University of Southern Denmark
IMADA

DM566: Data Mining and Machine Learning

Spring term 2022

Exercise 15

Exercise 15-1 k-Fold Cross Validation

Take an SVM with squared-exponential kernel and $C = 1$.
Take the breast cancer diagnostics data set, which can be loaded in to python using

```
X, y = datasets.load_breast_cancer(return_X_y=True)
```

Perform k-fold cross validation with $k = 3, 4, 5$.
Perform 10 random train-test splitting with test set ratio 20%.

Make a box plot of mean and standard deviation of test error across repetitions for all four situations above. Let the x-axis contain the following labels {3-Fold CV, 4-Fold CV, 5-Fold CV, 10 random}, and let the y-axis be the values of the means and standard deviations. Comment on the observed outcome. Is random splitting more or less conservative compared to k-fold cross validation? How is the test error affected by increasing k?