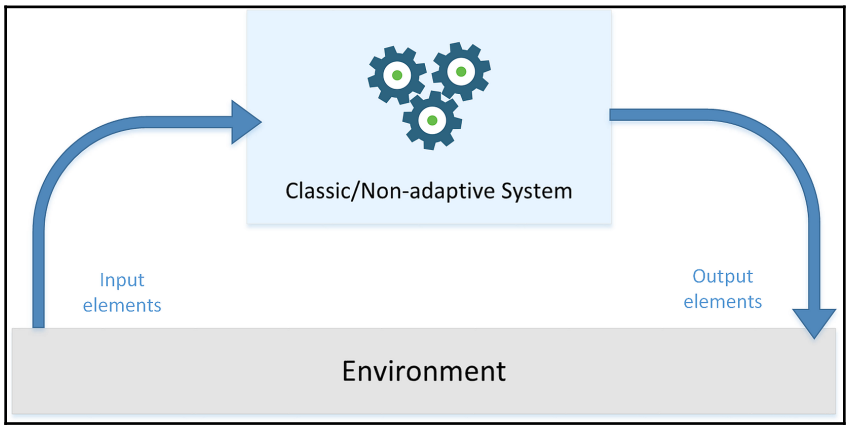
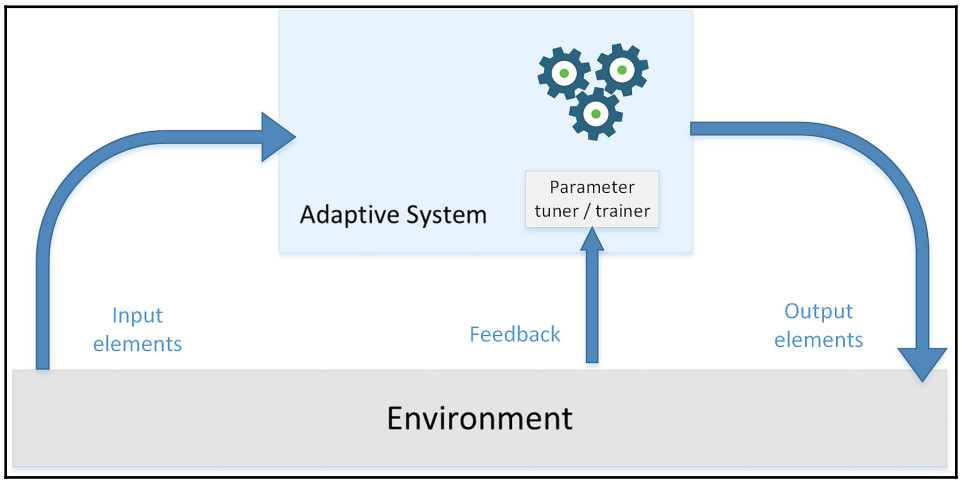
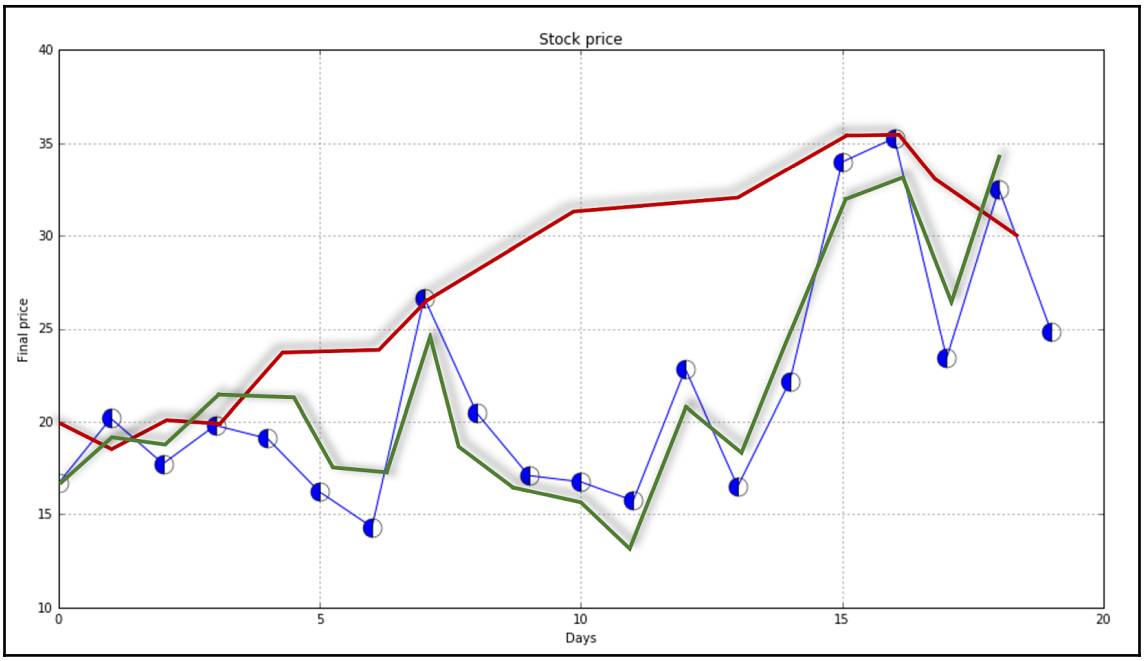
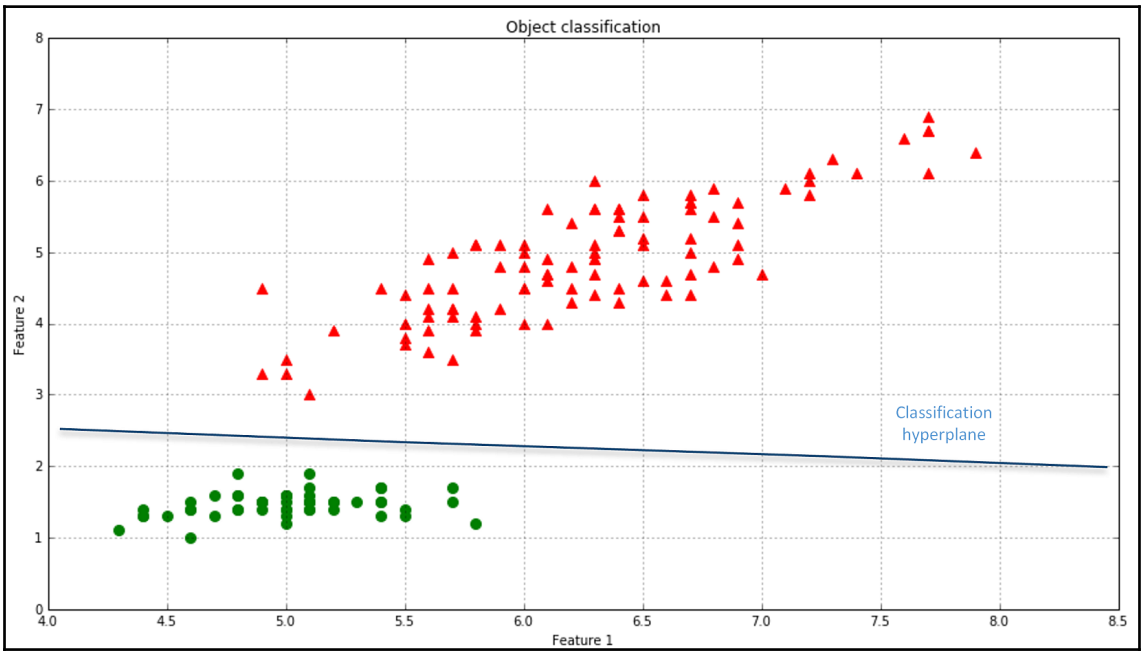
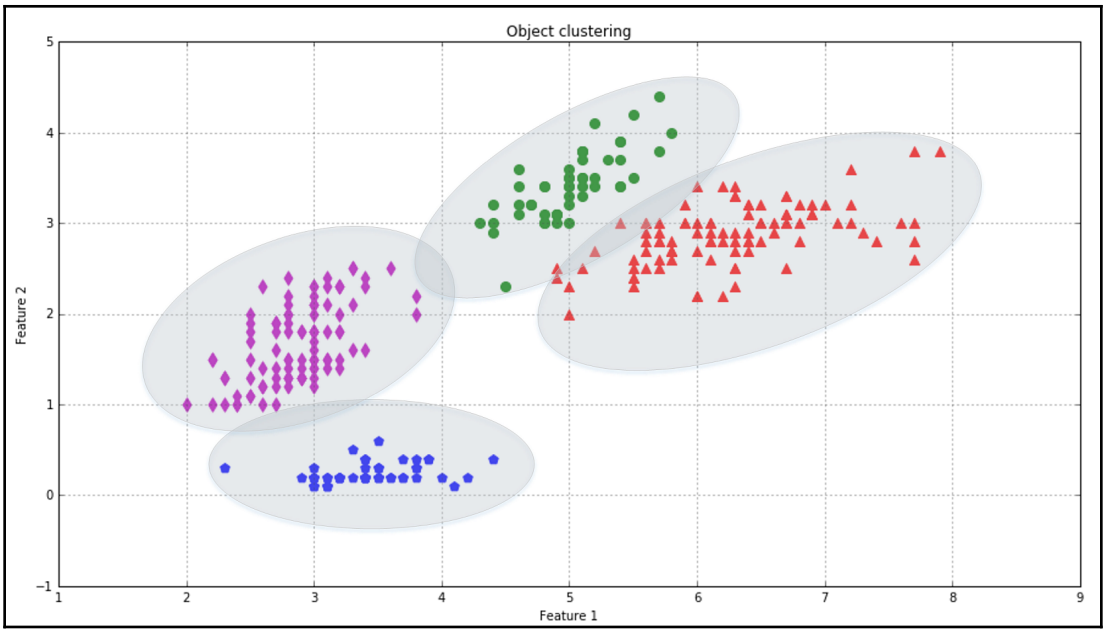
Chapter 01: A Gentle Introduction to Machine Learning

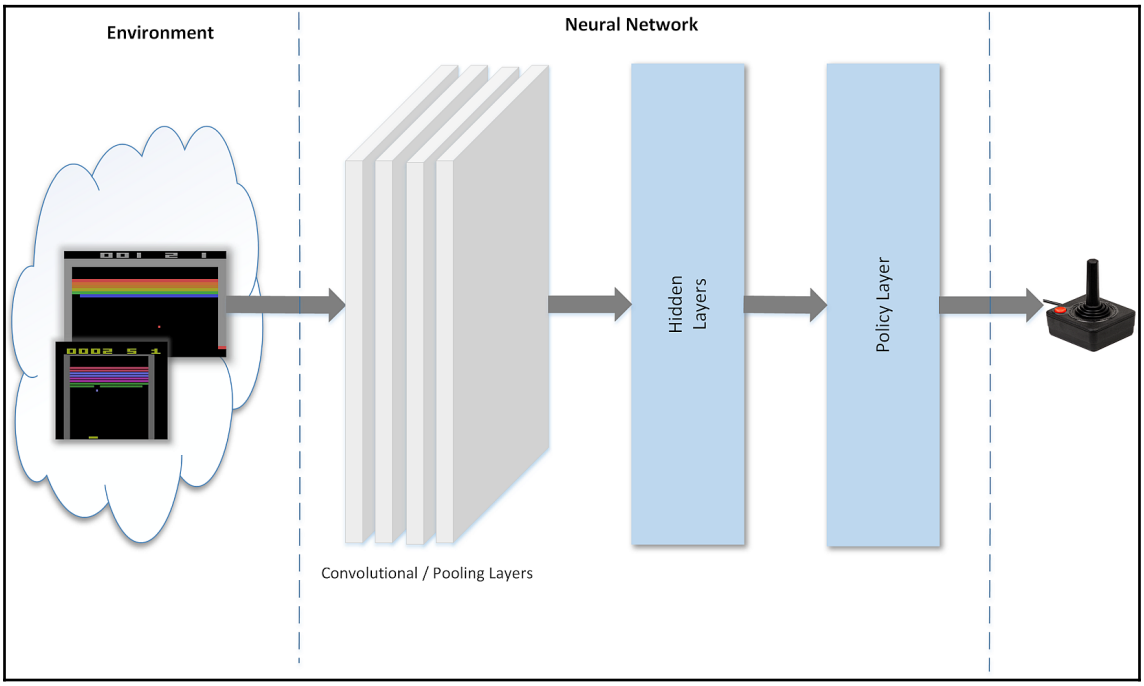




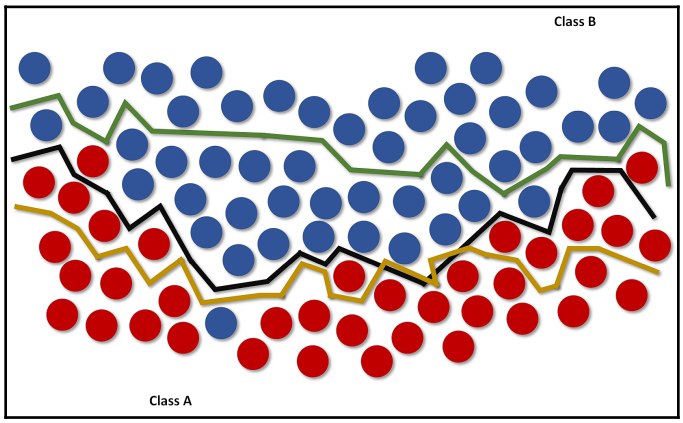

Stock price
Final price
Days
40
35
30
25
20
15
10
0
5
10
15
20

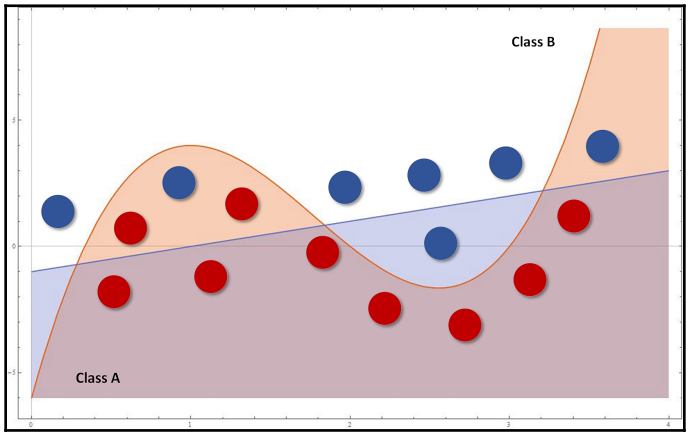

Object classification
Feature 2
Feature 1
Classification
hyperplane
8
7
6
5
4
3
2
1
0
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5

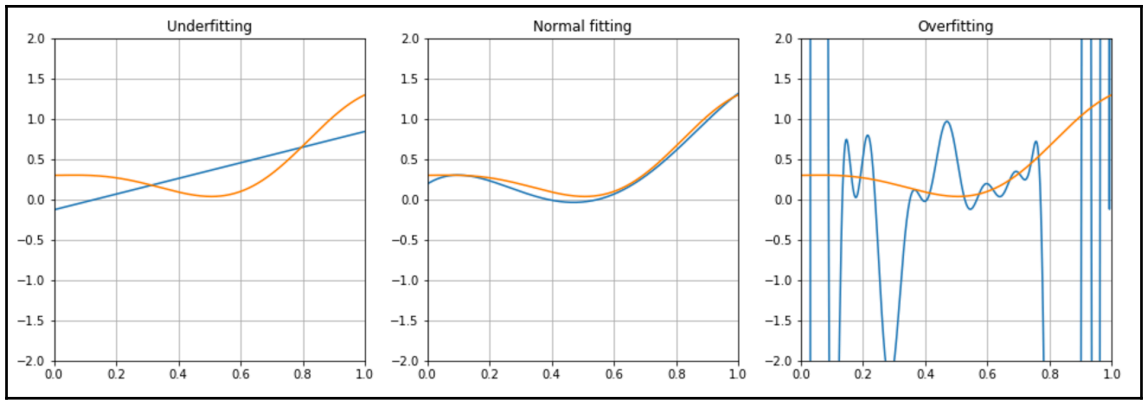

Object clustering
Feature 2
Feature 1
5
4
3
2
1
0
−1
1
2
3
4
5
6
7
8
9

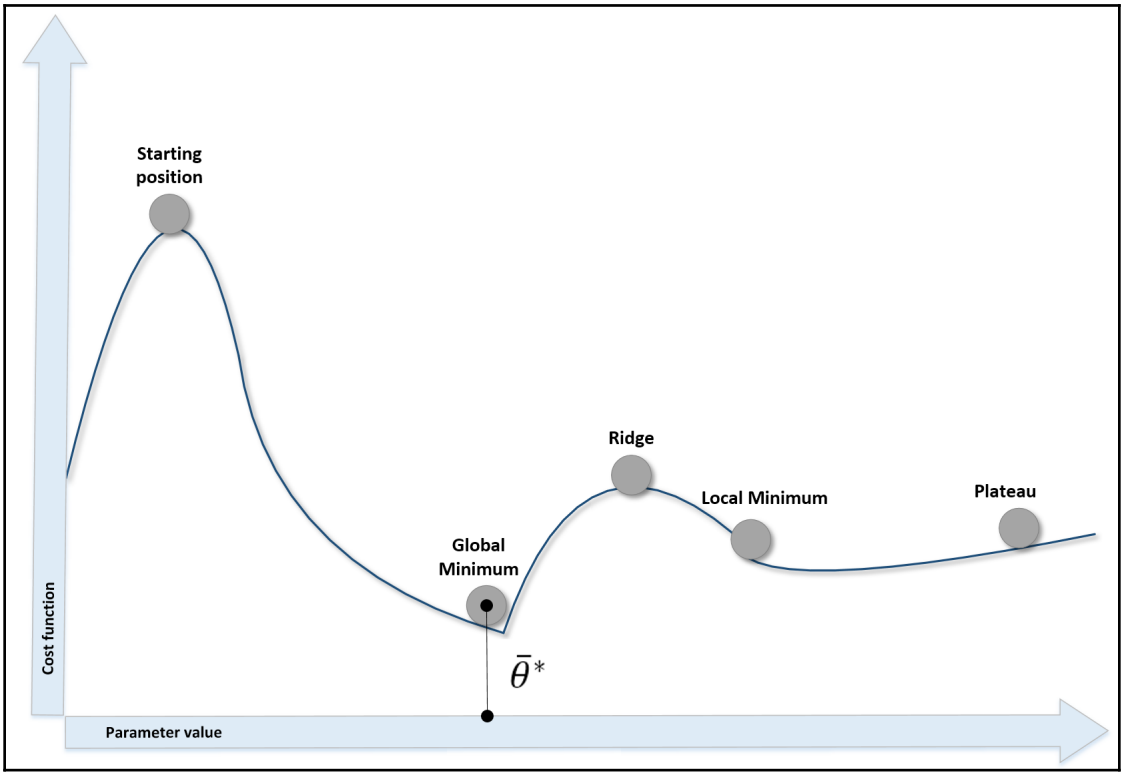

Environment
Neural Network
Hidden Layers
Policy Layer
Convolutional / Pooling Layers

