Linnéuniversitetet
Kalmar Växjö

Report

Assignment 1

Author: Qingqing Dai
Supervisor: Jonas Lundberg
Supervisor: Jonas Nordqvist
Semester: VT 2022
Discipline: Software technology
Course code: 2DV516

Contents

1 Exercise 1 k-NN Classification **1**

1.1 Part 1 Original Microchip Data 1

1.2 Part 2 Prediction of 3 Unknown Microchips 1

1.3 Part 3 The Decision Boundary and the Training Errors 2

2 Exercise 2 k-NN Regression **3**

2.1 Part 1 Divide the dataset into a training set and a test set 3

2.2 Part 2 the Training and Test Set in a 1 x 2 plot 3

2.3 Part 3 the k-NN Regression Result and the MSE Training Error 4

2.4 Part 4 the k-NN Regression Result and the MSE Test Error 5

2.5 Part 5 The k That Gives the Best Regression 6

3 Exercise 4 k-NN Classification using scikit-learn **7**

3.1 Part 1 Original Microchip Data 7

3.2 Part 2 Prediction of 3 Unknown Microchips 7

3.3 Part 3 The Decision Boundary and the Training Errors 8

1 Exercise 1 k-NN Classification

1.1 Part 1 Original Microchip Data

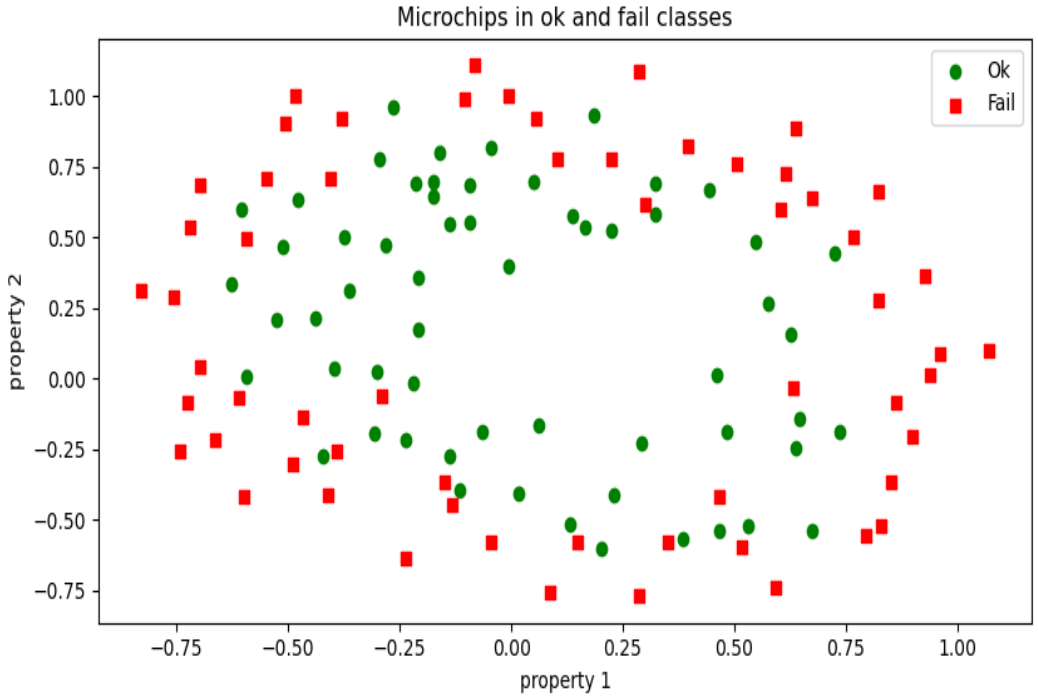

Figure 1: The image shows the plot of the original microchip data using different markers for the two classes OK and Fail

