UiT

THE ARCTIC
UNIVERSITY
OF NORWAY

FYS-2021 - Exercise Set 7

Dimensionality Reduction and Clustering
Department of Physics and Technology
Faculty of Science and Technology

From book:

6.1

Clustering

Problem 1

(1a) Implement your own version of the k-means algorithm

(1b) Test your implementation on the datasets provided in `blobs.csv` and `flame.csv`. Plot and comment on the results.

(1c) Test your implementation on the optdigits dataset provided in `optdigits.csv`. Plot the centroids of each cluster and use these to determine which digits they represent. Plot some of the wrongly assigned digits, and explain why they were misclassified.

Dimensionality reduction

Problem 2

(2a) Implement your own version of the multidimensional scaling algorithm.

(2b) The file `city-distances.csv` provides pairwise geodesic distances between 34 Norwegian cities (The names are available in `city-names.csv`). Use your MDS implementation to embed the cities in two dimensions. Plot the result.

If you want to annotate the plot with the city names, you can do the following:

```python
fig, ax = plt.subplots()
ax.scatter(y[:,0], y[:,1]) # y is the MDS output
for i, name in enumerate(names):
    ax.annotate(name, y[i])
plt.show()
```

Problem 3

(3a) In linear regression, $E_{RSE} = 1 - R^2$ can be used to measure the model error ($1-$ goodness of fit). If E_{RSE} is computed using the training set, it will decrease with the number of predictors. Explain why this happens. Why does this behavior make E_{RSE} unsuitable as an error measure in the Subset-Selection algorithm?

Problem 4

(4a) Implement your own version of the subset selection algorithm. Here you can choose whether to use forward- or backward selection.

A truncated version of the *Optdigits* dataset is provided in `optdigits-012.csv`. In this file, each row represents a flattened 8 × 8 image of a hand-written 0, 1 or 2. The first element is the label. The image can be recovered by doing `row.reshape((8,8))`.

(4b) Test your subset selection algorithm on the Optdigits-data, using the logistic discriminator from Chapter 10. Use classification error ($1-$Accuracy) as the error-measure. Remember to split the data set into training- and validation-sets, and use the latter to compute the classification error.

(4c) Plot a binary 8 × 8 image, where a pixel has value 1 if it was included by the subset-selection procedure. Comment on the location of these pixels, with respect to the input-data characteristics.