors $\mathbf{p_l}$ is added n times, therefore:

$$||\mathbf{p_i}||^2+\frac{1}{n}\left(\sum_{l=1}^{n}||\mathbf{p_l}||^2\right)-\frac{1}{2n^2}\left(n\sum_{l=1}^{n}||\mathbf{p_l}||^2+n\sum_{k=1}^{n}||\mathbf{p_k}||^2\right) \tag{9}$$

The sum of k and l for these vectors are the same, therefore:

$$||\mathbf{p_i}||^2+\frac{1}{n}\left(\sum_{l=1}^{n}||\mathbf{p_l}||^2\right)-\frac{2n}{2n^2}\left(\sum_{l=1}^{n}||\mathbf{p_l}||^2\right)=||\mathbf{p_i}||^2 \tag{10}$$

Now equation 4 is shown to be correct.

2.4 d

First the distance is expanded:

$$d_{ij}=||\mathbf{p_i}-\mathbf{p_j}||^2=2\langle\mathbf{p_i},\mathbf{p_j}\rangle+||\mathbf{p_i}||^2+||\mathbf{p_j}||^2 \tag{11}$$

Now equation 4 is substituted in 11 for $||\mathbf{p_i}||^2$ and $||\mathbf{p_j}||^2$:

$$d_{ij}=2\langle\mathbf{p_i},\mathbf{p_j}\rangle+\sum_{l=1}^{n}\frac{d_{il}}{n}-\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{d_{kl}}{n^2}+\sum_{l=1}^{n}\frac{d_{jl}}{n} \tag{12}$$

Therefore:

$$-2\langle\mathbf{p_i},\mathbf{p_j}\rangle=d_{ij}-\sum_{l=1}^{n}\frac{d_{il}}{n}-\sum_{l=1}^{n}\frac{d_{jl}}{n}+\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{d_{kl}}{n^2} \tag{13}$$

3 Locality-sensitive hashing

3.1 a

If the size of the intersection of the sets is zero, than there is no common row with both ones. Therefore the minhashing for each permutation is zero, because every row is different (except with two zeros). Mathematically:

$$\mathrm{sim}_g(S_i,S_j)=\frac{|S_i\cap S_j|}{|S_i\cup S_j|-|S_i\cap S_j|}=Pr[h(S_i)=h(S_j)] \tag{14}$$

Because the intersect is zero the Probability is zero.

$$\mathrm{sim}_g(S_i,S_j)=\frac{0}{|S_i\cup S_j|-|S_i\cap S_j|}=0 \tag{15}$$

Therefore the min hashing is always the correct estimate, namely zero.

3.2 b

In order to maximally amplify the hash family we look at the equation: $1-(1-s^k)^L$. As can been seen in this equation is that in order to maximize the probability of becoming a candidate pair we should maximize L and minimize k, because $L = \frac{m}{k}$. With m the number of hash functions. Therefore the AND constructor (L) must be equal to m and the OR constructor must be equal to 1.

3.3 c

Using the definition of false positives and the function $Y(p)$:

$$Y(p) = \begin{cases} 1 \text{ if } \{\mathbf{p} \in \mathbf{P} : h(\mathbf{p}) = h(\mathbf{q}) \bigwedge d(\mathbf{q}, \mathbf{p}) \leq d_2\} \\ 0 \text{ otherwise} \end{cases} \tag{16}$$

The expected number of false positive is given by:

$$E(x) = \sum_{\mathbf{p} \in \mathbf{P}} Y(p) \tag{17}$$