1.2 Part 2 Prediction of 3 Unknown Microchips

k = 1

Chip 1 : [-0.3, 1.0] ==> OK

Chip 2 : [-0.5, -0.1] ==> Fail

Chip 3 : [0.6, 0.0] ==> Fail

k = 3

Chip 1 : [-0.3, 1.0] ==> Fail

Chip 2 : [-0.5, -0.1] ==> Fail

Chip 3 : [0.6, 0.0] ==> OK

k = 5

Chip 1 : [-0.3, 1.0] ==> Fail

Chip 2 : [-0.5, -0.1] ==> OK

Chip 3 : [0.6, 0.0] ==> OK

k = 7

Chip 1 : [-0.3, 1.0] ==> Fail

Chip 2 : [-0.5, -0.1] ==> Fail

Chip 3 : [0.6, 0.0] ==> OK

1.3 Part 3 The Decision Boundary and the Training Errors

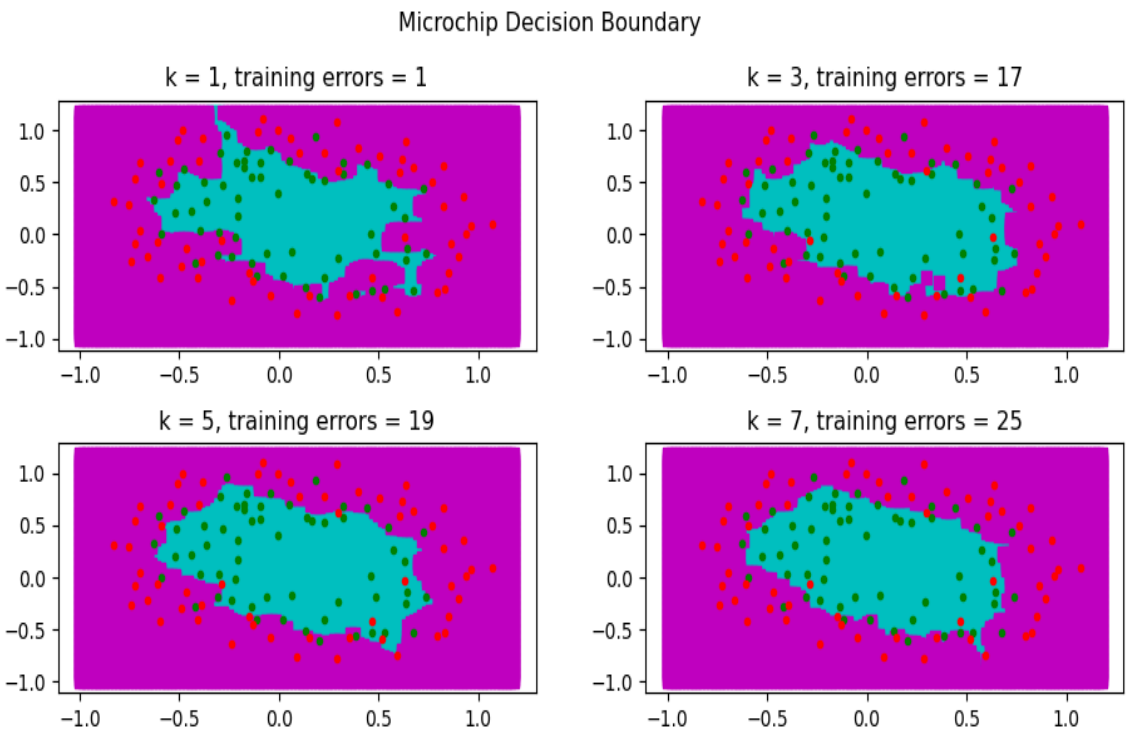

Figure 2: The image shows the decision boundary and the training error for k = 1, 3, 5, 7

Note: the plots may show errors more than stated in the titles. However, if zooming in, some seemingly error points are actually located in correct areas.

2 Exercise 2 k-NN Regression

2.1 Part 1 Divide the dataset into a training set and a test set

read file "polynomial200.csv"

rawData = np.loadtxt("polynomial200.csv", delimiter=",", usecols=(0, 1))

`training_set` = rawData[:100, :]

`test_set` = rawData[100:, :]