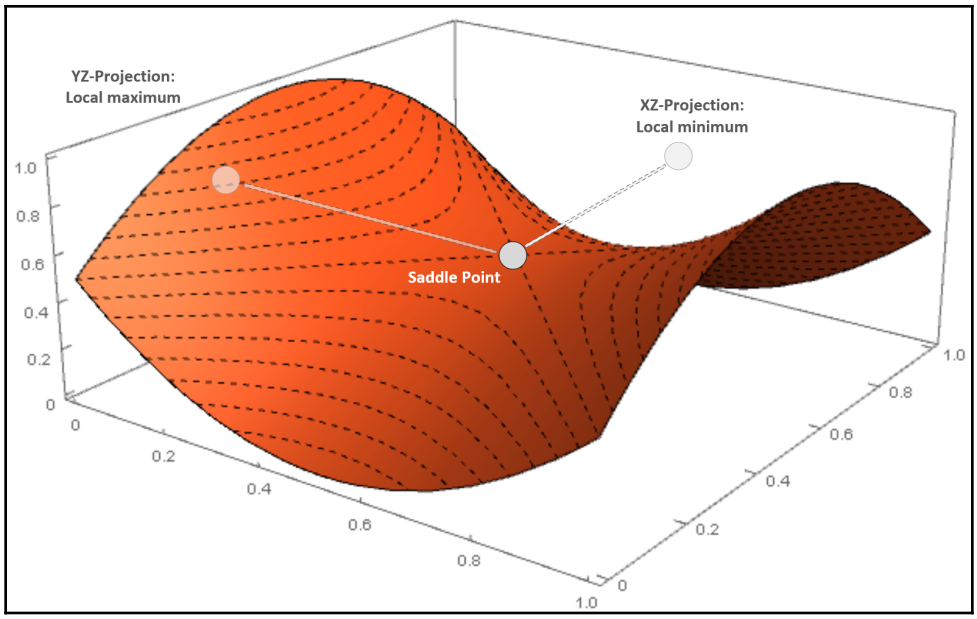
Chapter 02: Important Elements in Machine Learning

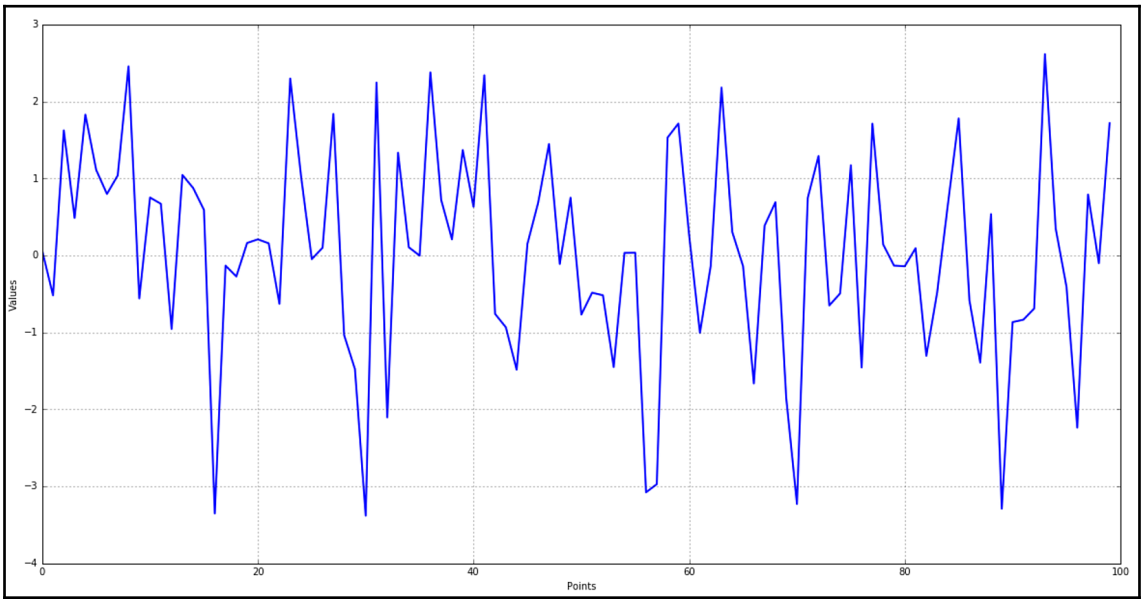


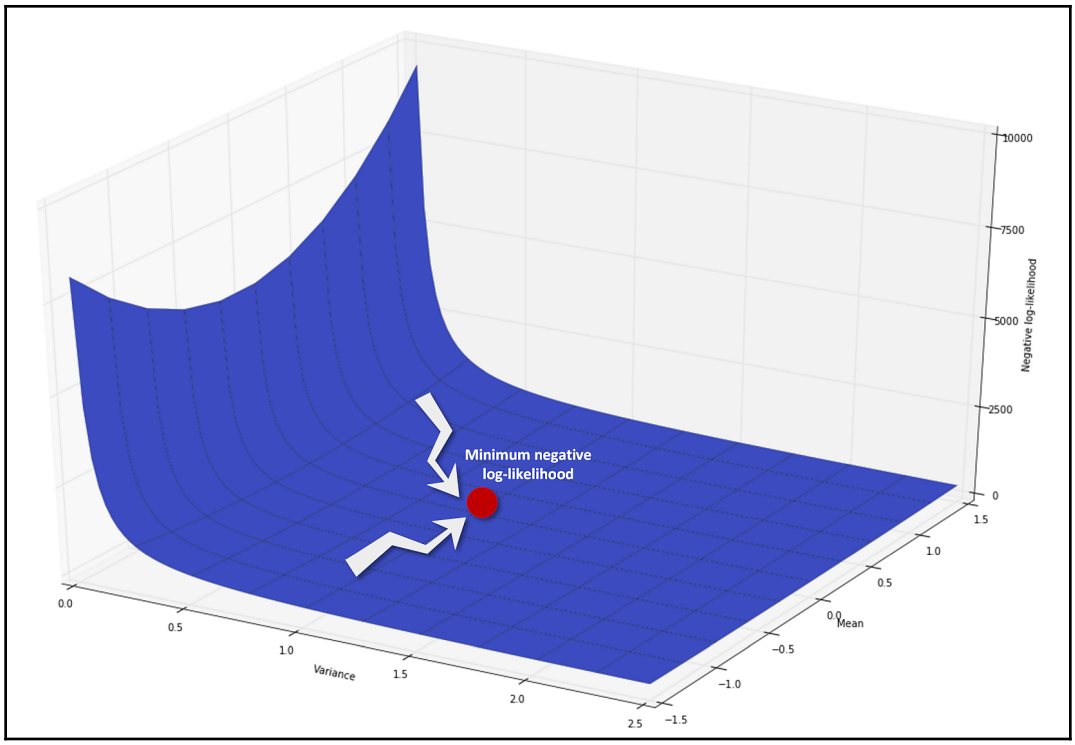


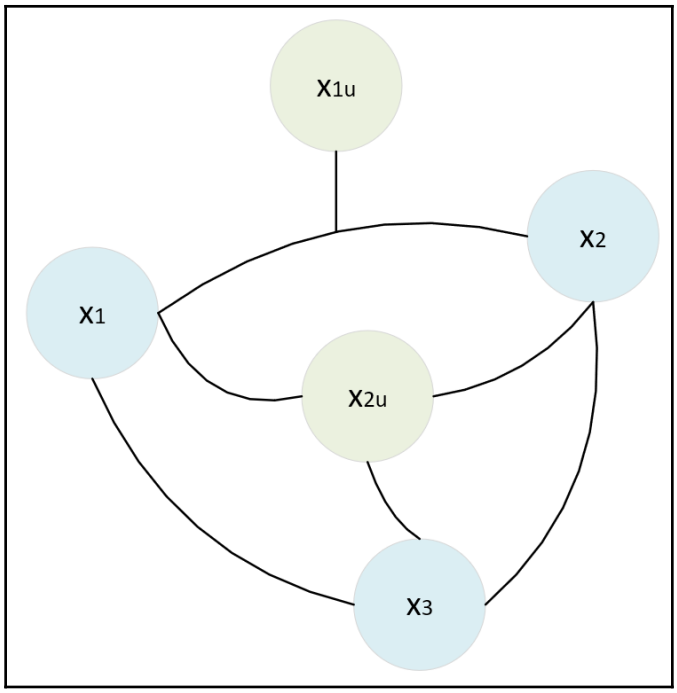

Underfitting
Normal fitting
Overfitting

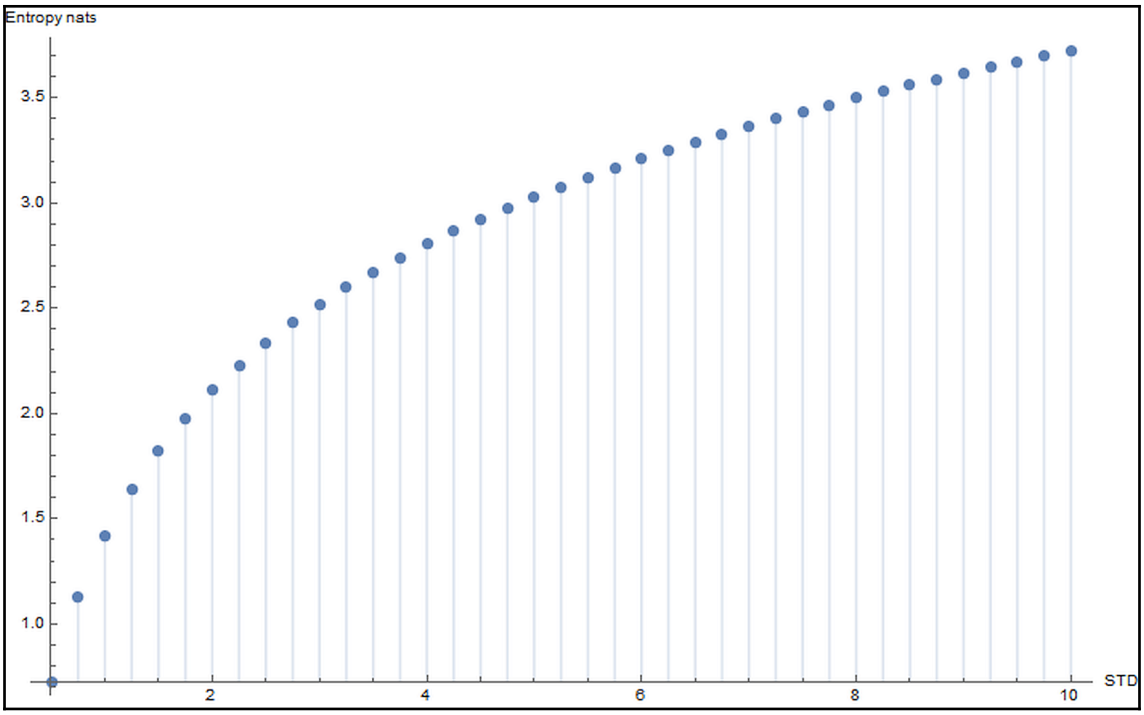

Starting position
Ridge
Local Minimum
Plateau
Global Minimum
$\bar{\theta}*$
Cost function
Parameter value


YZ-Projection:
Local maximum
XZ-Projection:
Local minimum
Saddle Point
1.0
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1.0

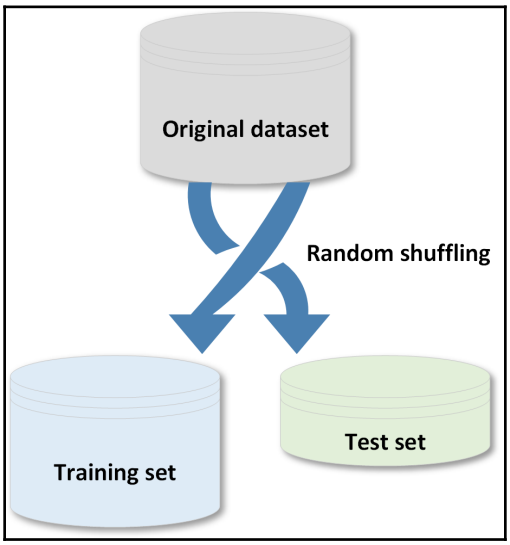

Values
Points
3
2
1
0
−1
−2
−3
−4
0
20
40
60
80
100

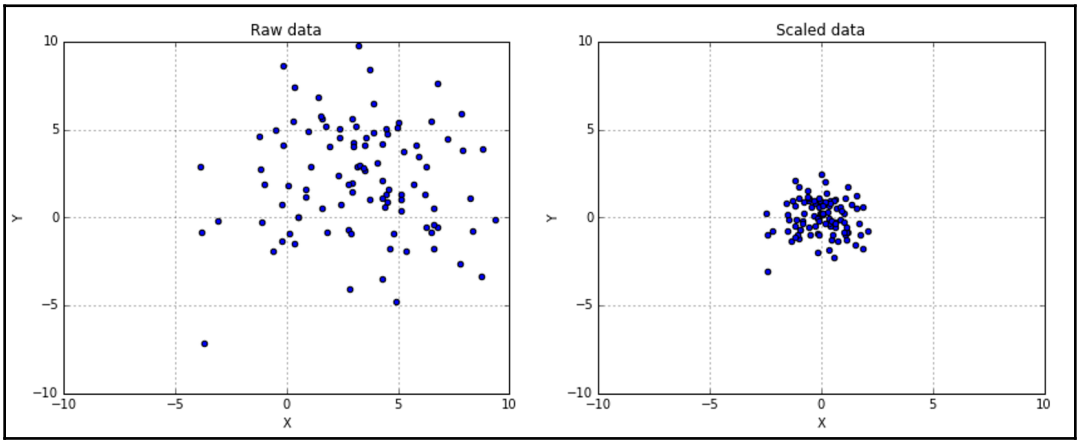

Minimum negative
log-likelihood
Negative log-likelihood
10000
7500
5000
2500
0
Variance
0.0
0.5
1.0
1.5
2.0
2.5
Mean
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

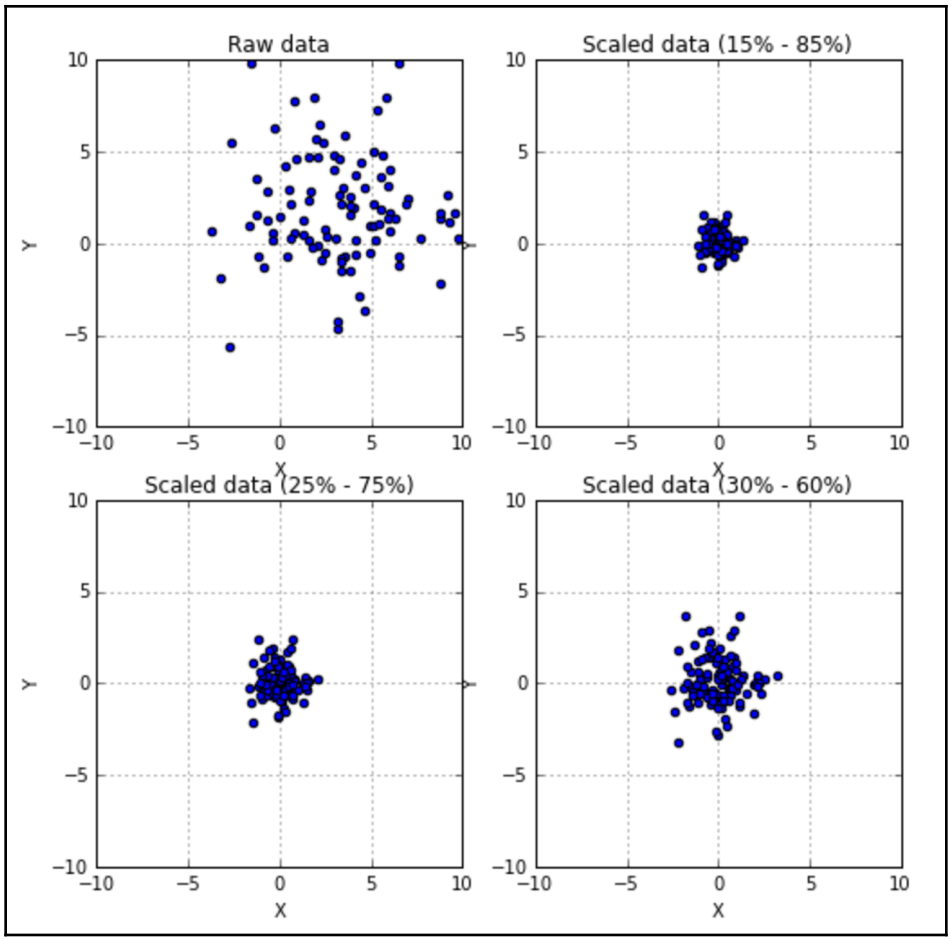

X1u
X2
X1
X2u
X3

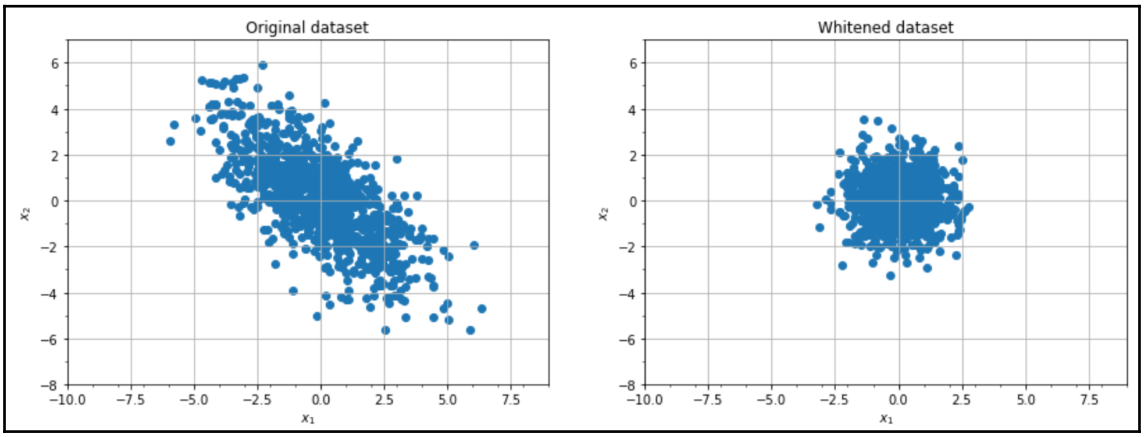

Entropy nats
3.5
3.0
2.5
2.0
1.5
1.0
2
4
6
8
10
STD

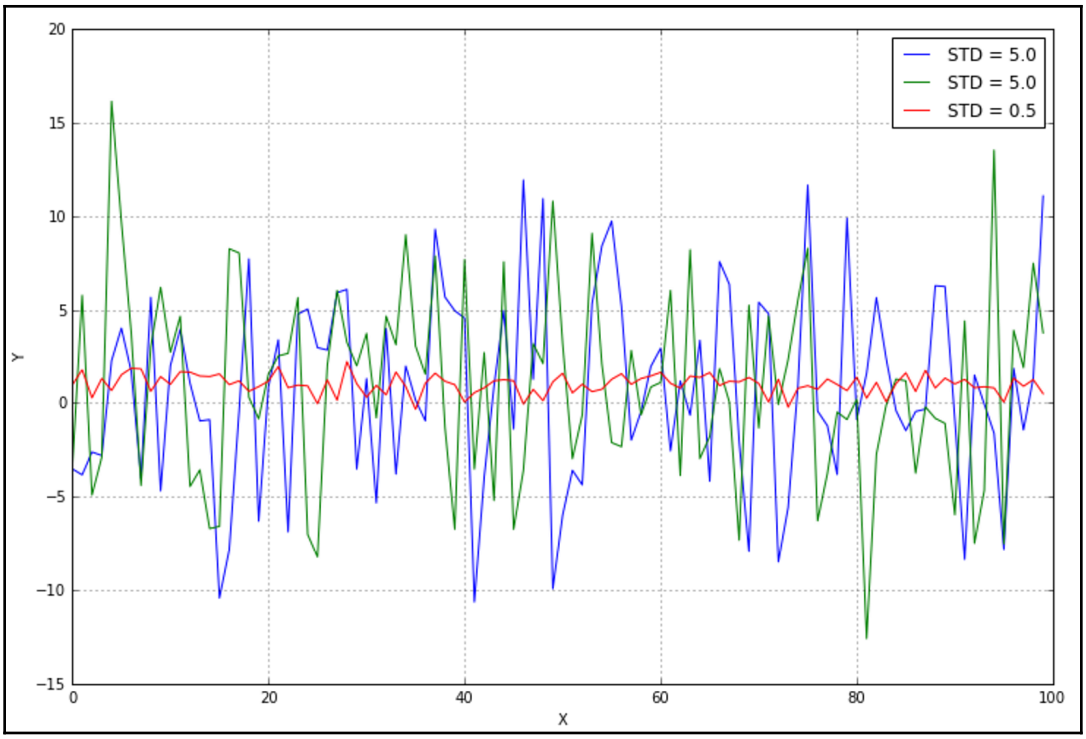
Chapter 03: Feature Selection and Feature Engineering

