**Assignment 3**

Machine Learning, SS23

|  |  |  |
| --- | --- | --- |
| **Team members** | | |
| Last name | First name | Matriculation Number |
| Heng | Chloe Yi Ning | 12239126 |
| Lim | Yan Qi | 12239271 |

## K-Nearest Neighbors

## Varying value of k

Plot the training and validation accuracy for varying values of kA picture containing text, line, diagram, plot

Description automatically generatedA picture containing text, line, diagram, plot

Description automatically generatedA picture containing line, diagram, text, plot

Description automatically generated for each data set

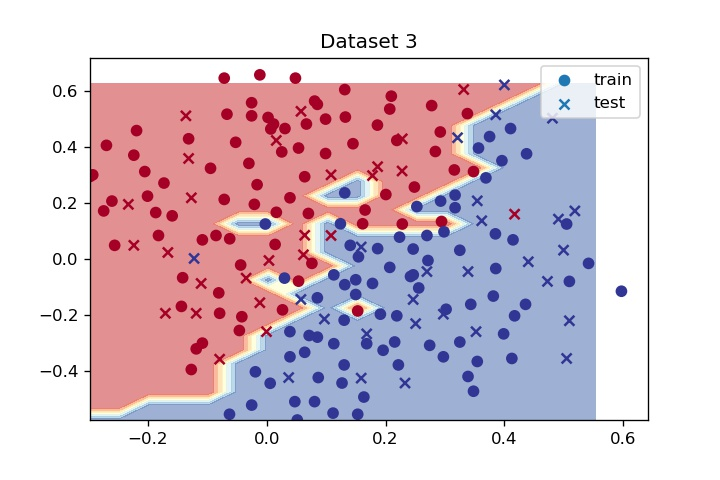
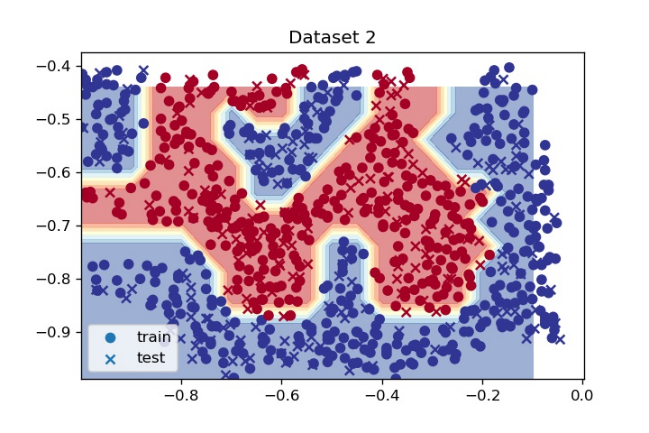
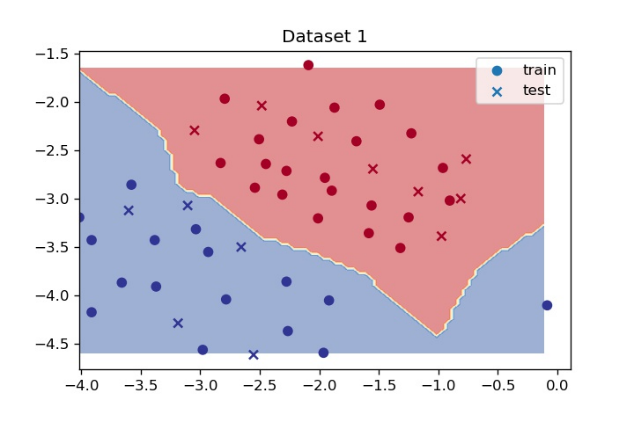
Best value for k using 5-fold cross-validation using a grid search for different values of k ∈ [1, 100]