2.2 Part 2 the Training and Test Set in a 1 x 2 plot

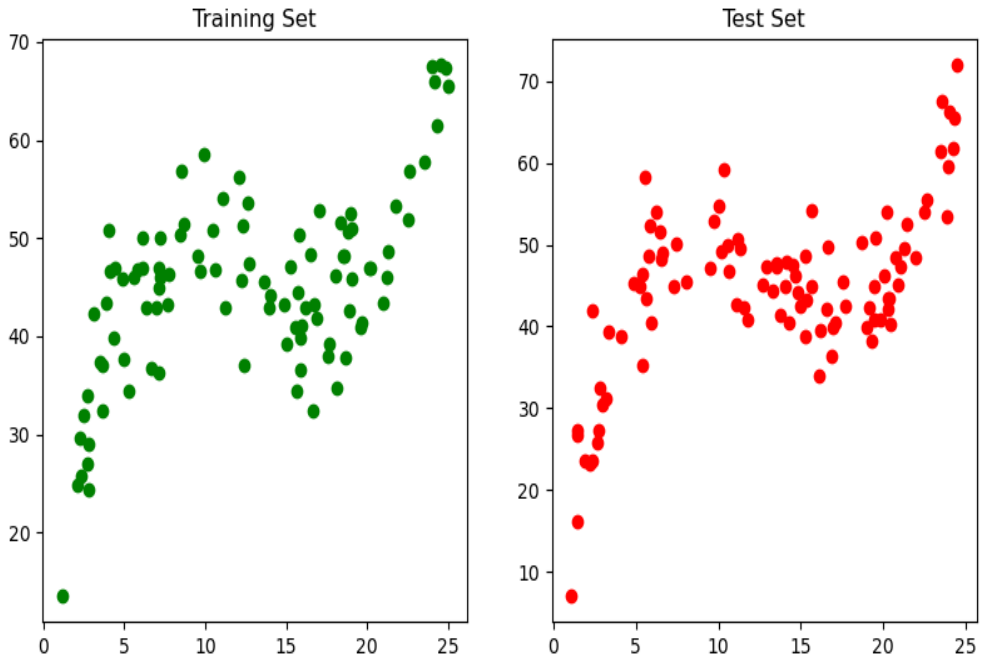

Figure 3: The image shows the training and test sets

2.3 Part 3 the k-NN Regression Result and the MSE Training Error

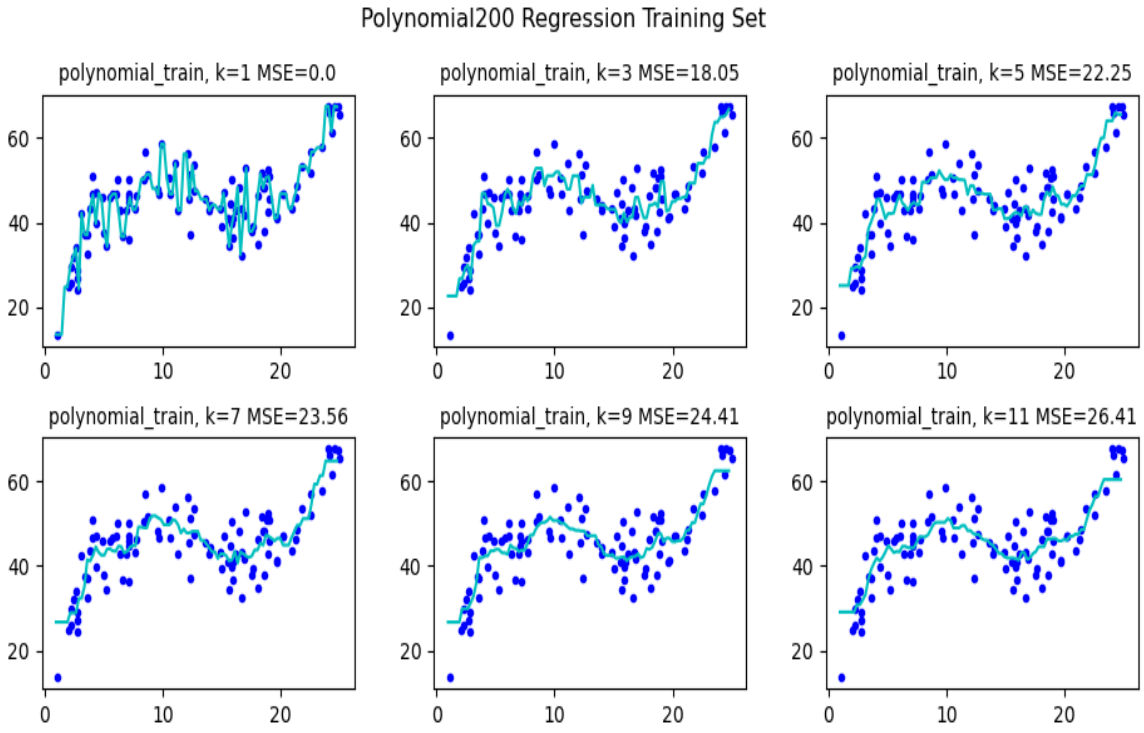

Figure 4: The image shows the k-NN regression result and the MSE training error for k = 1, 3, 5, 7, 9, 11