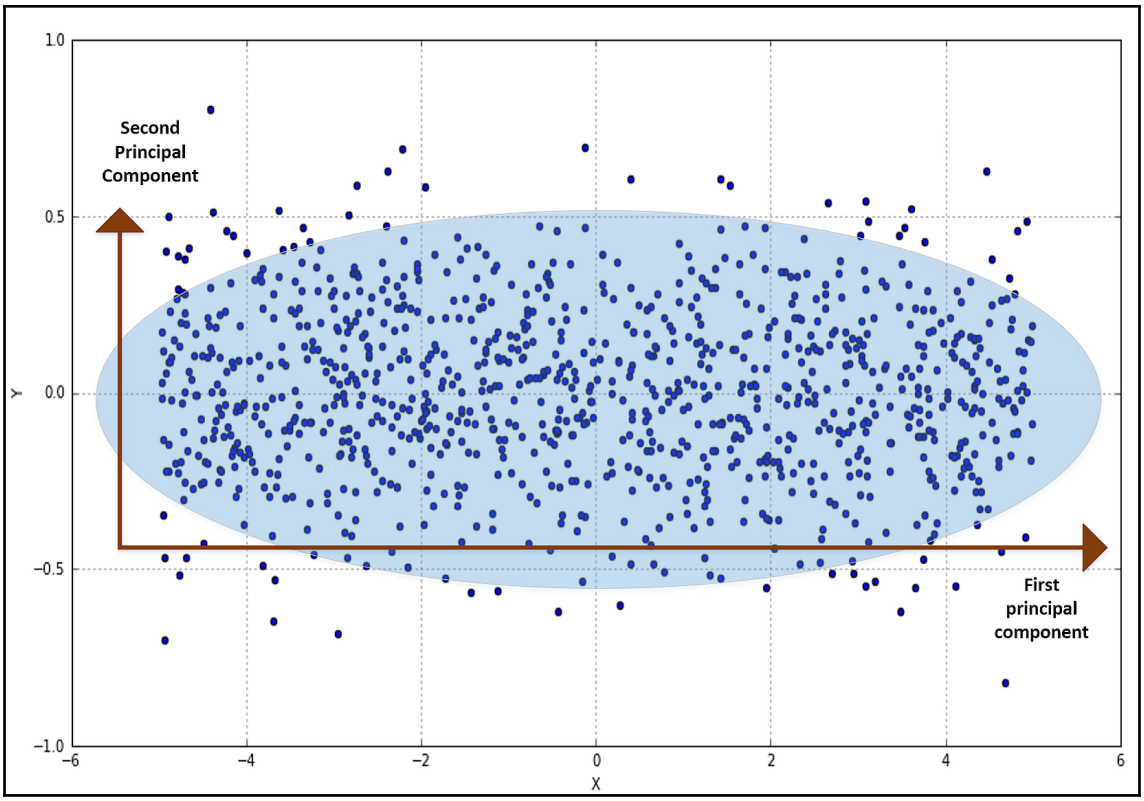


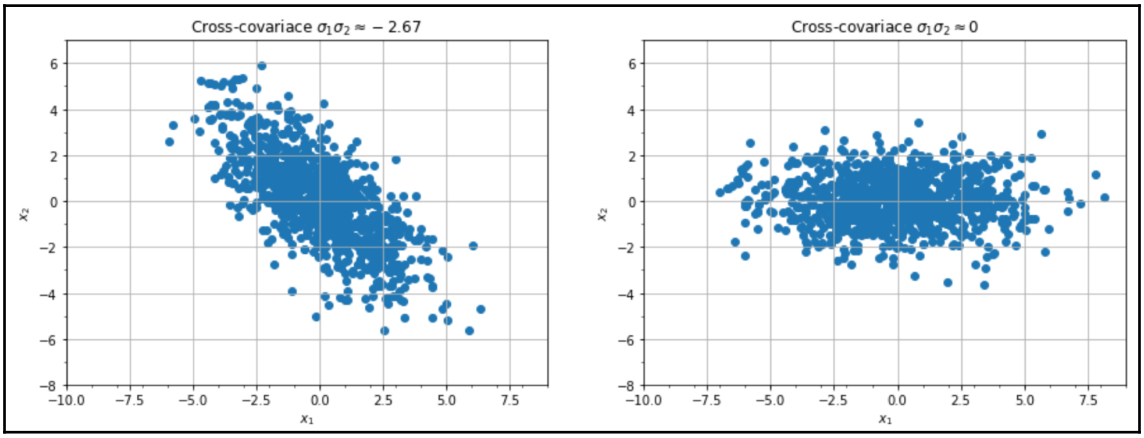


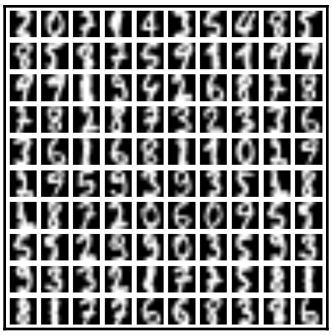

Raw data
Scaled data (15% - 85%)
Scaled data (25% - 75%)
Scaled data (30% - 60%)
X
Y

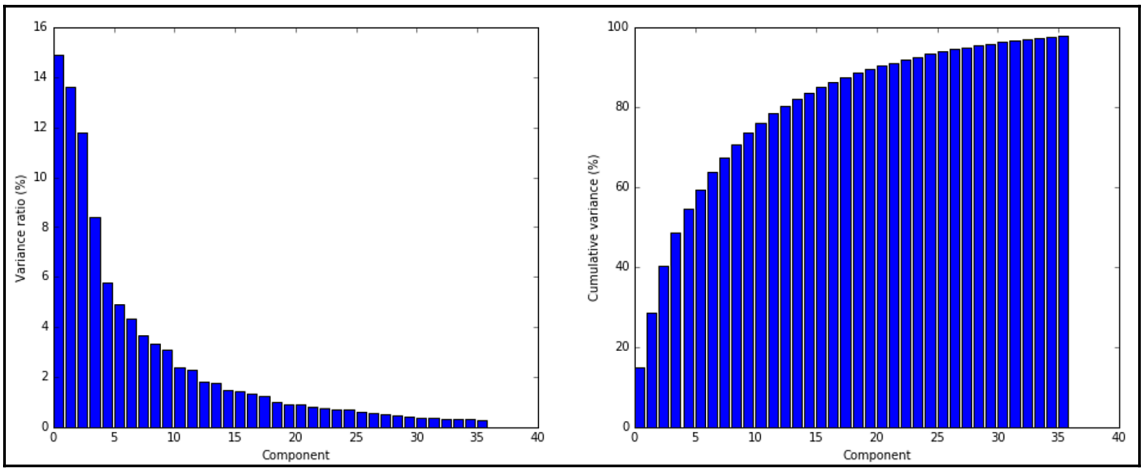

Original dataset
Whitened dataset
x_1
x_2

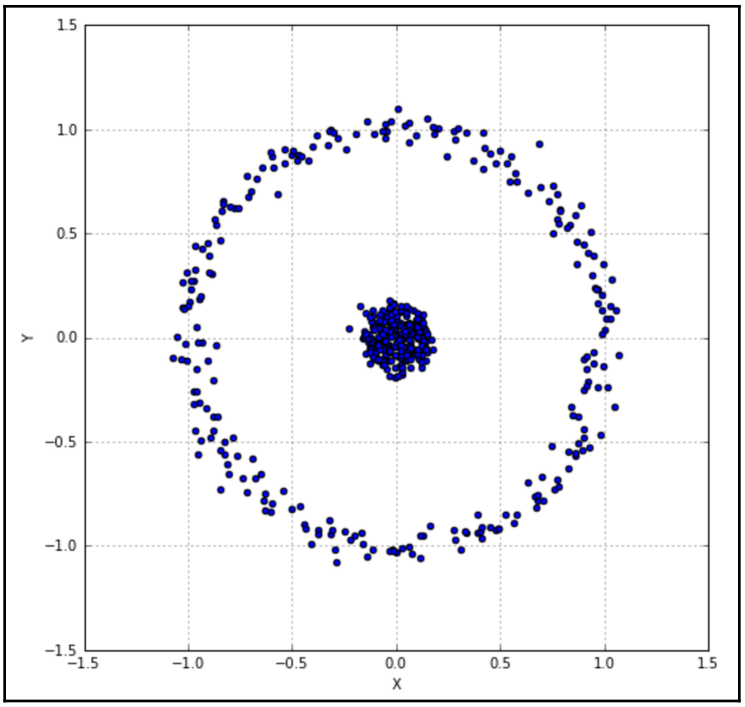

STD = 5.0
STD = 5.0
STD = 0.5
20
15
10
5
0
−5
−10
−15
0
20
40
60
80
100
X
Y

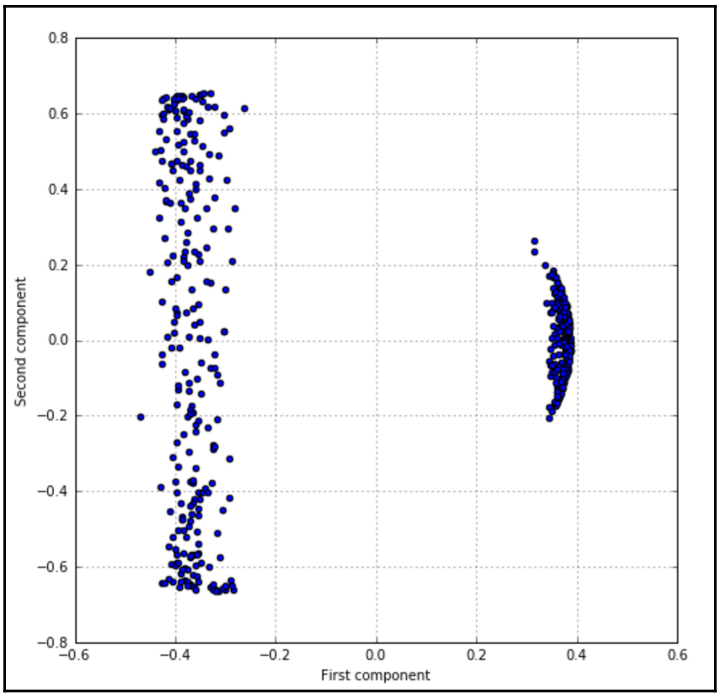

Second
Principal
Component
First
principal
component
Y
X

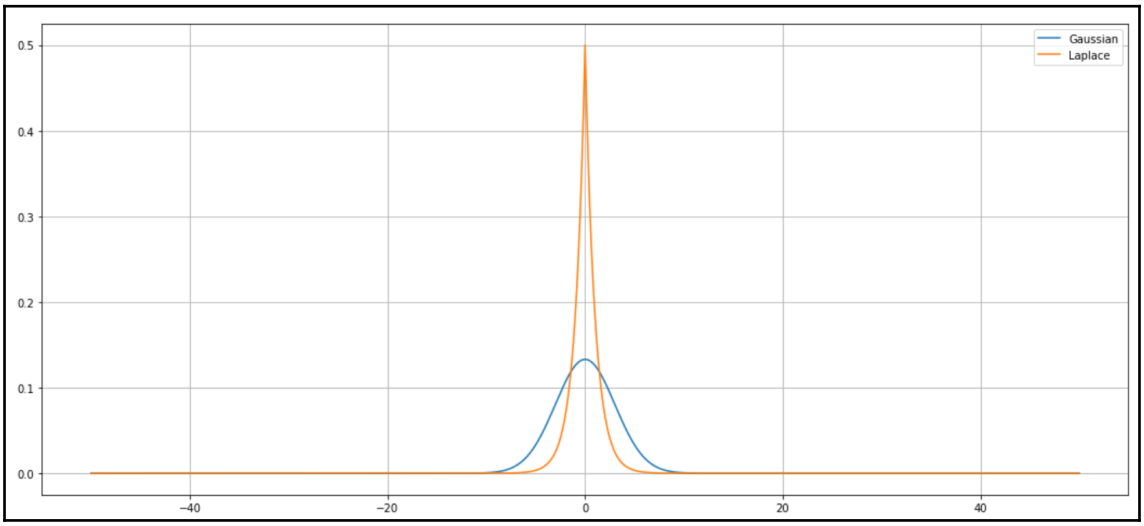

Cross-covariace $\sigma_1\sigma_2 \approx -2.67$
Cross-covariace $\sigma_1\sigma_2 \approx 0$
x_1
x_2

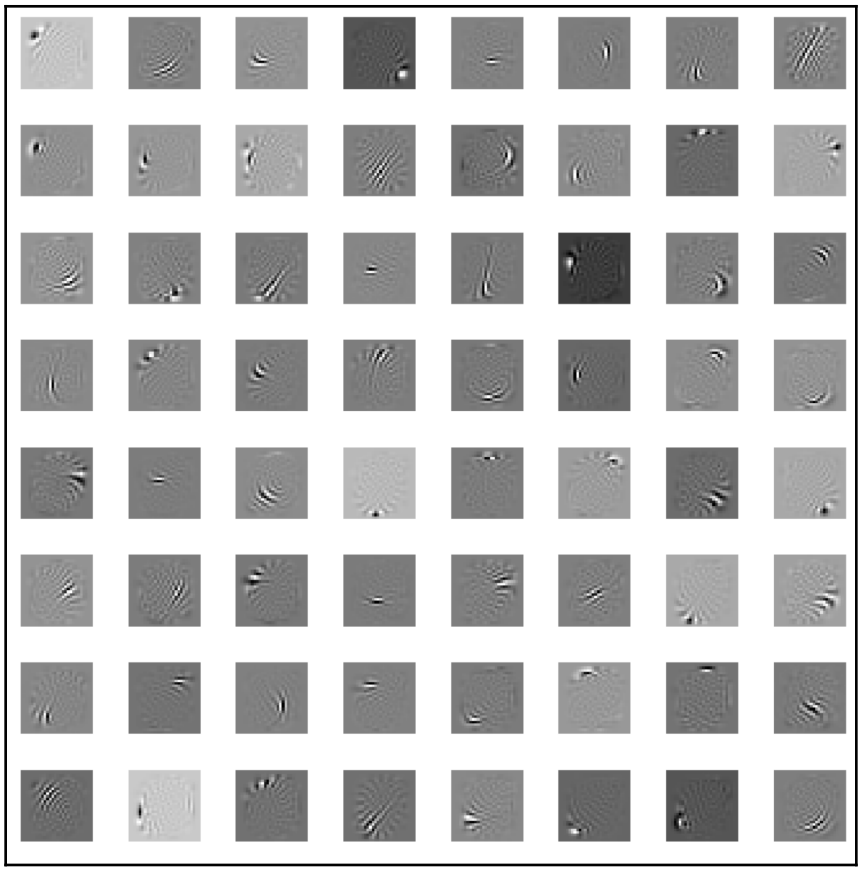


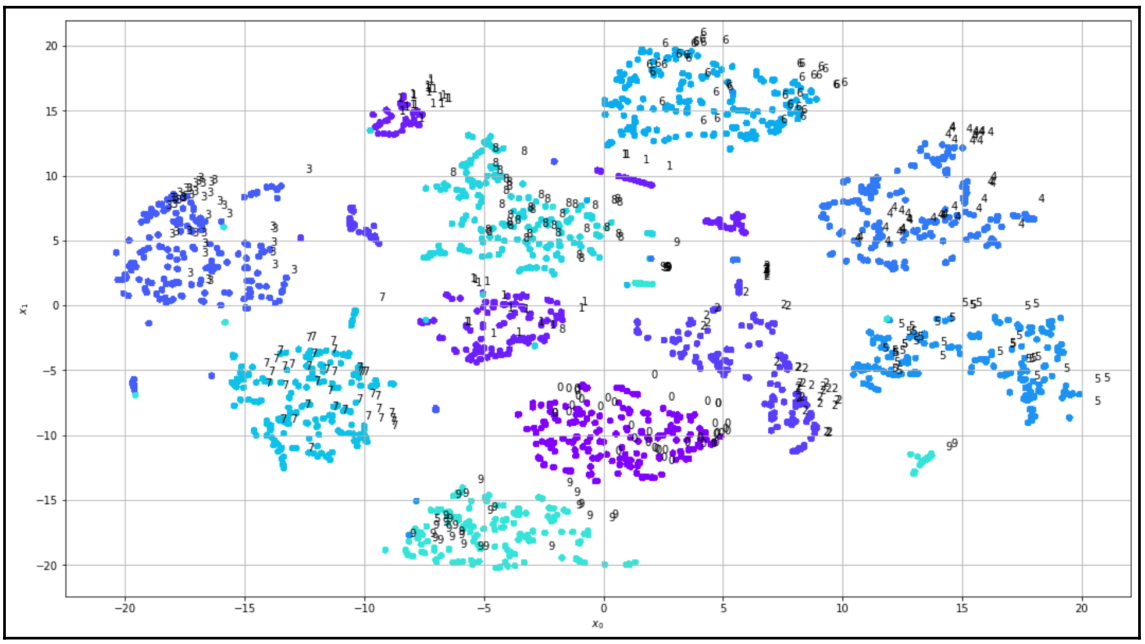

Variance ratio (%)
Component
Cumulative variance (%)
Component


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
X

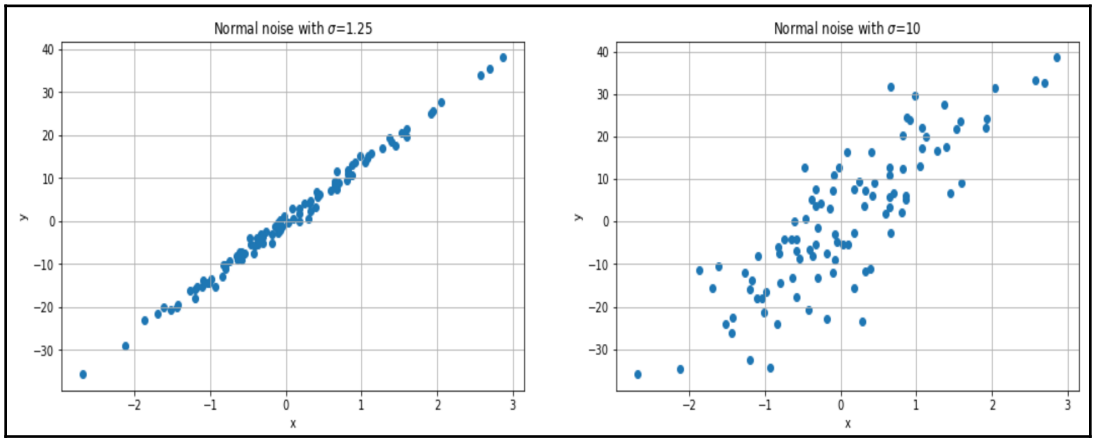

Second component
First component
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