* k = 1

Performance of the classifier

* Dataset 1: 1.0
* Dataset 2: 0.99537
* Dataset 3: 0.86792

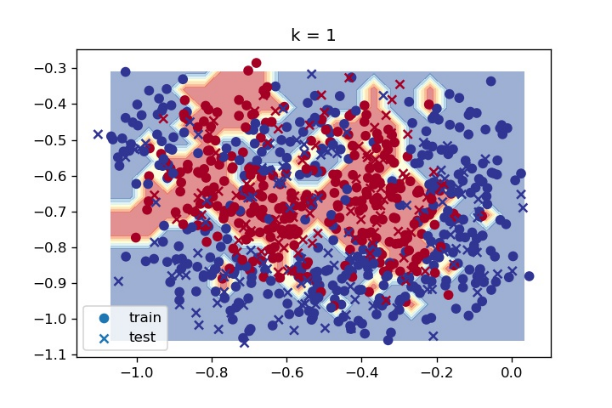
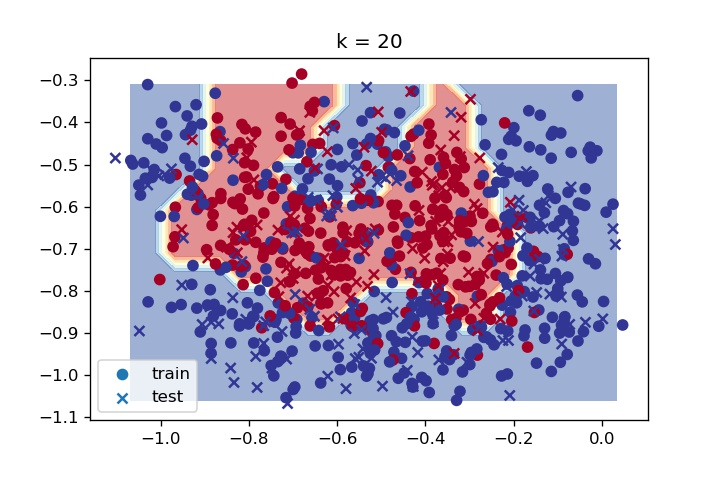
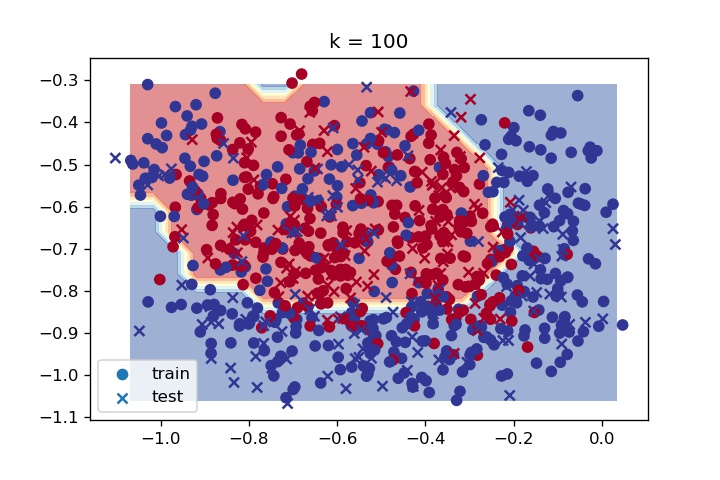
Decision boundaries for k = 1 for each data set



### Effect of low or high values of k

Effect of low or high values of k, k = 1

* With low values of k, the model will not be able to generalize well. As we are overfitting the model, the result has high train set accuracy but rather low test set accuracy.
* With high values of k, the model becomes too generalized instead. As we are underfitting the model, resulting in both low train set accuracy and test set accuracy.
* When k = 1, this means that algorithm will only pick the value closest to the data sample. This will result in a very complex decision boundary that does not generalize well, hence resulting in overfitting as well.

Plot the decision boundaries for k ∈ {1, 20, 100} for the noisy variant of dataset 2.

Best value for k using 5-fold cross-validation using a grid search for k = [1,100]

* k = 5

Plot the training and validation accuracy for varying values of k for noisy variant of data set 2

A picture containing line, plot, diagram, slope

Description automatically generated

Performance of the classifier with chosen value of k on the test set

* Test Score for k = 5: 0.80093

### Support Vector Machines

### Derivation of gradients for w and b

### Decision boundary after training classifier until convergence

parameters: {'C': 0.5, 'eta': 0.001, 'max\_iter': 2000}

A picture containing text, line, diagram, screenshot

Description automatically generatedA picture containing screenshot, diagram, text, colorfulness

Description automatically generated

How do you know that the classifier converged on a solution, i.e., the loss will not improve anymore? (Try to describe your findings in terms of equation 1)

* From the convergence graph, if it reaches plateau. Thus, in terms of equation 1, we should differentiate the equation and equate it to 0 while the rate of change is negative.