2.4 Part 4 the k-NN Regression Result and the MSE Test Error

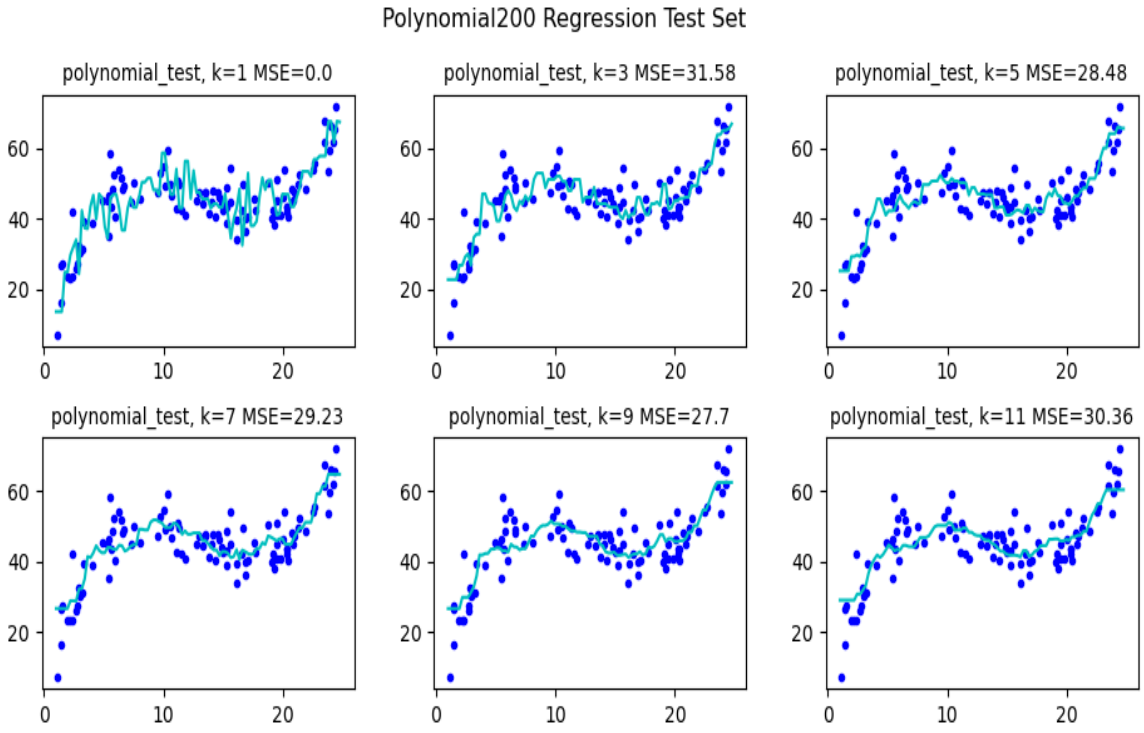

Figure 5: The image shows the k-NN regression result and the MSE test error for k = 1, 3, 5, 7, 9, 11

2.5 Part 5 The k That Gives the Best Regression

k = 5 gives the best regression.

When k = 5, its Mean Square Error(MSE) in training set is the second smallest (22.25) and its MSE is the second smallest in test set as well (28.48). The smaller MSE is, the better quality regression is. Therefore, combining two results on both training and test sets, k=5 is the best choice.