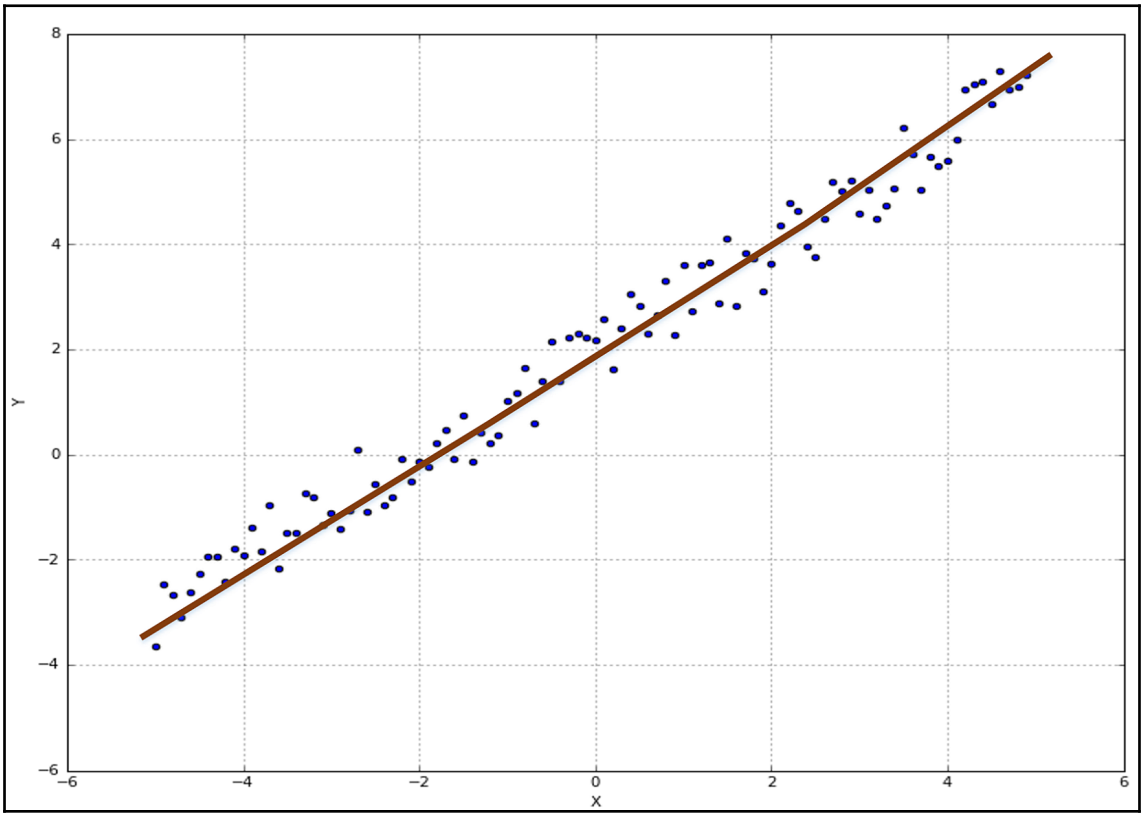

Gaussian
Laplace
0.5
0.4
0.3
0.2
0.1
0.0
−40
−20
0
20
40

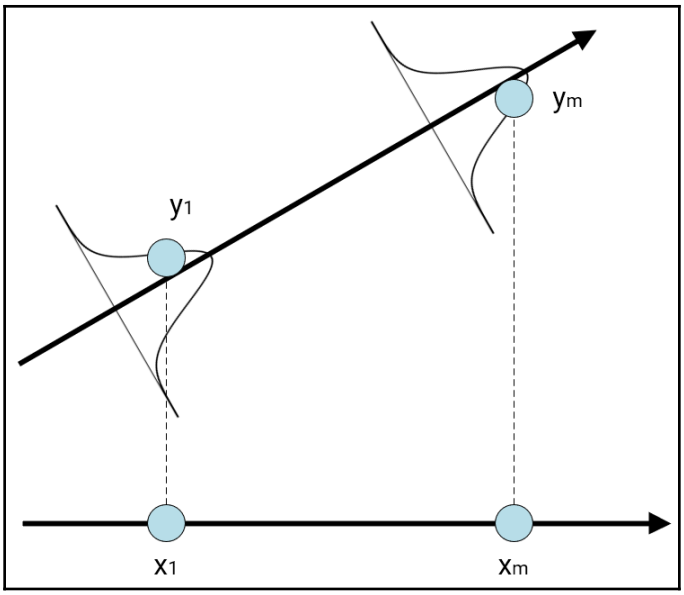


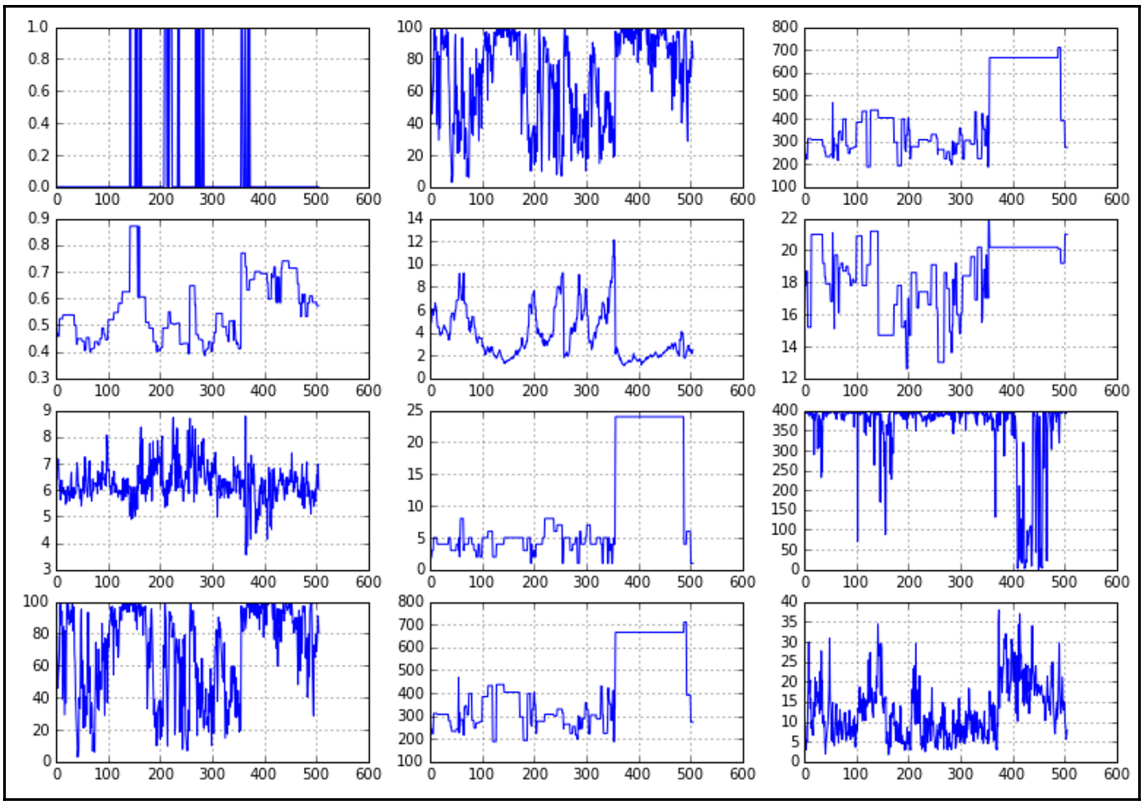

20
15
10
5
0
−5
−10
−15
−20
x_1
−20
−15
−10
−5
0
5
10
15
20
x_0

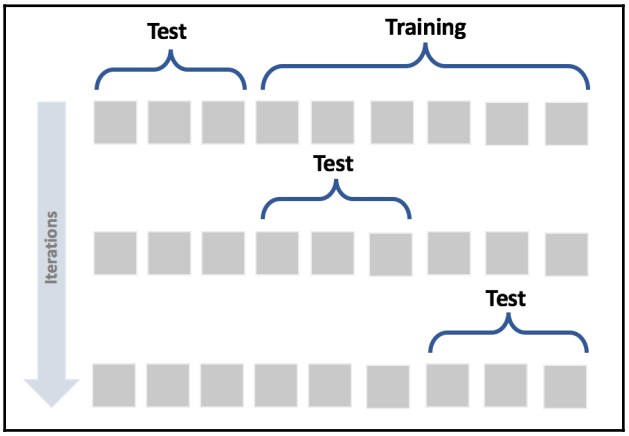
Chapter 04: Regression Algorithms

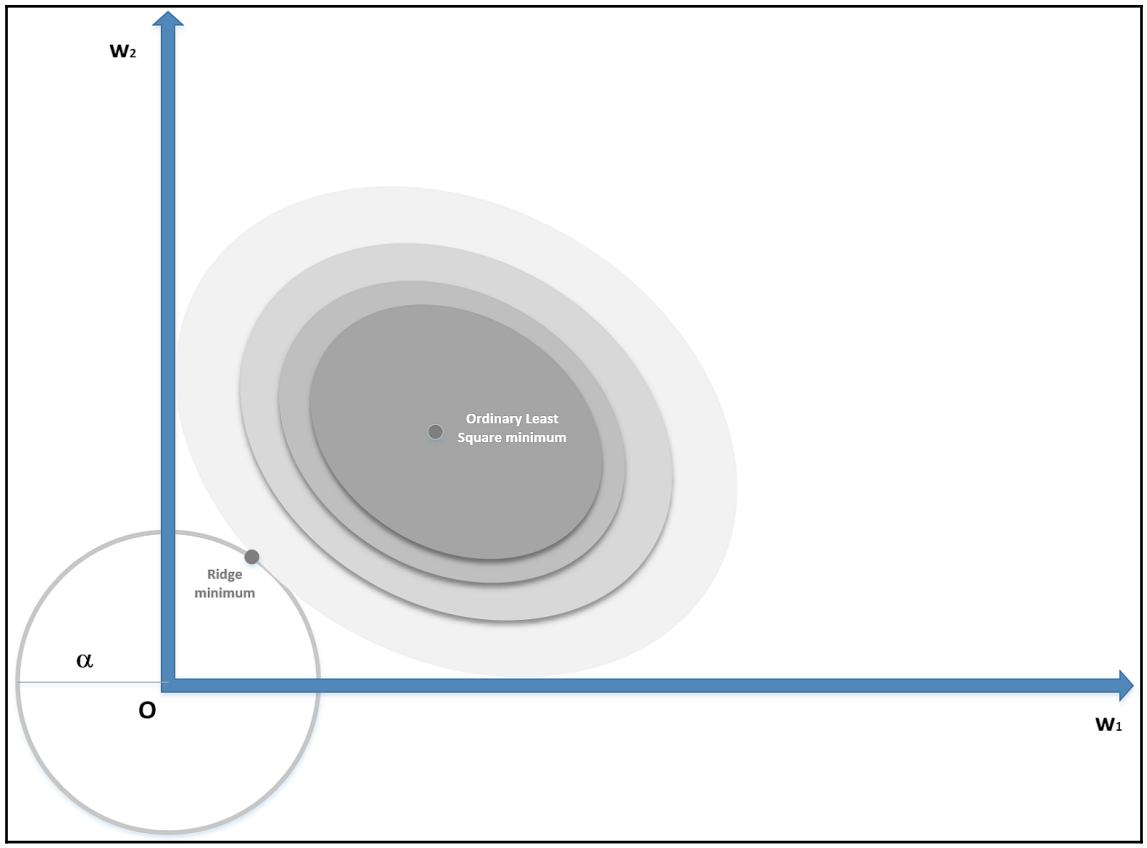


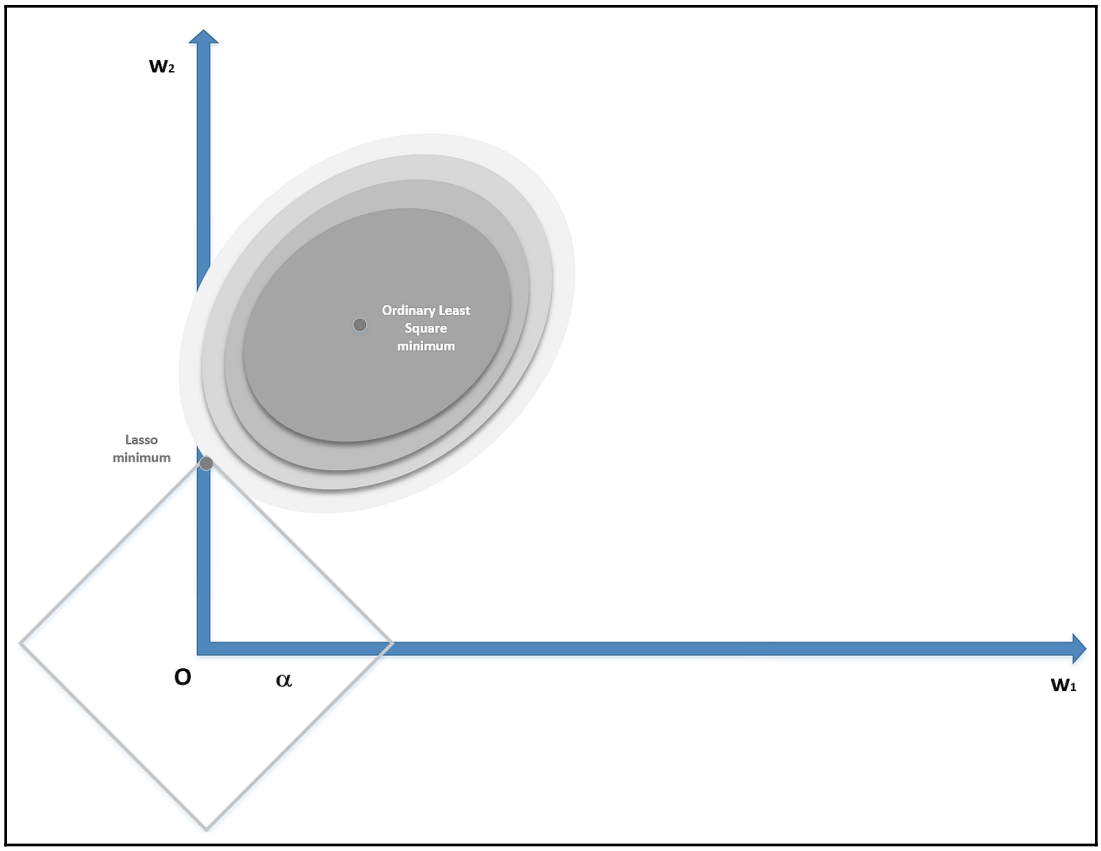


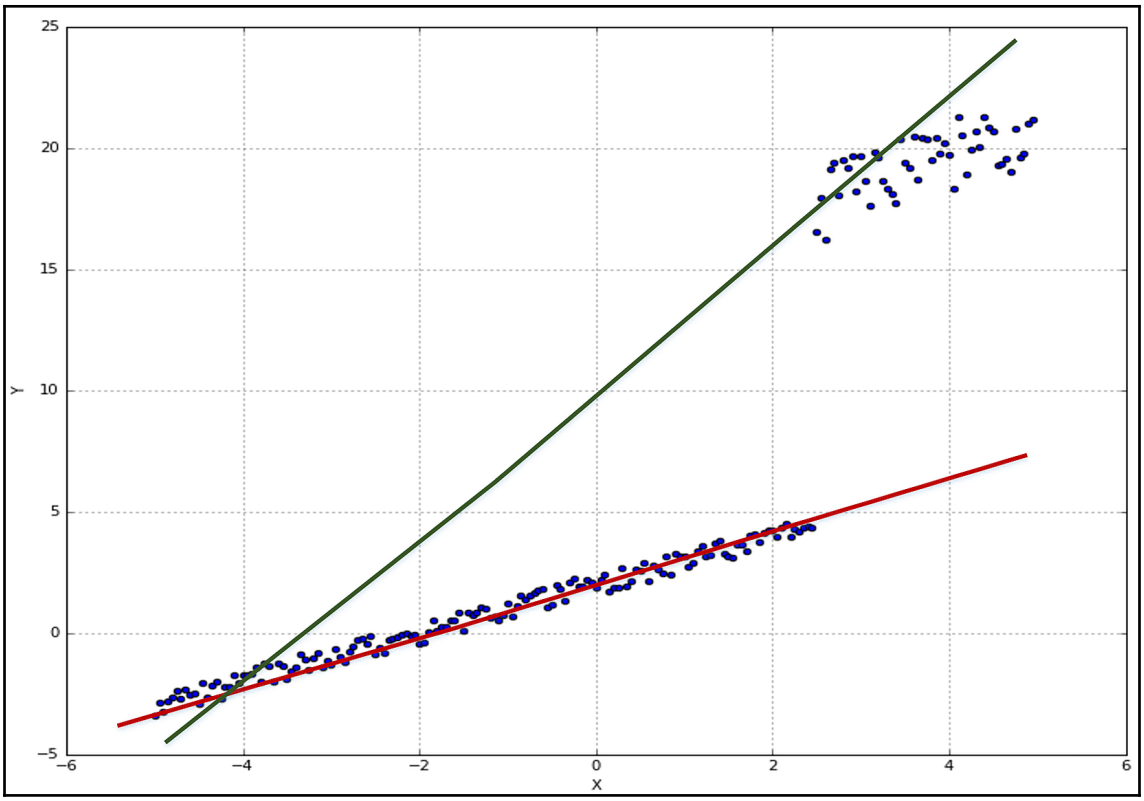

ym
y1
x1
xm

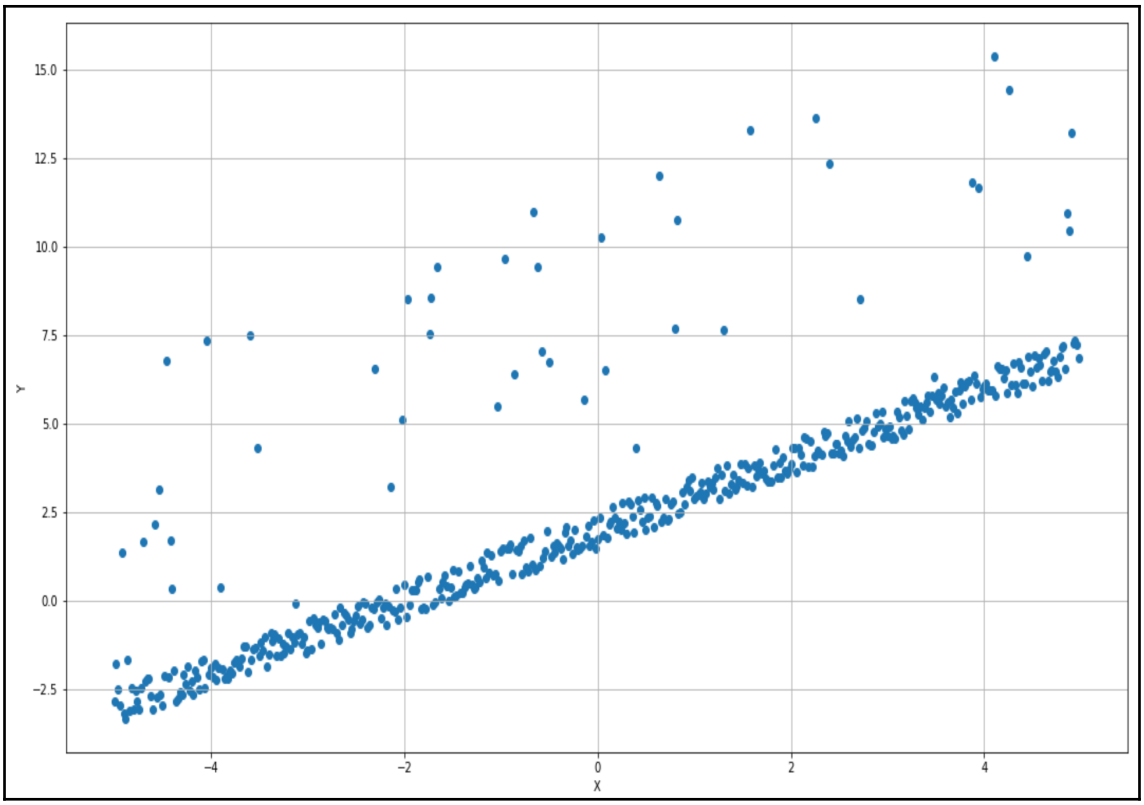


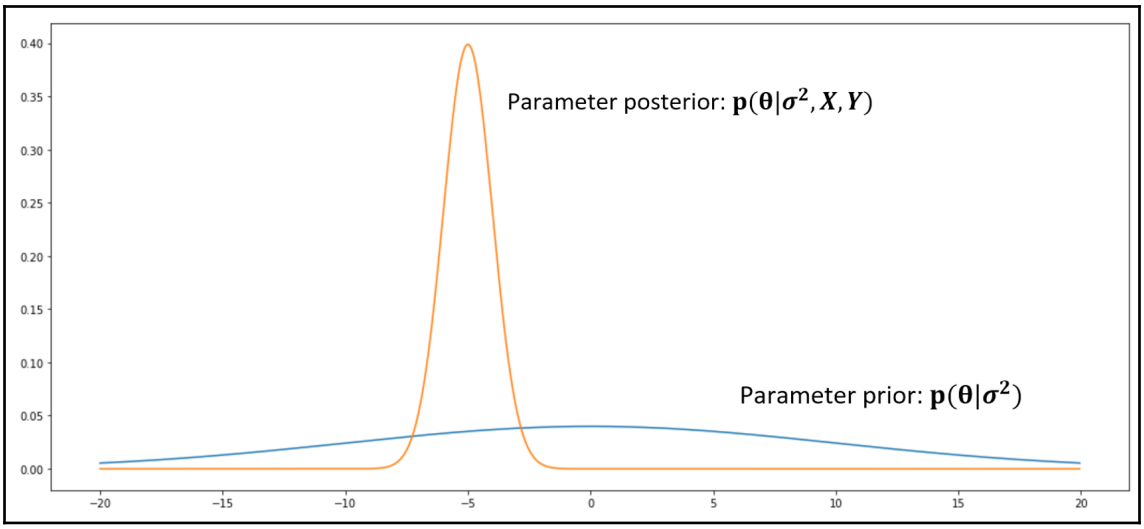

Test
Training
Test
Test
Iterations

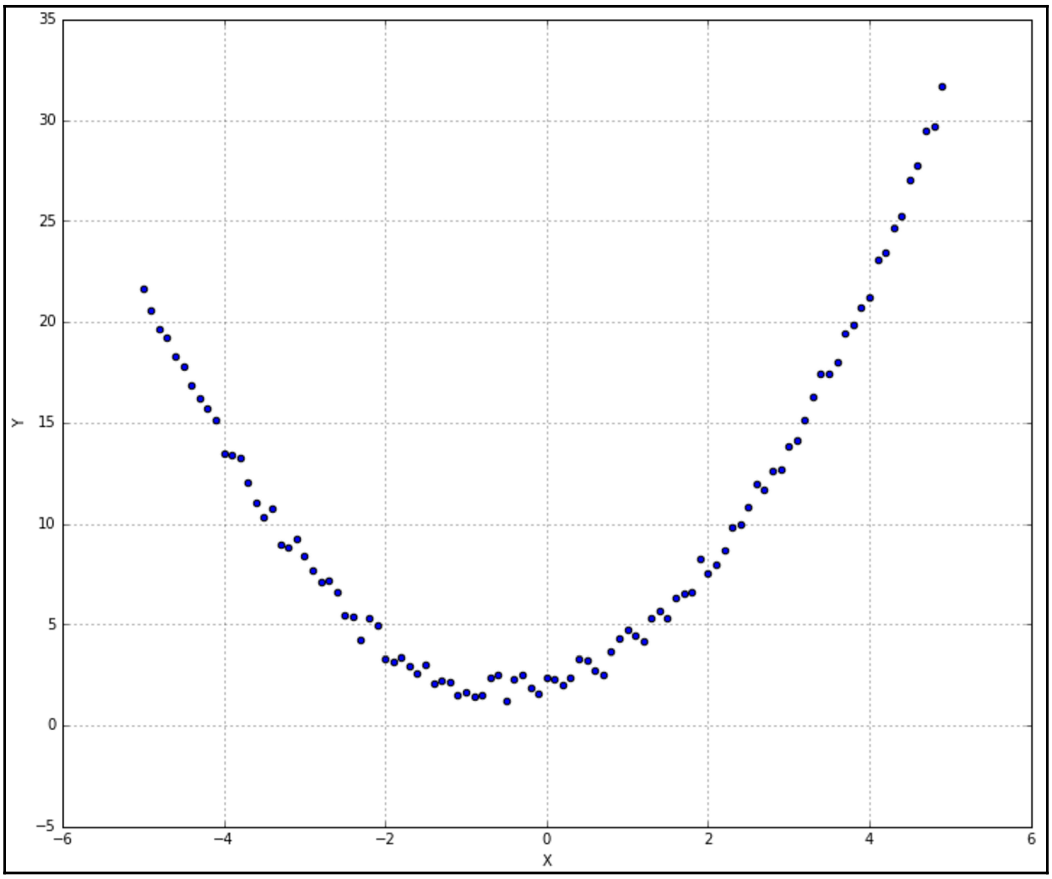