Does the decision boundary change when you vary C? Do you have to adapt the learning rate? (Try to describe your findings in terms of equation 1)

* The decision boundary changes when C is varied which leads to the learning rate being adapted. If C increases, values of w need to decrease. Thus, to minimise the cost function, the learning rate has to increase so that values of w can decrease at a faster rate within the same number of iterations.

### Results for grid search over C (and γ) for all datasets for linear and Gaussian kernel

A picture containing text, screenshot, diagram, plot

Description automatically generated

Mean Cross-Validated test Score: 1.0

Parameters:

{'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'}

A picture containing screenshot, colorfulness, text

Description automatically generated

Mean Cross-Validated test Score: 0.8148

Parameters:

{'C': 1.0, 'gamma': 'scale', 'kernel': 'rbf'}

A picture containing screenshot, diagram, colorfulness

Description automatically generated

Mean Cross-Validated test Score: 0.8679

Parameters:

{'C': 1.0, 'gamma': 'scale', 'kernel': 'rbf'}

### Decision Trees & Ensemble Methods

### RandomForestClassifiers (max\_depth & n\_estimators)

Discuss the effects of varying the number of trees and the maximum tree depths on the decision boundary and the performance

* The greater the number of trees, a more complex decision boundary will be produced and the performance on the test set would tend to increase but only until the point before the model starts to overfit.
* Likewise, the deeper the tree, the more complex the decision boundary becomes as well and the performance on the test set would also tend to increase up to the point before overfitting occurs, afterwards when overfitting has occurred, the deeper the tree would result in declining performance.

Report the mean cross-validated accuracy and accuracy on the test set for the best parameters.

