Exam answers. All questions 2-5 are presented here

Problem 2

a)

Forward selection is a wrapper based hill-climbing algorithm to perform feature selection for the dataset. The forward selection algorithm begins with an empty feature set, and on each iteration, the algorithm supplements the feature set by the one feature that increased the preformance of the model we are using (e.g. classification accuracy, regularized fit in regression, classification error rate). The algorithm stops when there is no feature left that at current state would increase the performance. Because of the greedy nature of this algorithm, forward selection does not usually found the global optimal solution.

Steps of the algorithm:

1. Initialize the feature set as empty

2. Find the feature that increases the performance most. If there is no such feature, end the algorithm

3. Add the found feature into feature set

4. Repeat steps 2 and 3 until no features left.

b)

Following the steps of above algorithm (a)

Initial state $E(\o) = 0.777$

Best performance increase: feature 3, $E(3) = 0.743$, feature set: $\{3\}$

Best performance increase: feature 3, $E(3) = 0.743$, feature set: $\{3\}$

Best performance increase: feature 5, $E(3,5) = 0.675$, feature set: $\{3,5\}$

Best performance increase: feature 4, $E(3,5,4) = 0.663$, feature set: $\{3,5,4\}$

No such feature left that decreases the error rate. Algorithm finishes with feature set $\{3,5,4\}$ and error rate of $E(3,5,4) = 0.663$

c)

In this case forward selection found the optimal feature set.

Problem 3:

a)

Cost function of absolute Multidimensional scaling to be minimized

$$\sigma_r = \sum_{i<j} (p_{ij} - d_{ij}(X))^2$$

On above, p_{ij} denotes the original distances, and $d_{ij}(X)$ the reduced distances representation X. Therefore Multidimensional scaling tries to preserve original distances also in projected space. Preserving all distances is however overly optimistic, and Multidimensional scaling is better at preserving the long original distances. This is because the distance errors between far-away samples contributes to the cost function more compared to the distance errors between close-by samples.

b)

$$\sigma_r = (AB_O - AB_R)^2 + (AC_O - AC_R)^2 + ... + (DE_O - DE_R)^2,$$

where subindex O denotes the distances in original space and R in reduced space.

Numerical value:

$$\sigma_r = (4-2)^2 + (6-5)^2 + (2-3)^2 + (5-7)^2 + (3-3)^2 + (4-1)^2 + (5-5)^2$$
$$+ (5-2)^2 + (4-2)^2 + (3-4)^2 = 33$$

c)

In the context of dimensionality reduction, precision and recall are measures of trustworthiness and distance/neighbourhood preservation in the reduced space. The initial objective is to obtain good measures for both precision and recall, but usually it is rare to get good values for both.

Both measures have a range from 0 to 1 where lower value denotes poorer performance. Good precision means that the dimensionality reduction method is trustworthy, meaning that points that are close in the reduced space are also close in the original space. However, precision cannot detect if originally close samples have become distant in the reduced space. This might cause missing neighbours in the sample neighbourhoods in reduced space.

Good recall measures the projections capability to preserve original sample distances/neighbourhoods. This measure has also its drawbacks since it can’t detect if some originally distant samples have become close by in the reduced space. This might cause false neighbours in one sample’s neighbourhood.

Mathematical definitions (in terms of neighbourhoods):

Precision

$$\frac{|N_i \cap R_i|}{|R_i|}$$

Recall

$$\frac{|N_i \cap R_i|}{|N_i|}$$

At above, N_i denotes the i:th point's true sample neighbourhood, and R_i denotes the retrieved neighbourhood after projection to low-dimensional space.

d)

$$N_A = (B, D), R_A = (B, D)$$
$$N_B = (A, C, D), R_B = (D, A)$$
$$N_C = (B.E), R_C = (E, D)$$
$$N_D = (A, E, B), R_D = (C, B)$$
$$N_E = (C, D), R_E = (C, D)$$

Average precision: Naapuruston määritelmä puuttuu??

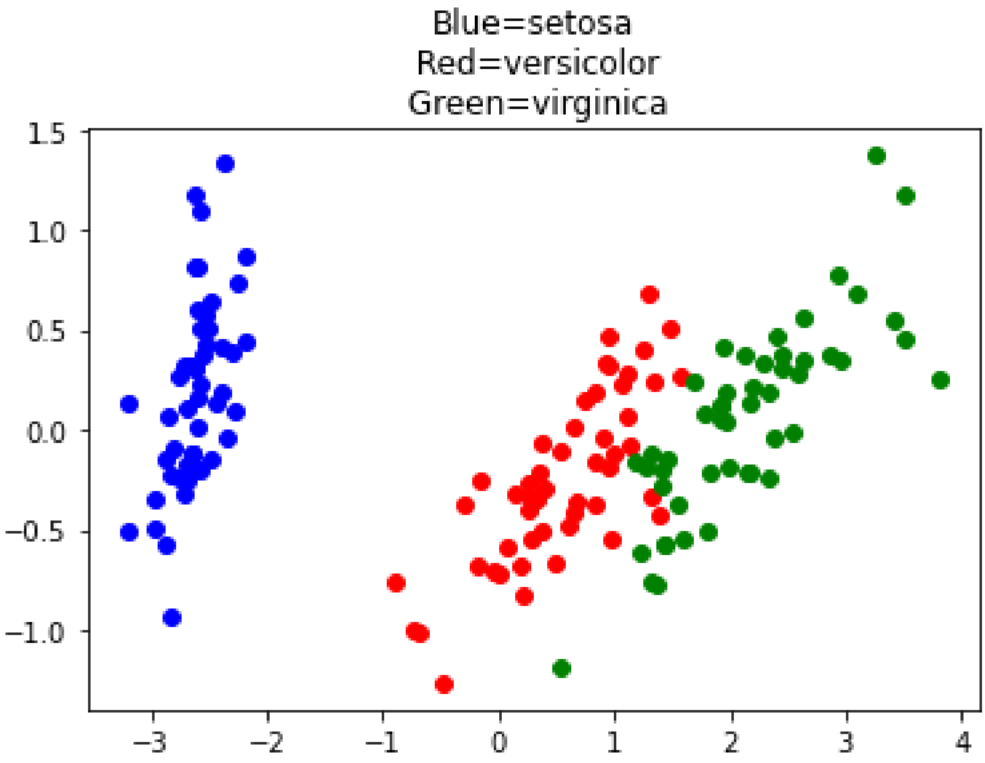
$$\frac{1}{5}\sum\frac{|N_i \cap R_i|}{|R_i|} = \frac{1}{5}(\frac{2}{2} + \frac{2}{2} + \frac{1}{2} + \frac{1}{2} + \frac{2}{2}) = 0.8$$

