UiT

THE ARCTIC UNIVERSITY OF NORWAY

5. Dimensionality Reduction

FYS-2021 Exercises

Department of Physics and Technology
Faculty of Science and Technology

Problem 1

The file `city-distances-sweden.csv` provides pariwise, geodesic distances between 34 Swedish cities, and the corresponding names are in the file `city-names-sweden.csv`. Based on these distances, we will in the following exercises estimate the coordinates of these cities relative to each other.

(1a) Describe the main difference between feature extraction and feature selection. Describe the multidimensional scaling (MDS) algorithm and comment on its areas of use.

(1b) Explain how you will use the provided distances to obtain the $\mathbf{B} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ matrix.

(1c) Implement the multidimensional scaling algorithm, and perform MDS on the provided data. Briefly discuss what would be a sensible number of dimensions of your output.

(1d) Plot the result of your MDS scaling. Comment on your result by visually comparing it to a map (e.g. Google maps[1]). Try to explain similarities and differences.

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

```
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()
```

[1] https://goo.gl/maps/hCoNRL1WQU82

Problem 2

In Exercise Set 2, we performed multivariate linear regression to predict a car's fuel consumption (miles per gallon) based on six predictors: (i) cylinders, (ii) displacement, (iii) horsepower, (iv) weight, (v) acceleration and (vi) model year. The dataset is named `auto-mpg.csv`, and can be found in Canvas (remember that the response is stored as the first column in the data matrix).

In this problem, we will explore *forward selection* to select a subset of the predictors (features) needed to predict the fuel consumption.

(2a) Split the dataset into a 80/20 training/test split. Train a multivariate linear regression model (on all features) and compute R^2 on the test set to evaluate the quality of the model.

(2b) Implement your own version of the forward selection algorithm. Partition the training data from (a) into a training set and a validation set and use the validation R^2 to evaluate the model's performance on each feature subset. Report the selected feature subset.

(2c) Use the training set from (a) to train a multivariate linear regression model on the selected set of features and compute R^2 on the test set to evaluate the quality of the model. Compare to the result in (a) and comment.

(2d) [Optional] The selected subset may depend on the training/validation split. Running the algorithm with different splits might thus result in different feature subsets. To achieve more robust results, implement k-fold cross validation to evaluate model performance on each feature subset. That means, split the training data from (a) into k folds and evaluate the model on the different splits. The R^2 score of each feature subset is averaged over the splits.

Problem 3

(3a) [Optional] 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 training E_{RSE} unsuitable as an error measure in the Subset-Selection algorithm?