w2
Ordinary Least
Square minimum
Ridge
minimum
α
O
w1

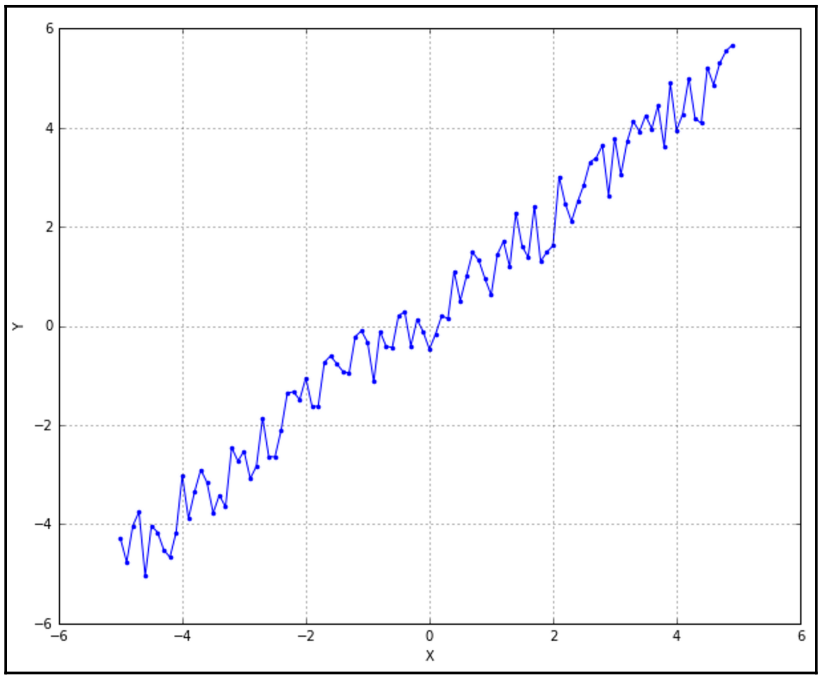

w2
Ordinary Least Square minimum
Lasso minimum
O
α
w1

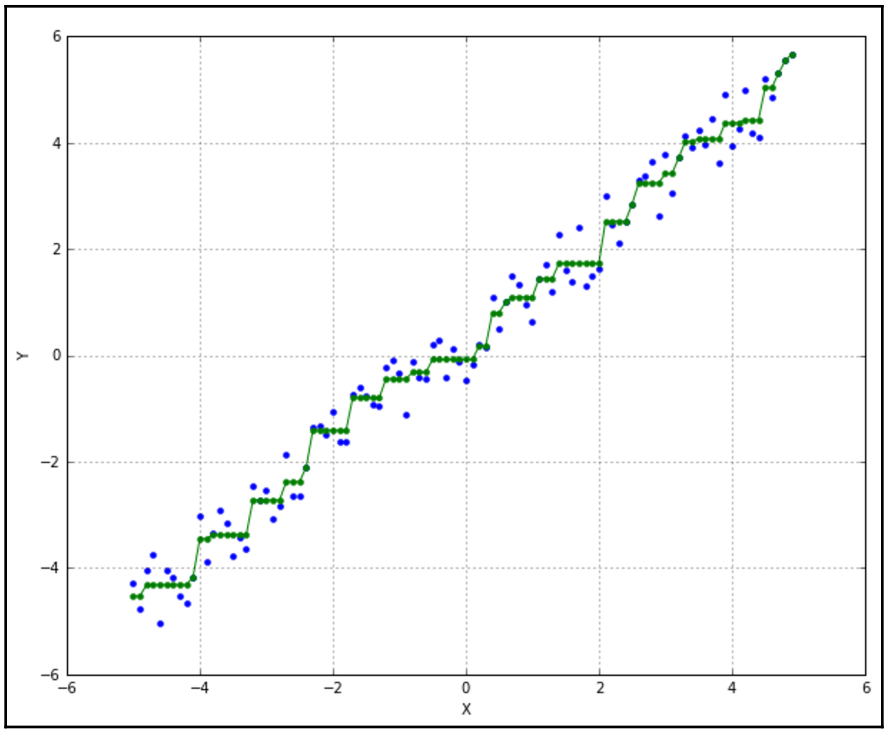

25
20
15
10
5
0
−5
Y
−6
−4
−2
0
2
4
6
X


15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
Y
−4
−2
0
2
4
X

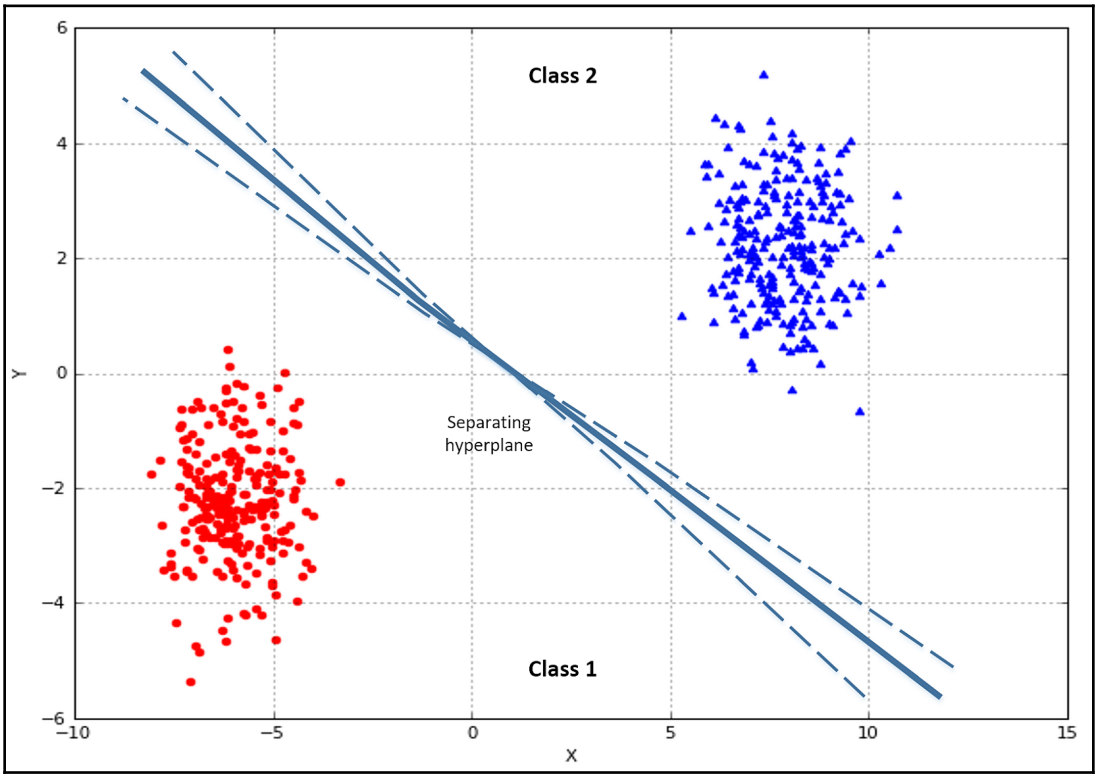

0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Parameter posterior: $\mathbf{p}(\boldsymbol{\theta}|\boldsymbol{\sigma}^2, \boldsymbol{X}, \boldsymbol{Y})$
Parameter prior: $\mathbf{p}(\boldsymbol{\theta}|\boldsymbol{\sigma}^2)$
−20
−15
−10
−5
0
5
10
15
20

