DATA.STAT.770

Exercise set 8

E6.1

Vector quantization update rule:

$$v(x(t)) = v(x(t)) + \alpha(t)(x(t) - v(x(t)))$$

SOM update rule:

$$u(t+1) = u(t) + h_{u,v}(t)\alpha(t)(x(t) - u(t))$$

where

$$h_{u,v}(t) = exp(\frac{-d_{grid}(u,v)}{\sigma(t)}), \sigma(t) = \sigma_0 exp(\frac{-t}{\lambda})$$

As iterations keep going

$$\lim_{t\to\infty} \sigma(t) = \sigma_0 \lim_{x\to\infty} exp(\frac{-t}{\lambda}) = 0$$

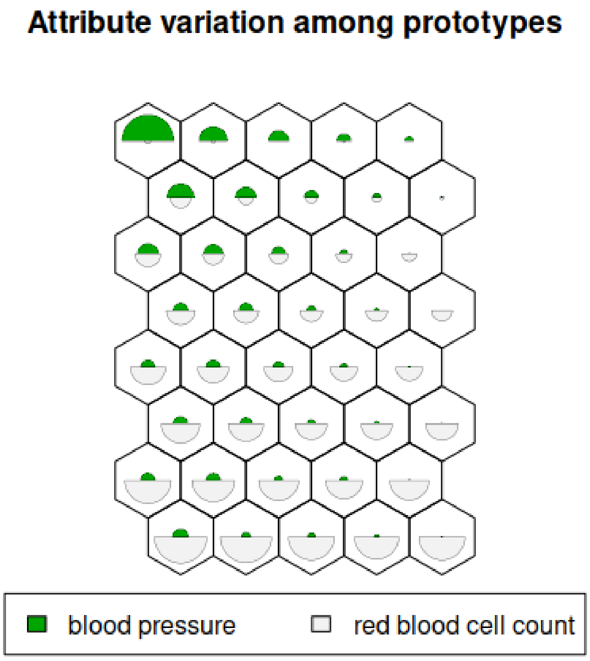
$$\lim_{t\to\infty} h_{u,v}(t) = \lim_{t\to\infty} exp(\frac{-d_{grid}(u,v)}{\sigma(t)}) = exp(-d_{grid}(u,v) \lim_{t\to\infty} \frac{1}{\sigma(t)}) = exp(0) = 1$$

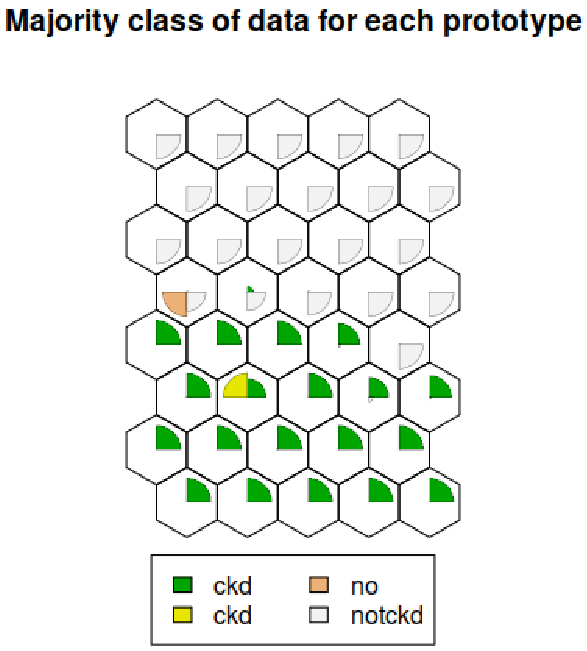
Therefore

$$\lim_{t\to\infty} u(t+1) = \lim_{t\to\infty} u(t) + h_{u,v}(t)\alpha(t)(x(t) - u(t)) = \lim_{t\to\infty} u(t) + \alpha(t)(x(t) - u(t))$$