* n\_estimators = 1

Dataset 1: Mean cross-validated accuracy: 0.9500, Test accuracy: 0.8462

Dataset 2: Mean cross-validated accuracy: 0.9304, Test accuracy: 0.9167

Dataset 3: Mean cross-validated accuracy: 0.8730, Test accuracy: 0.8113

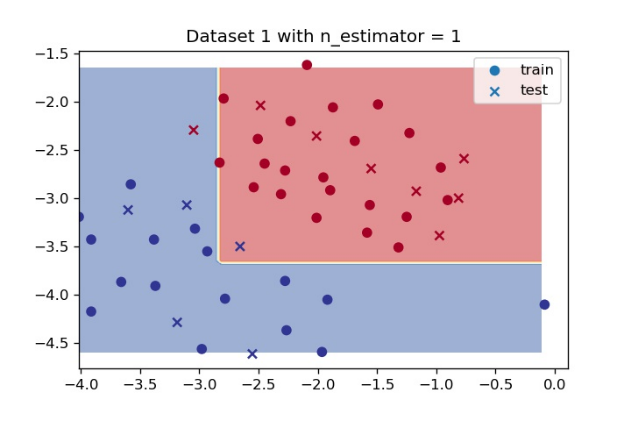
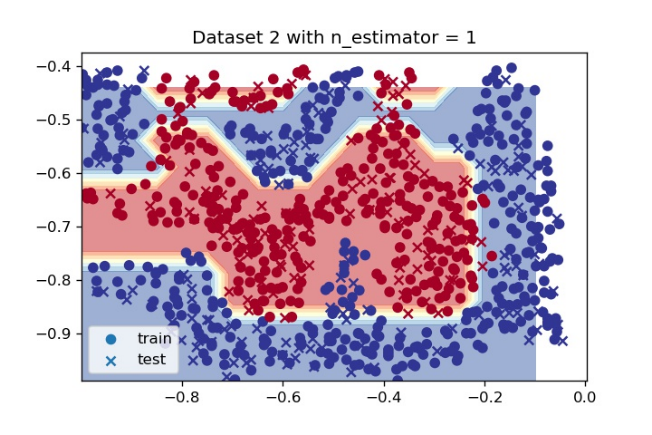
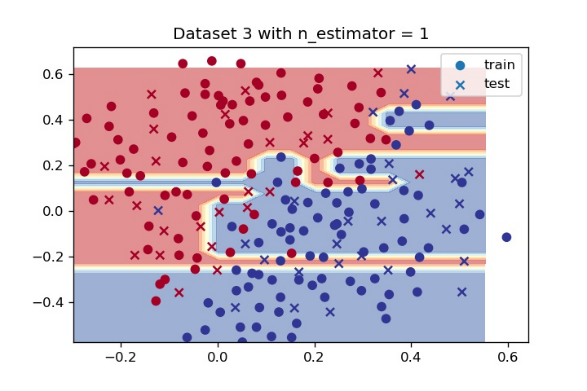
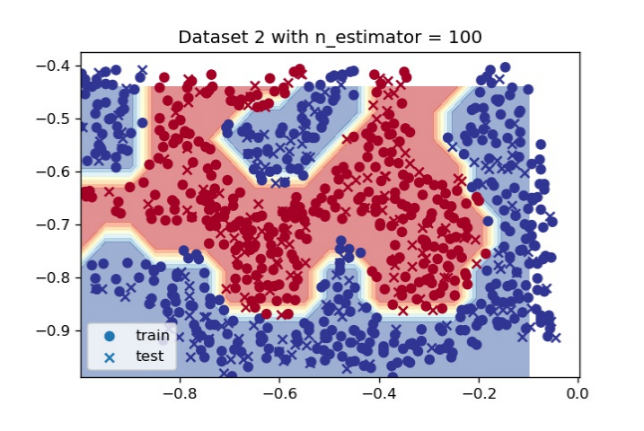
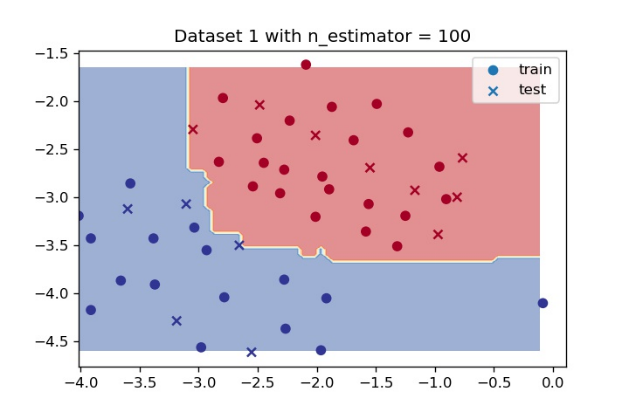
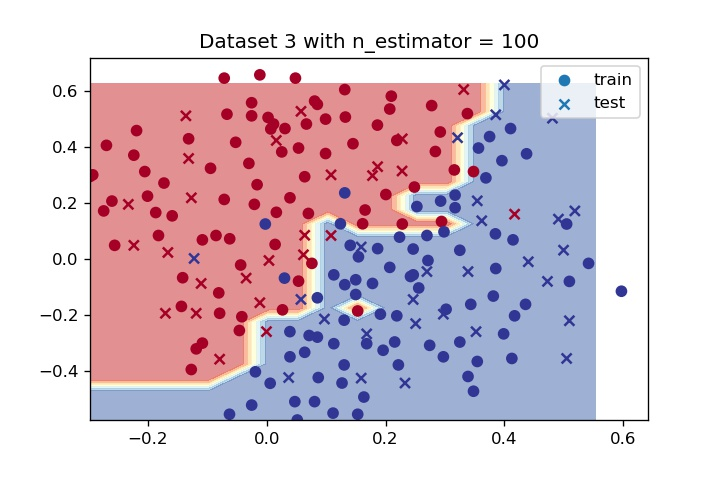
* n\_estimators = 100

Dataset 1: Mean cross-validated accuracy: 1.0, Test accuracy: 0.9231

Dataset 2: Mean cross-validated accuracy: 0.9722, Test accuracy: 0.9861

Dataset 3: Mean cross-validated accuracy: 0.8980, Test accuracy: 0.8868

Plot the decision boundaries for each dataset

* n\_estimator = 1
* n\_estimator = 100

Will the random forest be more or less affected by outliers or noise in the data as you increase the number of trees in the forest?

* The random forest would be less affected by outliers or noise in data as number of trees increase in the forest. As with greater number of trees, the effect of outliers or noise would diminish since there would be a greater number of trees aggregated all together to create the final performance/prediction of the random forest.
  1. **Feature Importance and RFE**

Performance of RandomForestClassifier

* Best parameters for Random Forest: {'max\_depth': 20}
* Performance of Random Forest Classifier on test set: 0.680

Relative Feature Importance Graph

A picture containing text, screenshot, plot, line

Description automatically generated

Performance of SVC classifier

* Best parameters for SVC: {'C': 1.0, 'gamma': 'scale'}
* Performance of SVC classifier on test set: 0.708

Performance of SVC classifier after RFE Transformation on dataset

* Mean Cross Validated Accuracy of best parameters after transformation: 0.748
* Accuracy on test set with best parameters after transformation: 0.748