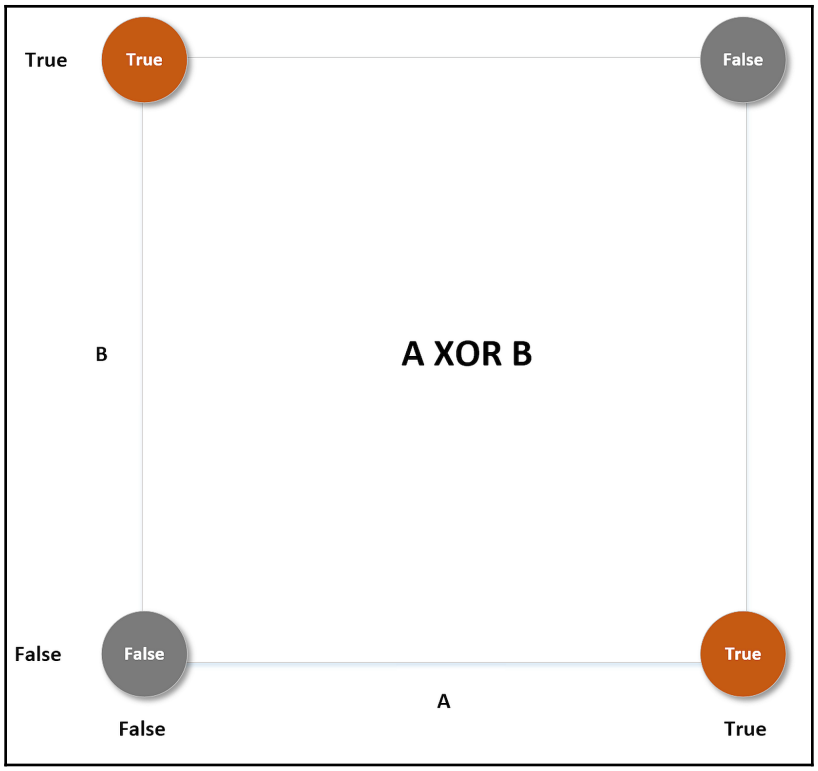

35
30
25
20
15
10
5
0
−5
Y
−6
−4
−2
0
2
4
6
X

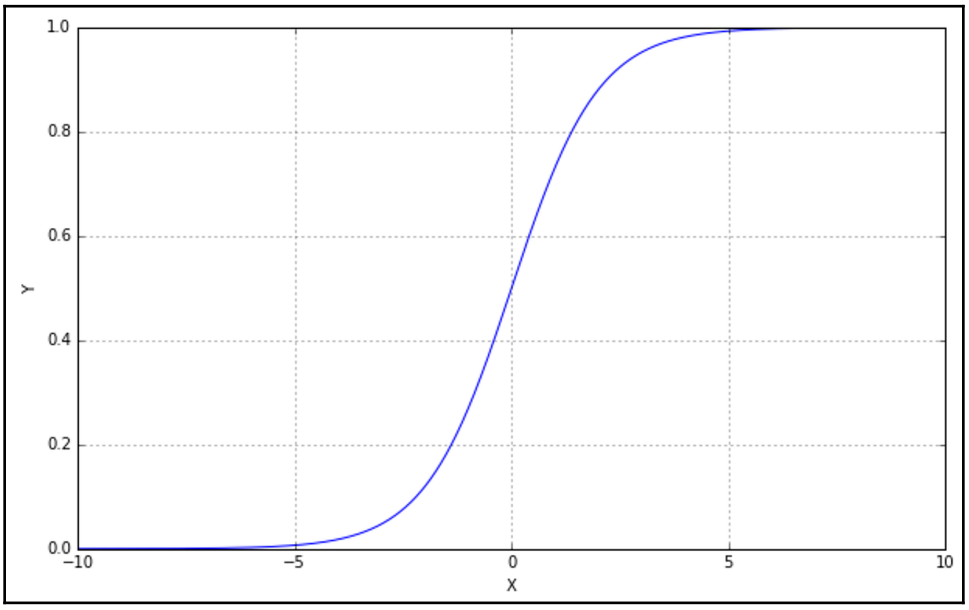

6
4
2
0
−2
−4
−6
Y
−6
−4
−2
0
2
4
6
X

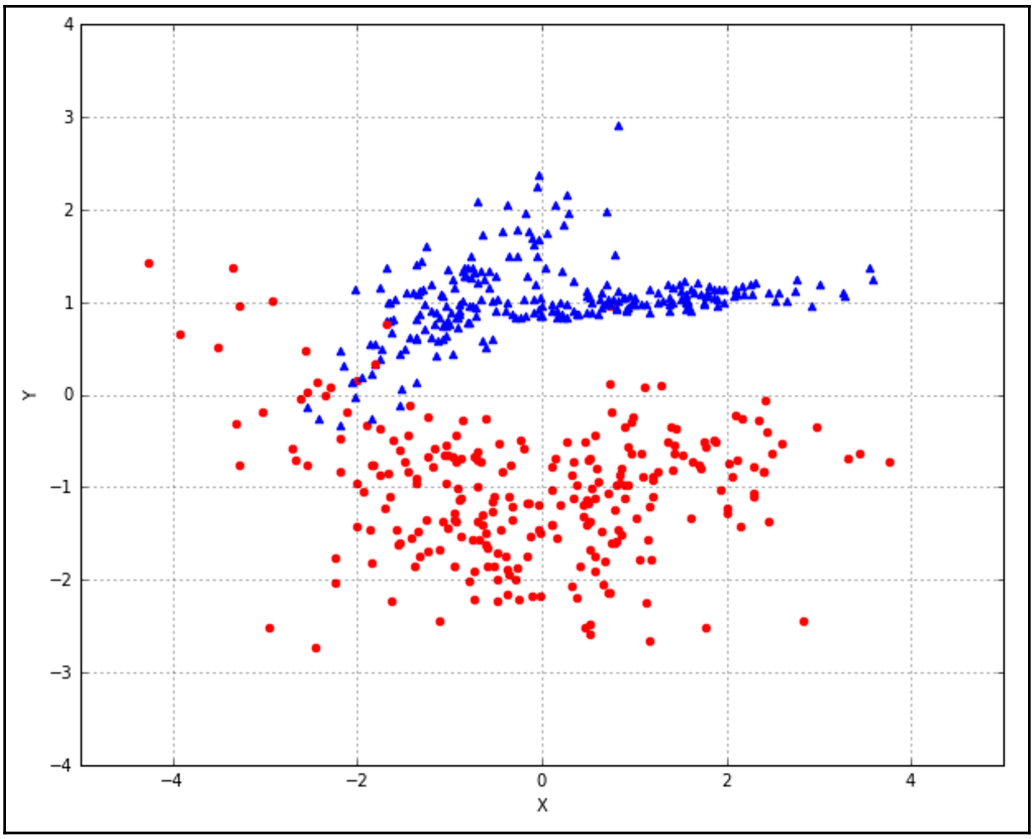

6
4
2
0
−2
−4
−6
Y
−6
−4
−2
0
2
4
6
X

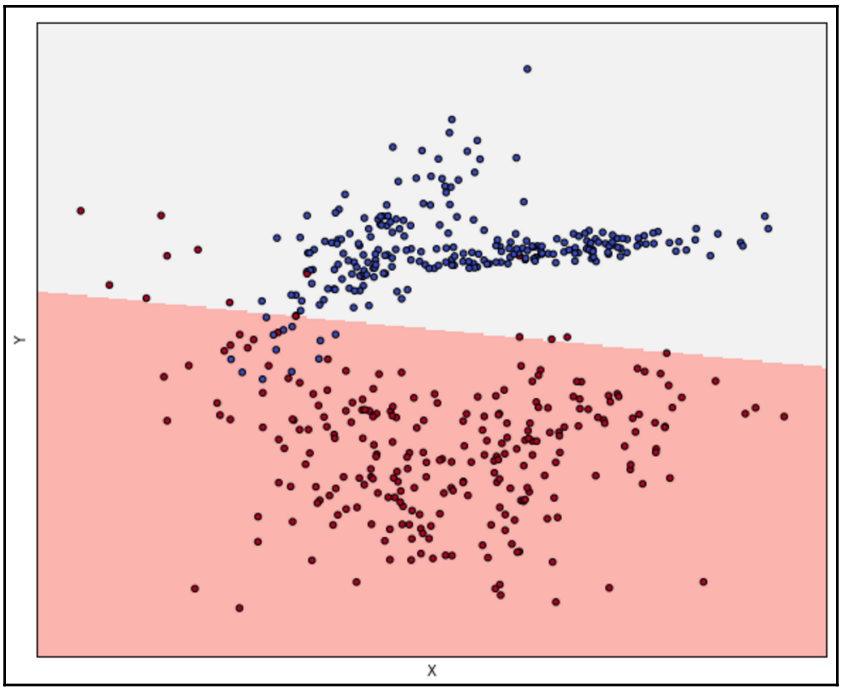
Chapter 05: Linear Classification Algorithms