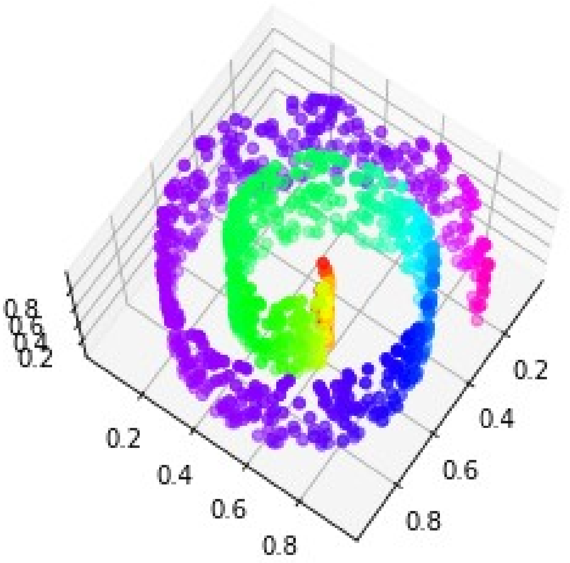
which is (conceptually) same as in vector quantization.

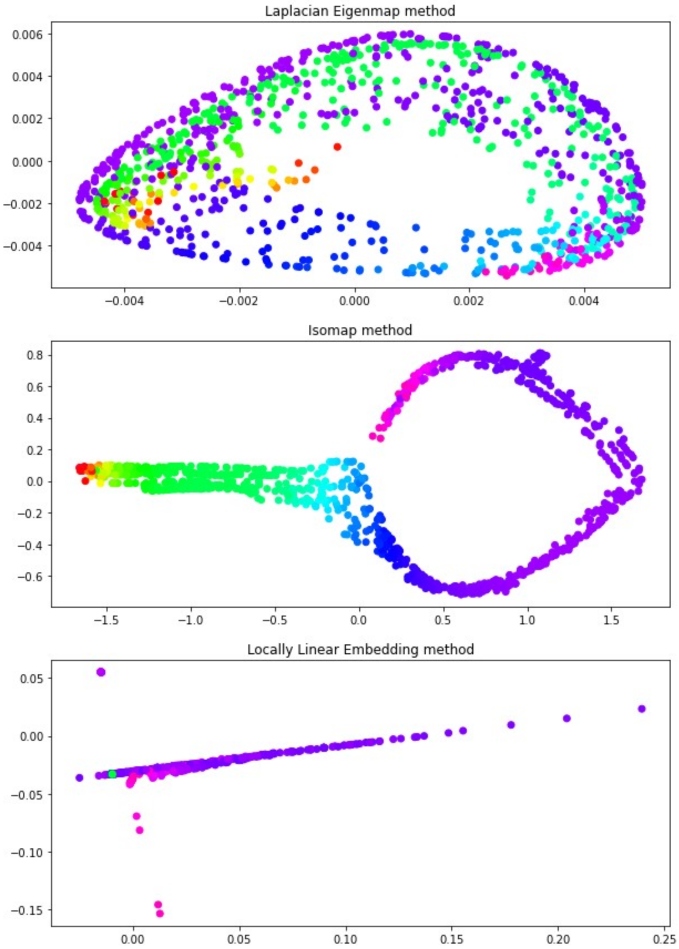
E6.2


Attribute variation among prototypes


Majority class of data for each prototype

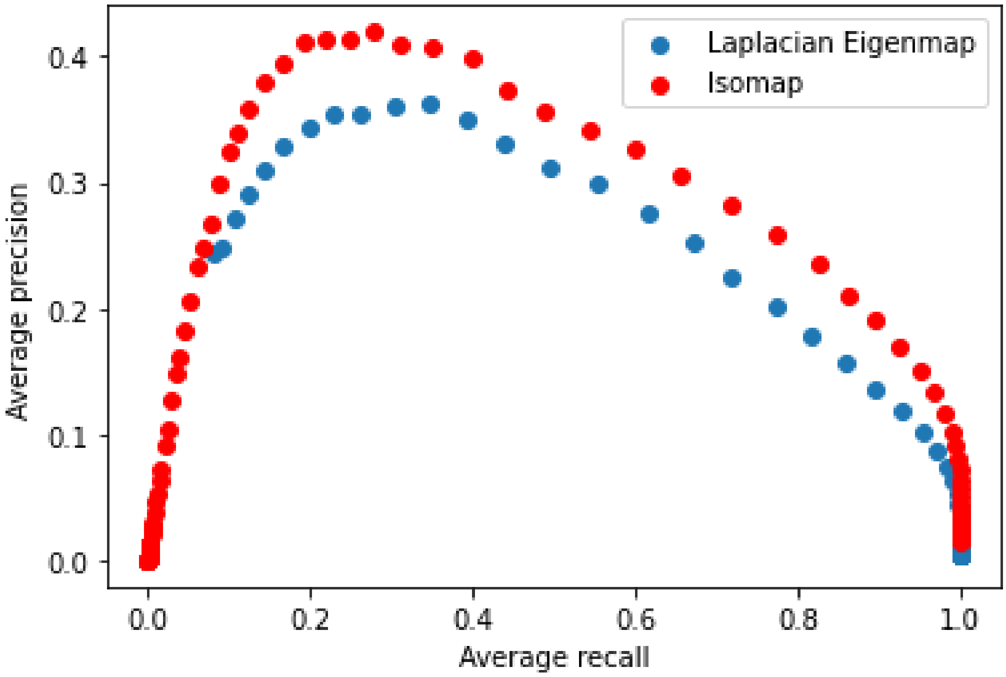
E7.1





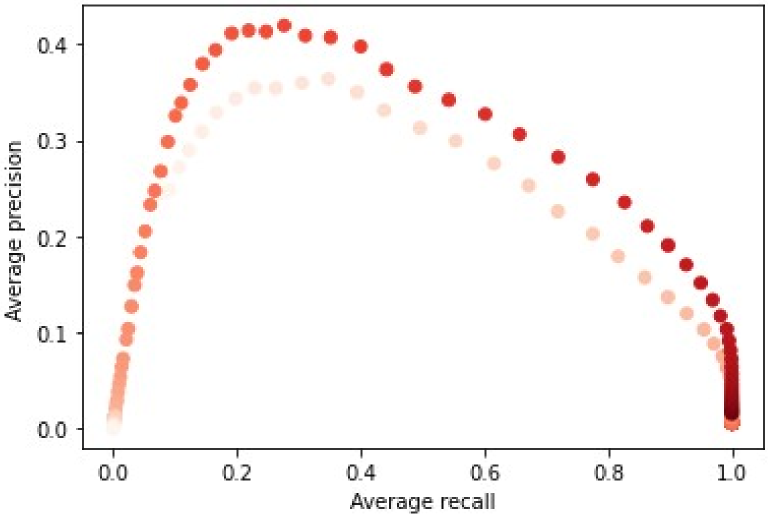
Obtained projections differs remarkably from each other subject to which dimension reduction method was used. It is easy to see that Isomap projection performs pretty well by preserving sample neighbourhoods from original space. In the case of Laplacian Eigenmap, sample neighbourhoods are preserved also quite well, although we can see that the original topology is partially lost. Depending on the context, this might or might not be unwanted property. It is hard to say anything specific about the last method, Locally Linear Embedding, since the projection seemed to lump majority of the samples in the same neighbourhood. Overall, I think that the Isomap method performed best in this projection task.

E7.2



In the plot above, each point corresponds to the average precision and recall values corresponding to single T-value. I constructed the threshold values by initializing it as 1 and on every iteration I multiplied it by 0.9, keeping the lowest possible value fixed at 0.0001. For both methods, I obtained notably better average recall values compared to precision, which is obvious if the threshold value is large.

If we plot the above figure simultaneously with T-values, we obtain the following plot



Here darker color means larger T-value (larger neighbourhood). Now we can also see that with LE-method and small T-values, we obtain better results in terms of both precision and recall. However, with mid-range T-values, the precision is really bad compared to the IM-method. Vice versa, the recall seems to be remarkably worse with IM-method in same situation.