Average recall:

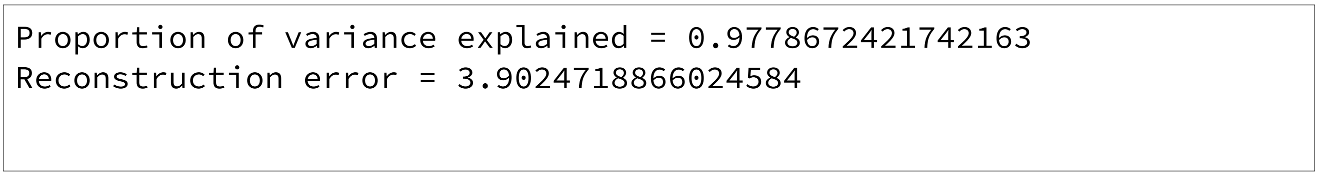
$$\frac{1}{5}\sum\frac{|N_i \cap R_i|}{|N_i|} = \frac{1}{5}(\frac{2}{2} + \frac{2}{3} + \frac{1}{2} + \frac{1}{3} + \frac{2}{2}) = 0.7$$

Problem 4.

a)

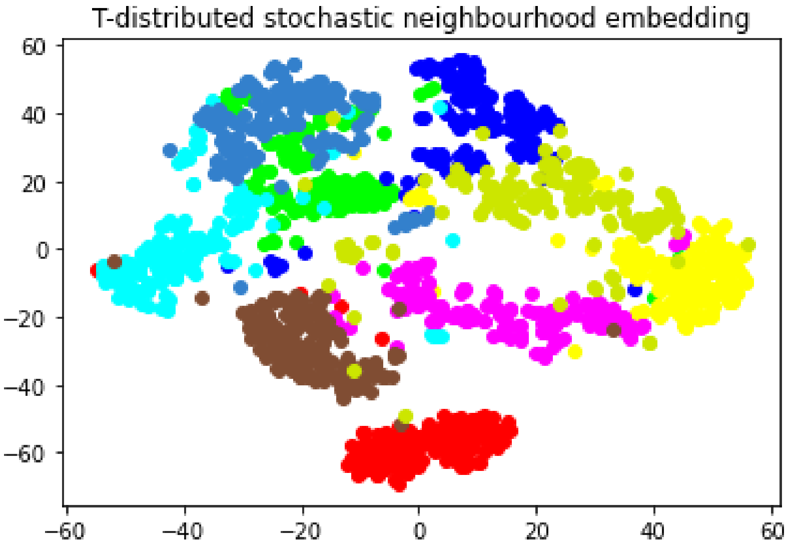


b)



```
Proportion of variance explained = 0.9778672421742163
Reconstruction error = 3.9024718866024584
```

c) I used t-distributed Stochastic Neighbourhood Embedding (t-SNE)



Problem 5.

a)

In the case of analysing the overall population of patients, there are many things that may be common between them. As a starting point, I would focus on analysing how how patients form a groups based on their background formalized as the demographic information. The question of interest, or hypothesis, would be that are similar patients treated in the same hospital. As a main method I would use Self-organizing maps with 10 prototypes to denote the 10 hospitals used in analysis. The self-organizing map tries to fit a graph surface onto the observed data samples and, where graph is represented as a grid of prototypes connected on each other. SOM is an iterative algorithm, and on each iteration we select a prototype and move it towards the closest sample present in the dataset. After running the algorithm, the assumption is that samples near to the specific prototype would be similar to each other. In the case of our research question, we would assume that given a prototype (representing individual hospital), all its neighbours (patients) would be treated in the hospital that the prototype represents. Based on the neighbours, we could for example plot the discrete distribution of the hospital information that was assigned for them. If this procedure produces a prototype-neighbourhood pairs that seem somewhat unintuitive (e.g. the neighbourhood consist of samples that are not treated in the hospital presented by prototype) we could further analyse that what makes these samples similar in the sense of hospital information. Alternatively we could create a report for the hospital itself about the patient information.