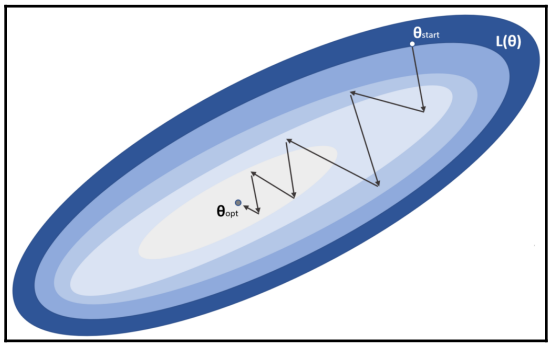


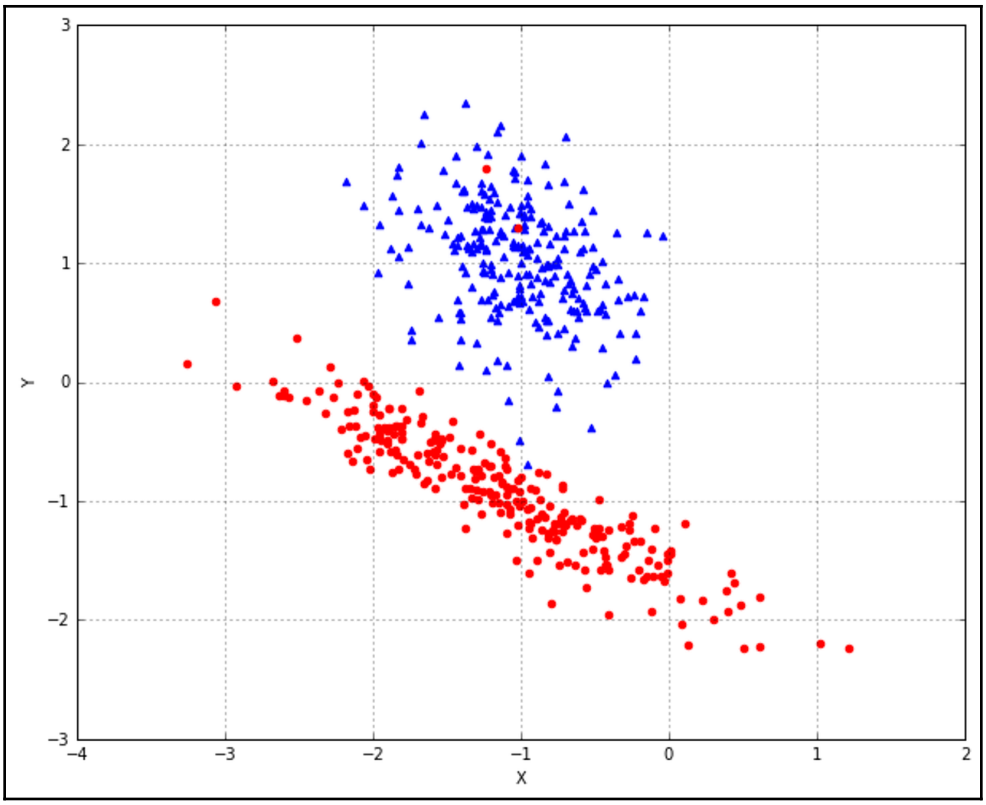

True
True
False
B
A XOR B
False
False
True
A
False
True

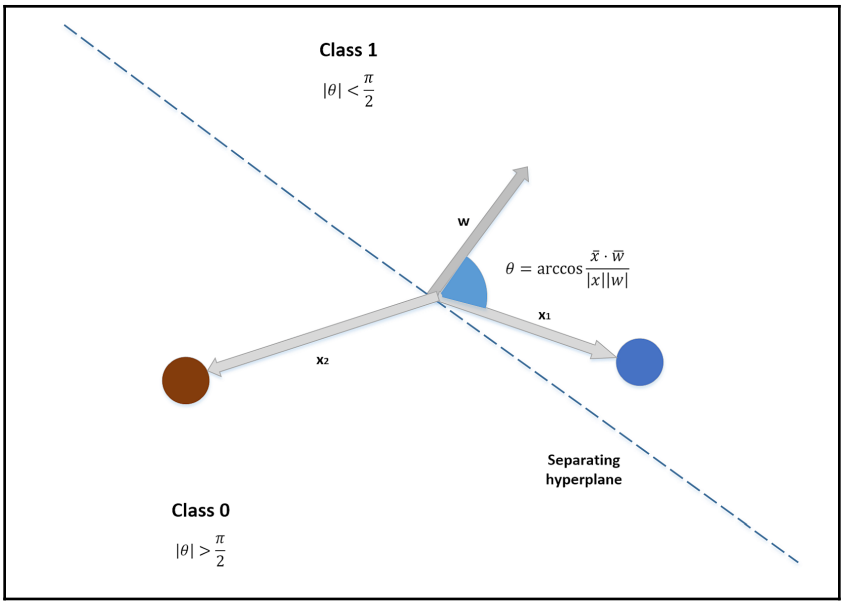

1.0
0.8
0.6
0.4
0.2
0.0
Y
−10
−5
0
5
10
X

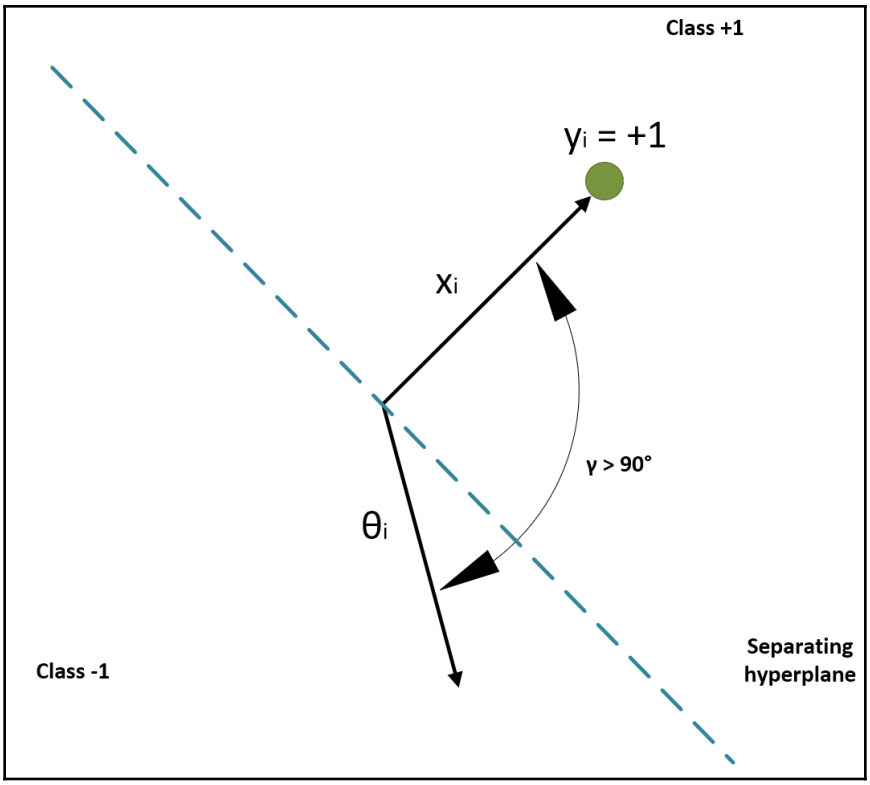

4
3
2
1
0
−1
−2
−3
−4
Y
−4
−2
0
2
4
X

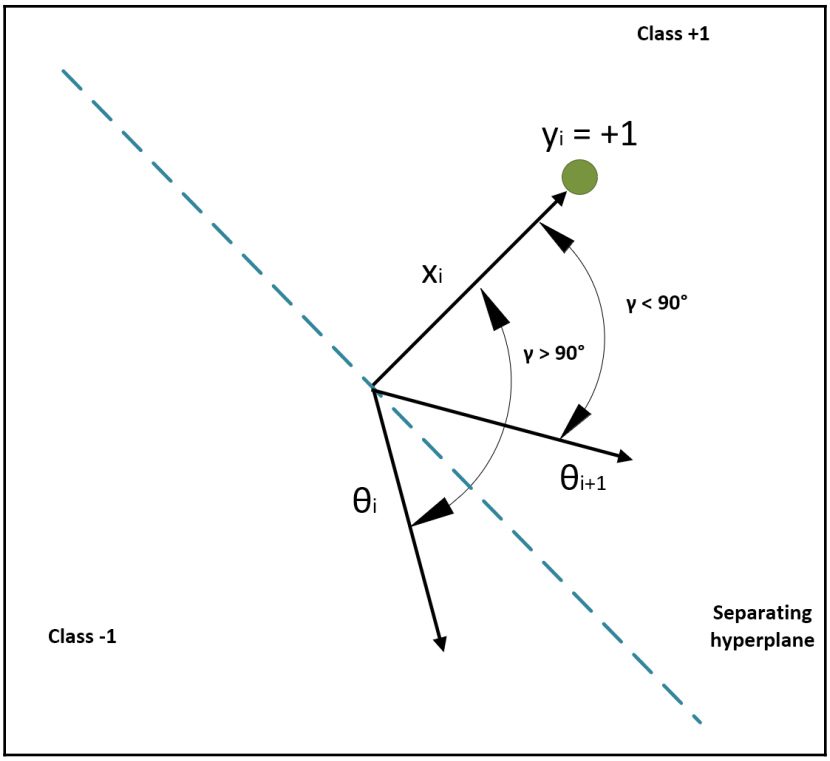

Y
X

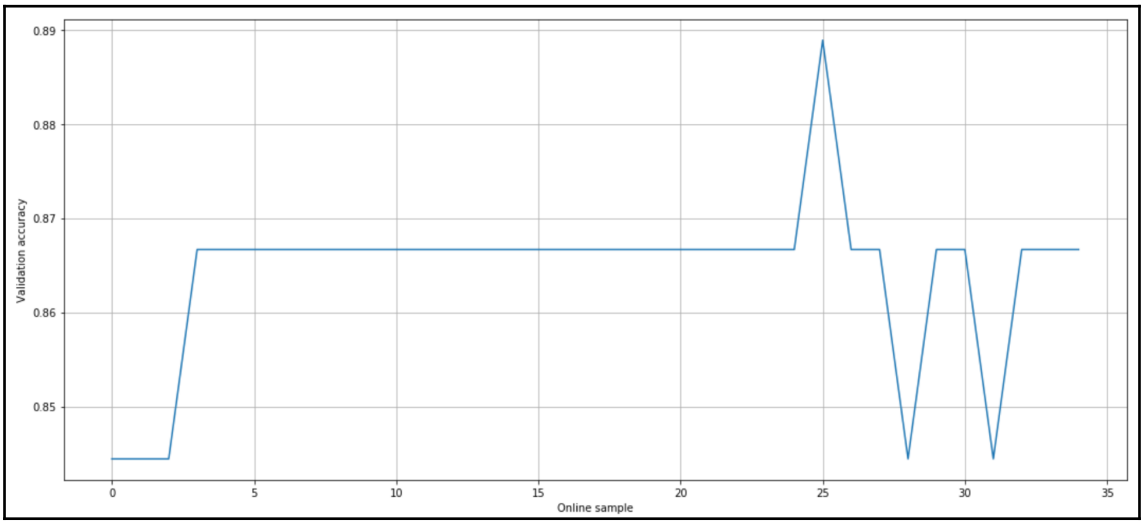

θstart
L(θ)
θopt

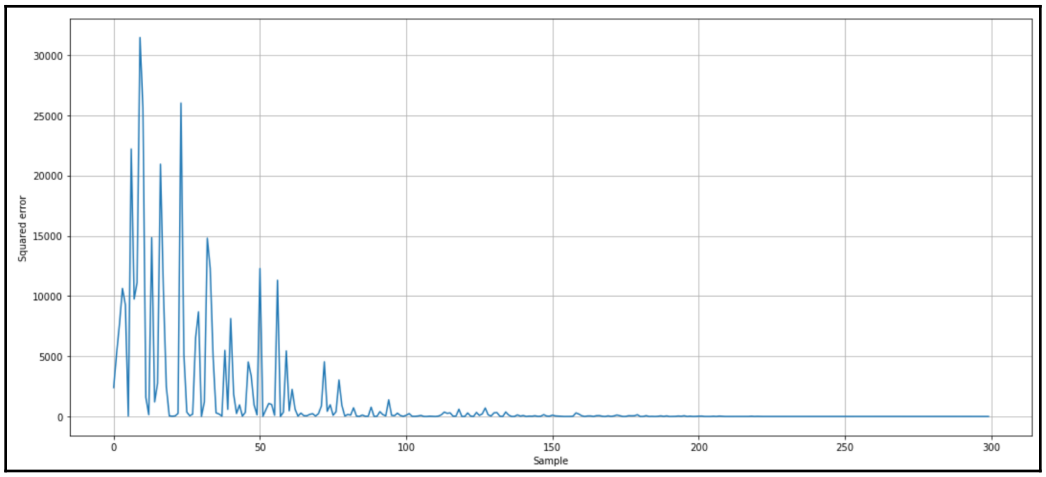

3
2
1
0
−1
−2
−3
Y
−4
−3
−2
−1
0
1
2
X

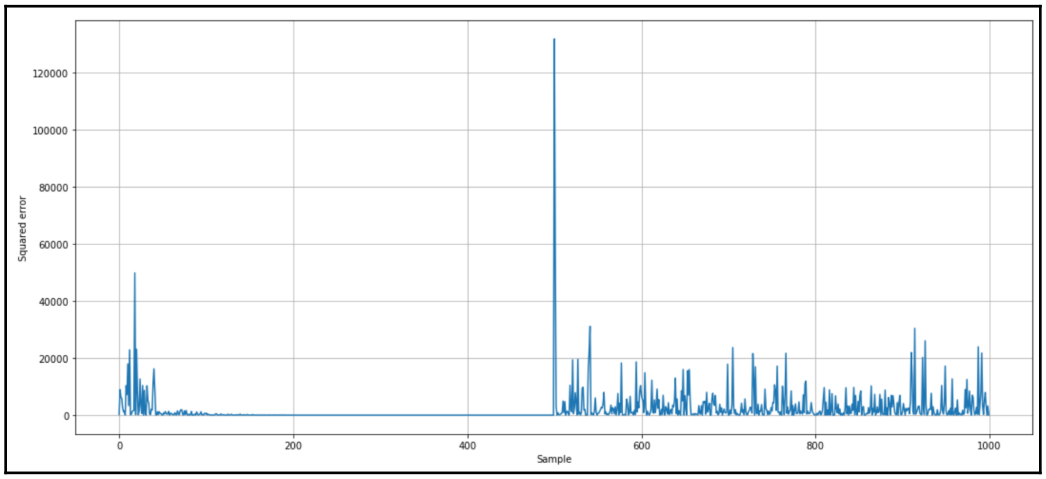

Class 1
$|\theta| < \frac{\pi}{2}$
w
$\theta = \arccos \frac{\vec{x} \cdot \vec{w}}{|x||w|}$
x1
x2
Separating hyperplane
Class 0
$|\theta| > \frac{\pi}{2}$

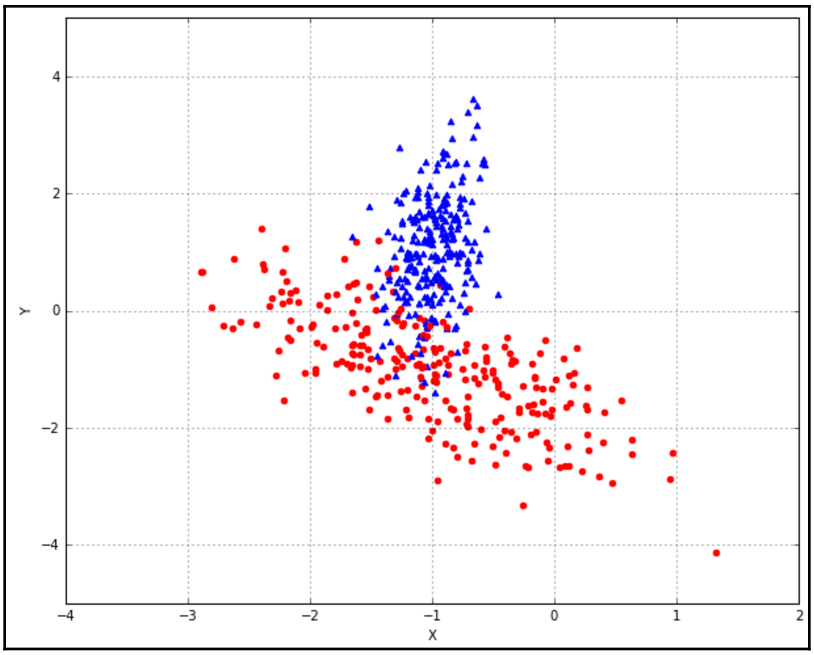

Class +1
$y_i = +1$
x_i
$\gamma > 90°$
θ_i
Class -1
Separating hyperplane


Class +1
$y_i = +1$
x_i
γ < 90°
γ > 90°
θ_{i+1}
θ_i
Class -1
Separating hyperplane

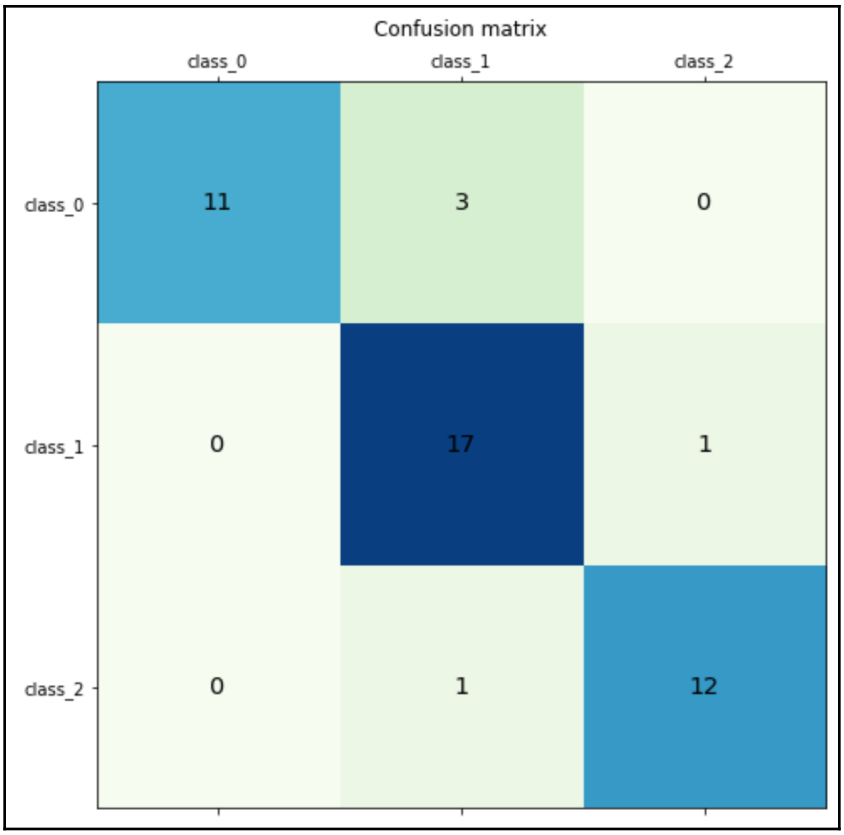

Validation accuracy
0.89
0.88
0.87
0.86
0.85
0
5
10
15
20
25
30
35
Online sample

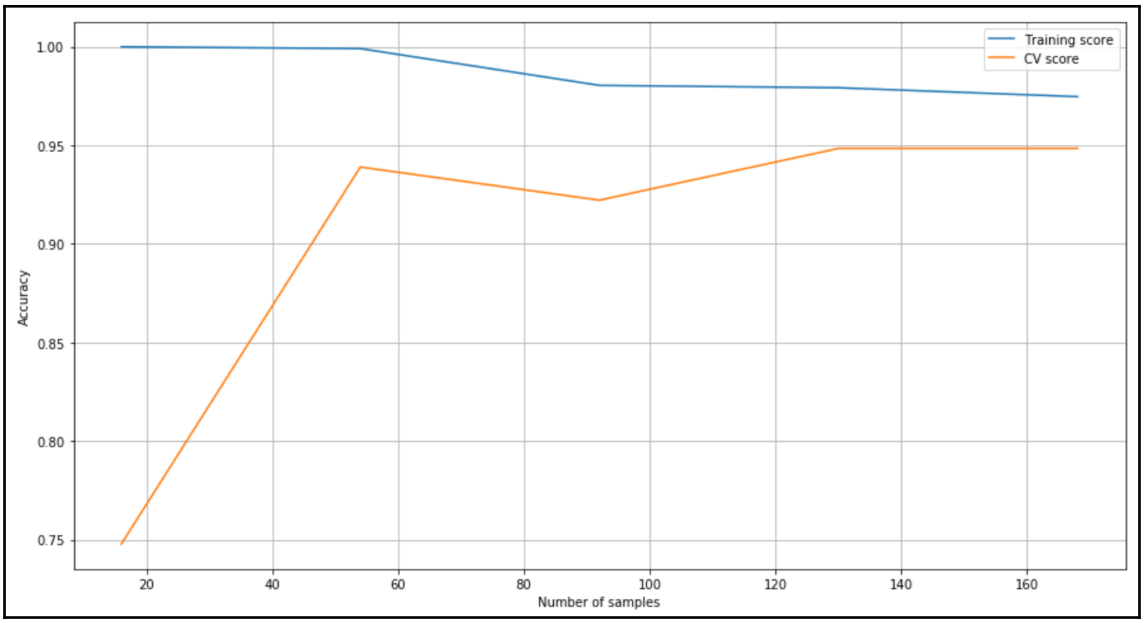

Squared error
Sample
0
5000
10000
15000
20000
25000
30000
0
50
100
150
200
250
300


Squared error
Sample
0
20000
40000
60000
80000
100000
120000
0
200
400
600
800
1000

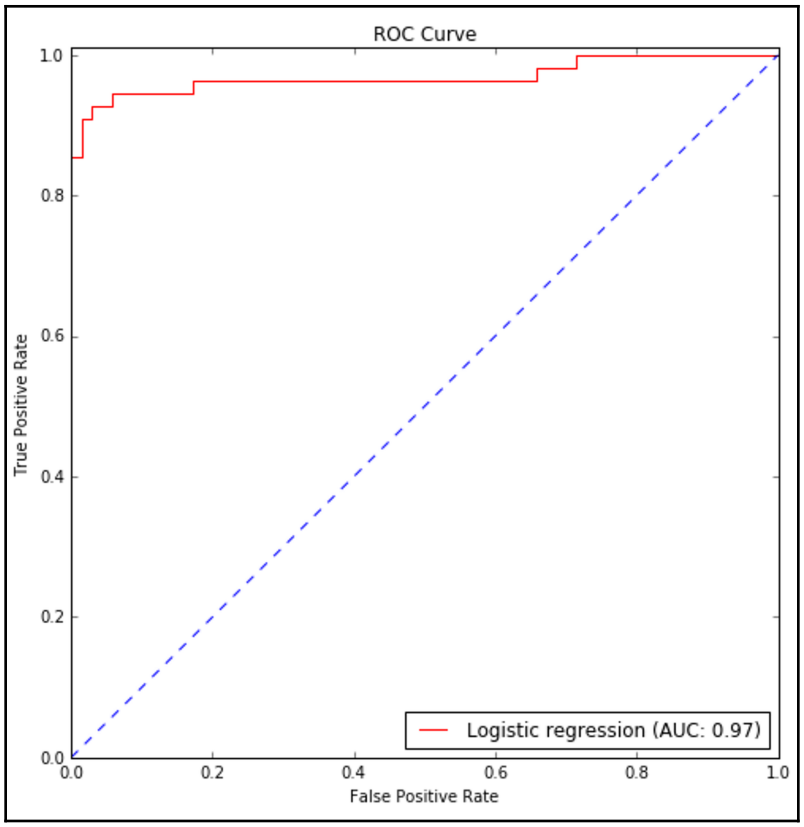


True Positives	False Negatives
False Positives	True Negatives

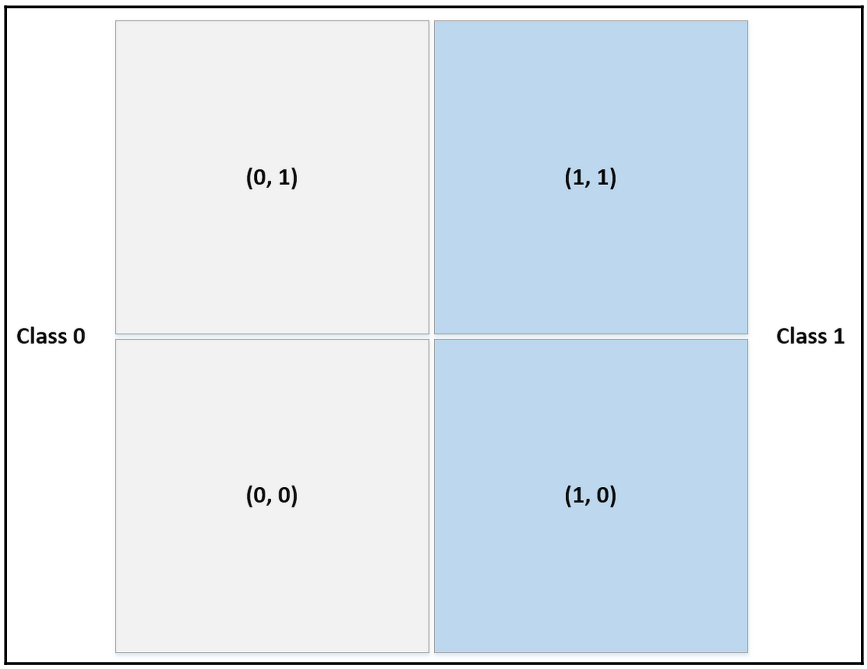

Confusion matrix
class_0
class_1
class_2
class_0
class_1
class_2
11
3
0
0
17
1
0
1
12

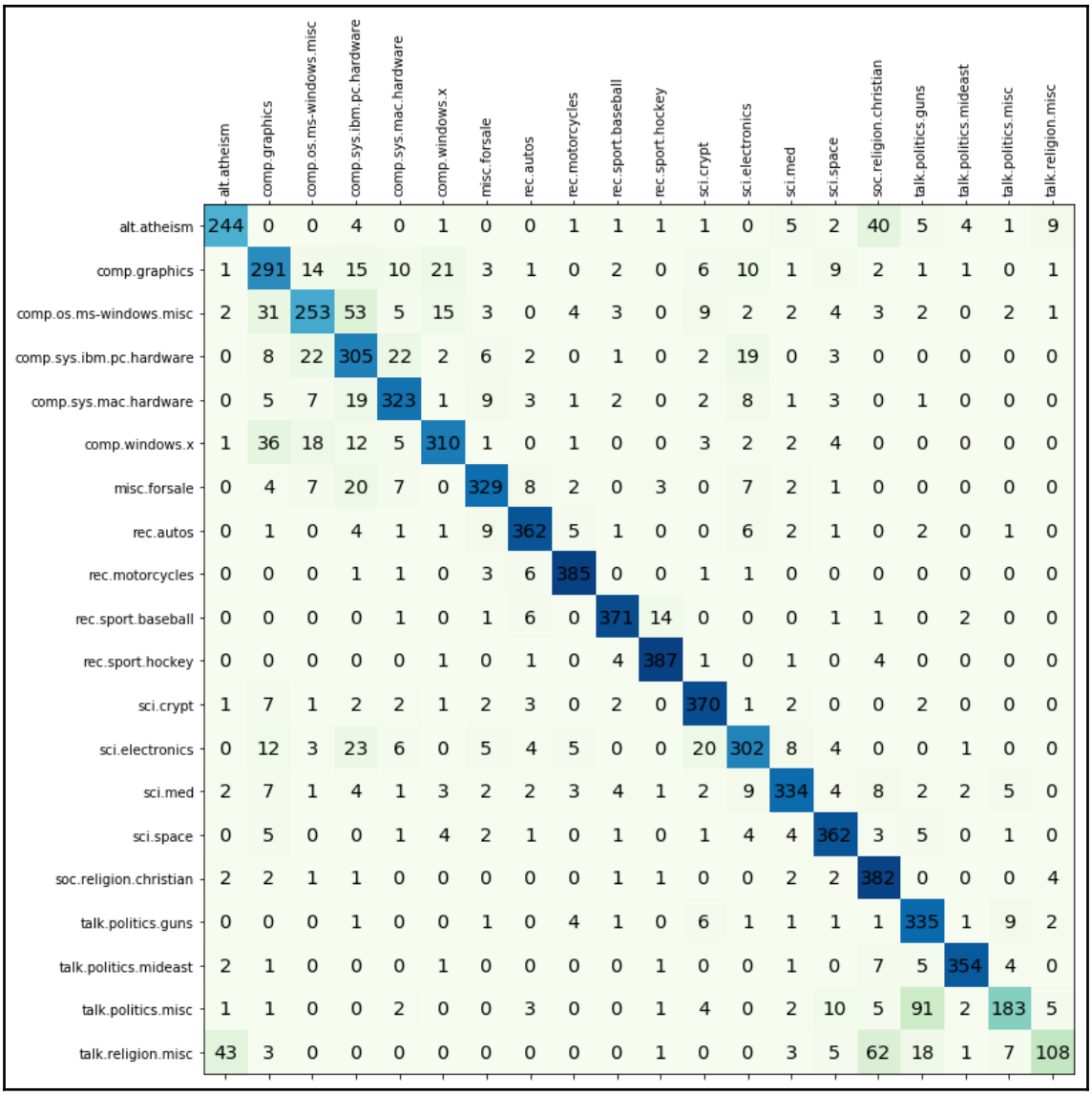

Training score
CV score
1.00
0.95
0.90
0.85
0.80
0.75
Accuracy
20
40
60
80
100
120
140
160
Number of samples