3 Exercise 4 k-NN Classification using scikit-learn

3.1 Part 1 Original Microchip Data

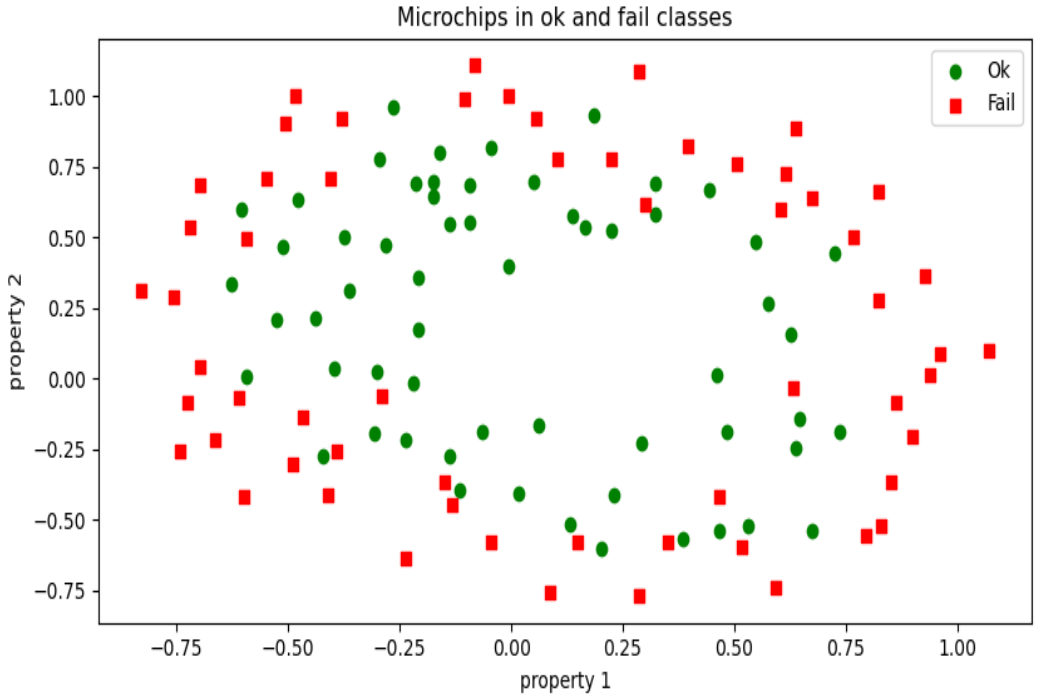

Figure 6: The image shows the plot of the original microchip data using different markers for the two classes OK and Fail

3.2 Part 2 Prediction of 3 Unknown Microchips

k = 1

Chip 1 : [-0.3, 1.0] ==> OK

Chip 2 : [-0.5, -0.1] ==> Fail

Chip 3 : [0.6, 0.0] ==> Fail

k = 3

Chip 1 : [-0.3, 1.0] ==> Fail

Chip 2 : [-0.5, -0.1] ==> Fail

Chip 3 : [0.6, 0.0] ==> OK

k = 5

Chip 1 : [-0.3, 1.0] ==> Fail

Chip 2 : [-0.5, -0.1] ==> OK

Chip 3 : [0.6, 0.0] ==> OK

k = 7

Chip 1 : [-0.3, 1.0] ==> Fail

Chip 2 : [-0.5, -0.1] ==> Fail

Chip 3 : [0.6, 0.0] ==> OK

3.3 Part 3 The Decision Boundary and the Training Errors

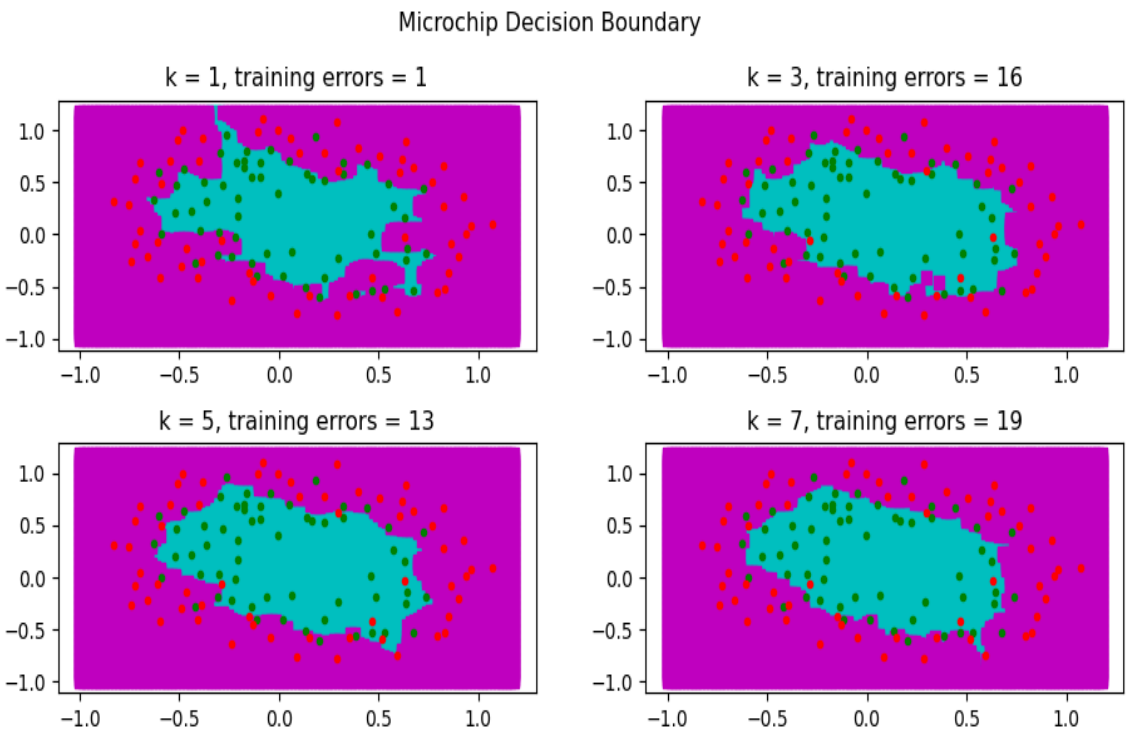

Figure 7: The image shows the decision boundary and the training error for k = 1, 3, 5, 7

Note: the plots may show errors more than stated in the titles. However, if zooming in, some seemingly error points are actually located in correct areas.