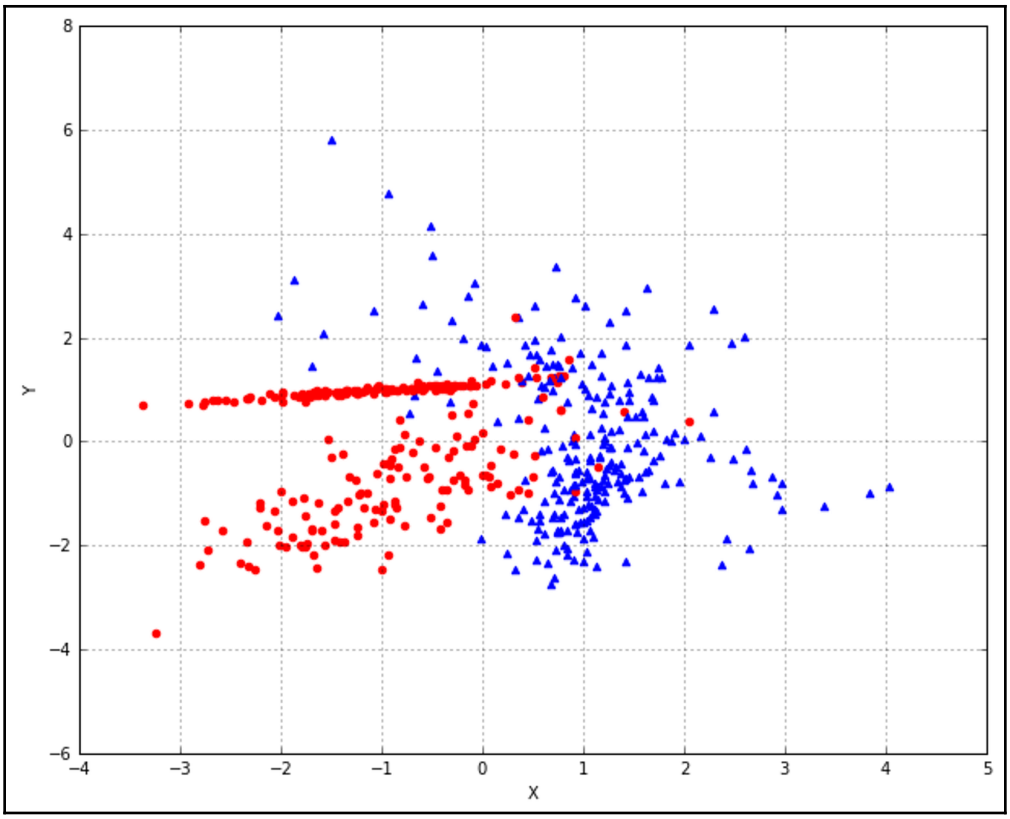

Better performances
Random performances
Worse performances
True positive rate (or Sensitivity)
0
0.5
1
0
0.5
1
False positive rate (or Specificity)

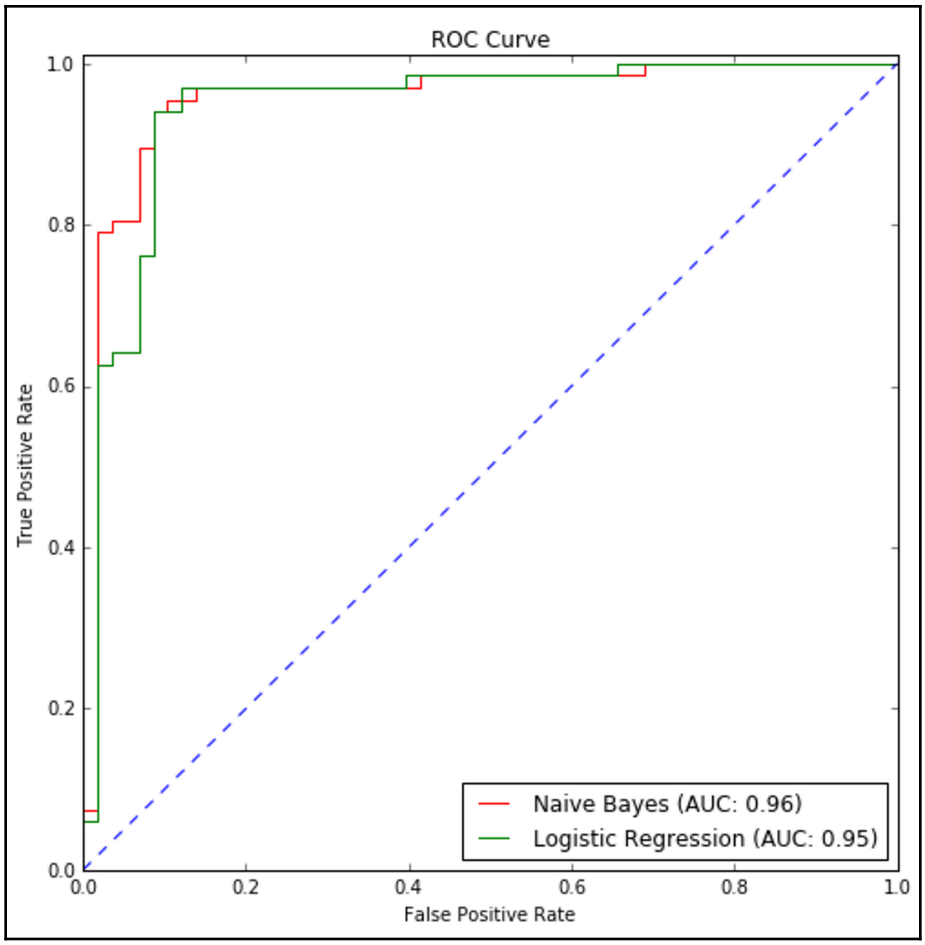

ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic regression (AUC: 0.97)

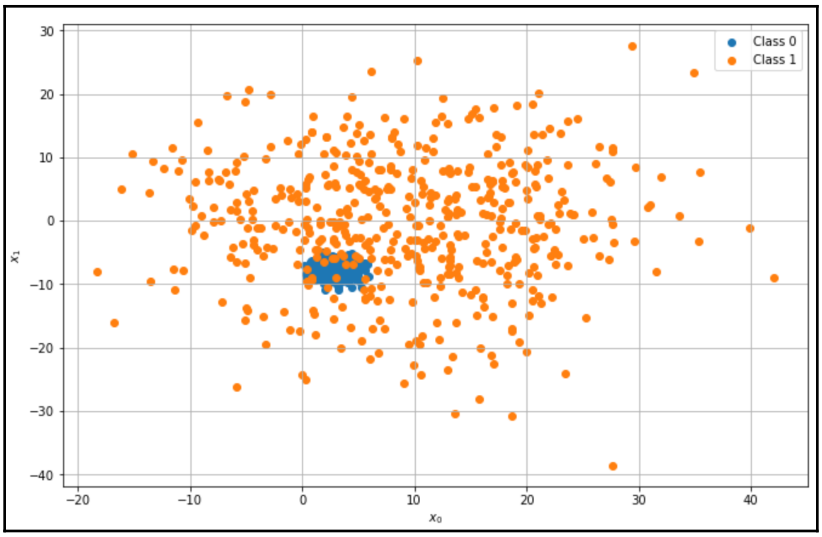
Chapter 06: Naive Bayes and Discriminant Analysis

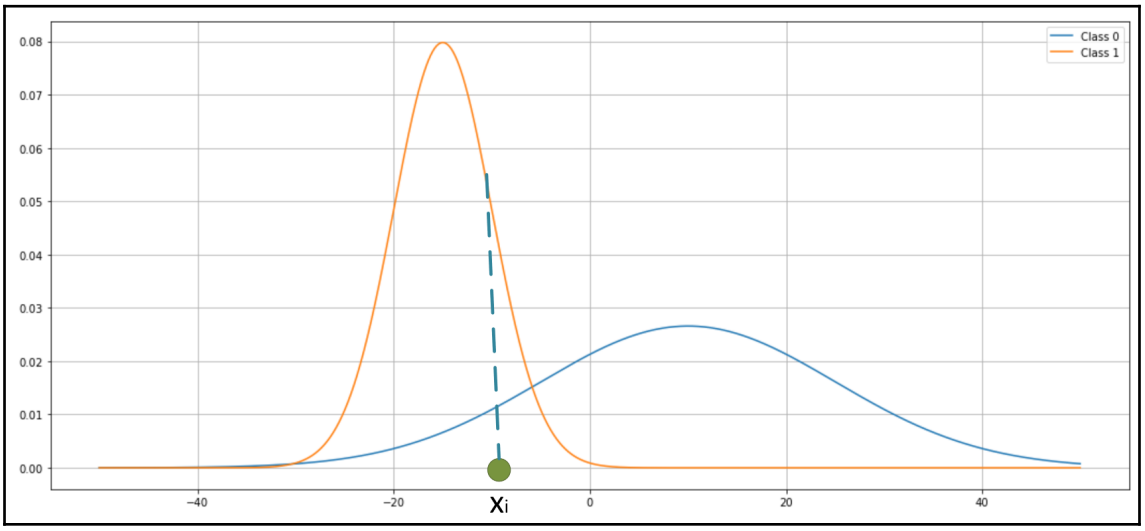



alt.atheism
comp.graphics
comp.os.ms-windows.misc
comp.sys.ibm.pc.hardware
comp.sys.mac.hardware
comp.windows.x
misc.forsale
rec.autos
rec.motorcycles
rec.sport.baseball
rec.sport.hockey
sci.crypt
sci.electronics
sci.med
sci.space
soc.religion.christian
talk.politics.guns
talk.politics.mideast
talk.politics.misc
talk.religion.misc
alt.atheism 244 0 0 4 0 1 0 0 1 1 1 1 0 5 2 40 5 4 1 9
comp.graphics 1 291 14 15 10 21 3 1 0 2 0 6 10 1 9 2 1 1 0 1
comp.os.ms-windows.misc 2 31 253 53 5 15 3 0 4 3 0 9 2 2 4 3 2 0 2 1
comp.sys.ibm.pc.hardware 0 8 22 305 22 2 6 2 0 1 0 2 19 0 3 0 0 0 0 0
comp.sys.mac.hardware 0 5 7 19 323 1 9 3 1 2 0 2 8 1 3 0 1 0 0 0
comp.windows.x 1 36 18 12 5 310 1 0 1 0 0 3 2 2 4 0 0 0 0 0
misc.forsale 0 4 7 20 7 0 329 8 2 0 3 0 7 2 1 0 0 0 0 0
rec.autos 0 1 0 4 1 1 9 362 5 1 0 0 6 2 1 0 2 0 1 0
rec.motorcycles 0 0 0 1 1 0 3 6 385 0 0 1 1 0 0 0 0 0 0 0
rec.sport.baseball 0 0 0 0 1 0 1 6 0 371 14 0 0 0 1 1 0 2 0 0
rec.sport.hockey 0 0 0 0 0 1 0 1 0 4 387 1 0 1 0 4 0 0 0 0
sci.crypt 1 7 1 2 2 1 2 3 0 2 0 370 1 2 0 0 2 0 0 0
sci.electronics 0 12 3 23 6 0 5 4 5 0 0 20 302 8 4 0 0 1 0 0
sci.med 2 7 1 4 1 3 2 2 3 4 1 2 9 334 4 8 2 2 5 0
sci.space 0 5 0 0 1 4 2 1 0 1 0 1 4 4 362 3 5 0 1 0
soc.religion.christian 2 2 1 1 0 0 0 0 0 1 1 0 0 2 2 382 0 0 0 4
talk.politics.guns 0 0 0 1 0 0 1 0 4 1 0 6 1 1 1 1 335 1 9 2
talk.politics.mideast 2 1 0 0 0 1 0 0 0 0 1 0 0 1 0 7 5 354 4 0
talk.politics.misc 1 1 0 0 2 0 0 3 0 0 1 4 0 2 10 5 91 2 183 5
talk.religion.misc 43 3 0 0 0 0 0 0 0 0 1 0 0 3 5 62 18 1 7 108

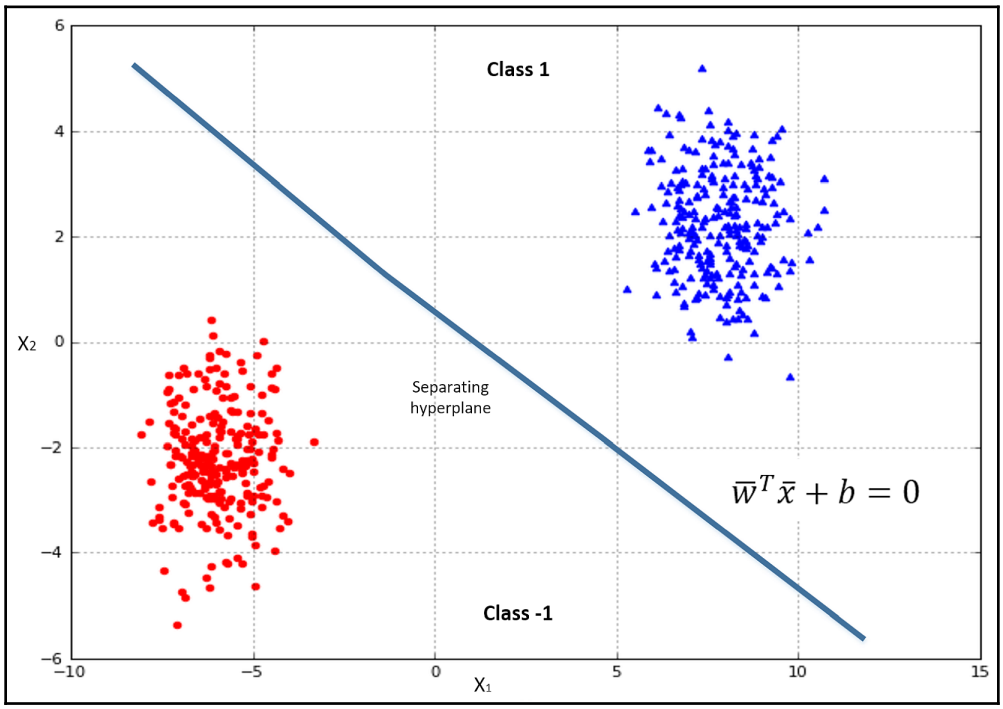

8
6
4
2
0
−2
−4
−6
Y
−4
−3
−2
−1
0
1
2
3
4
5
X

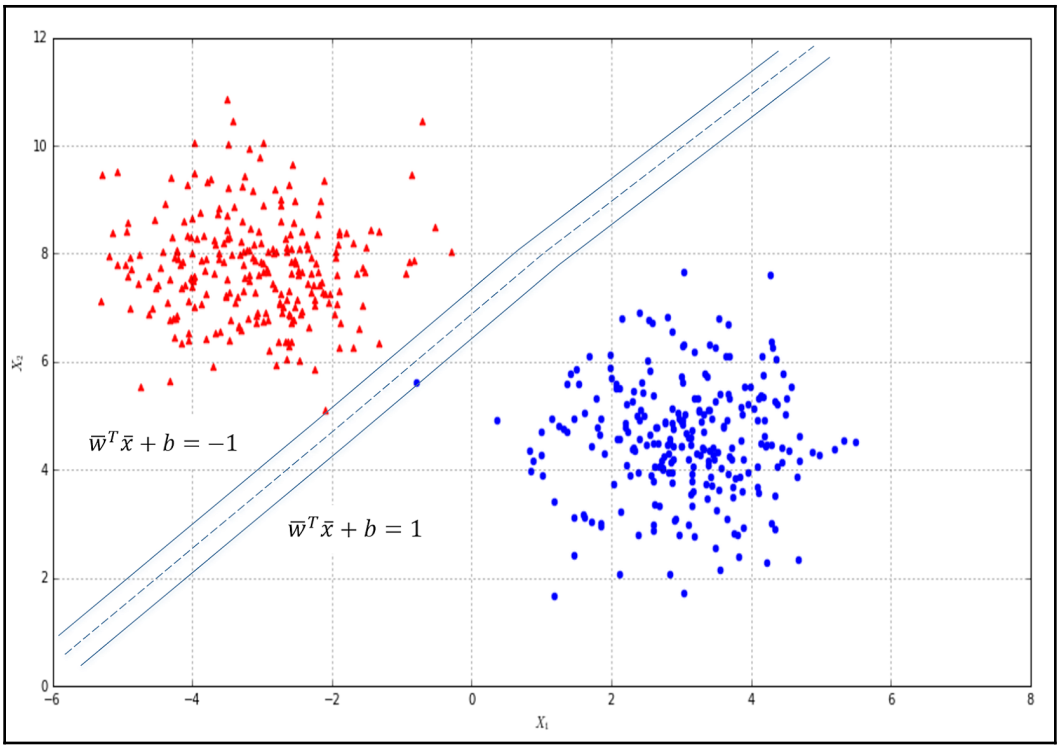

ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Naive Bayes (AUC: 0.96)
Logistic Regression (AUC: 0.95)

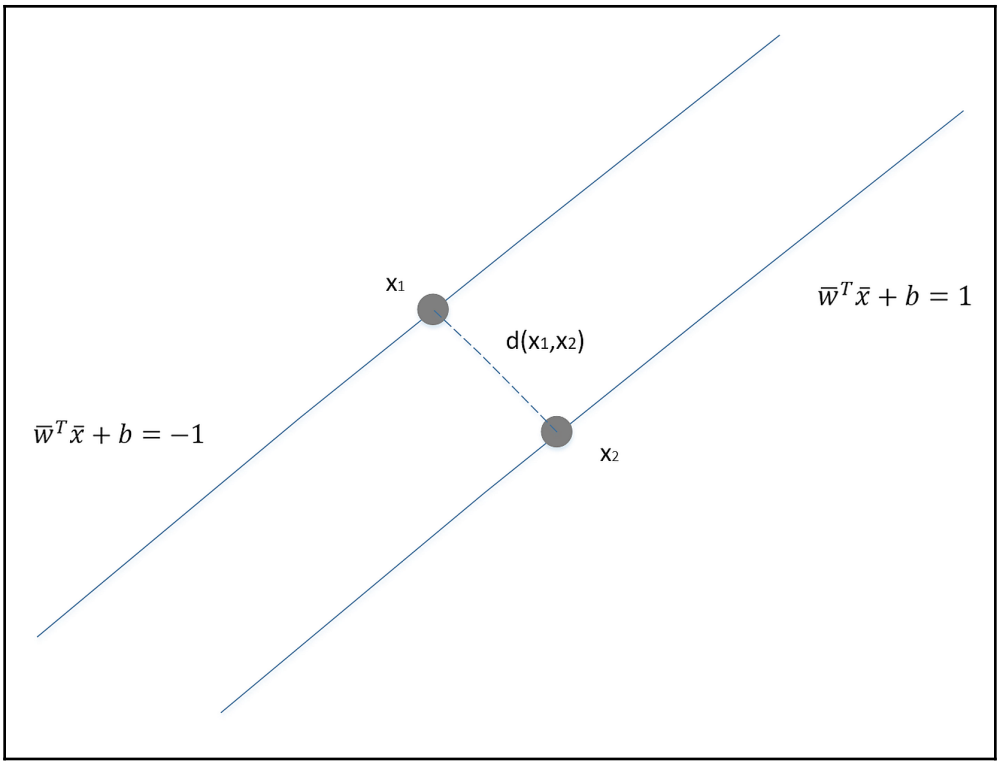

Class 0
Class 1
30
20
10
0
−10
−20
−30
−40
−20
−10
0
10
20
30
40
x_1
x_0

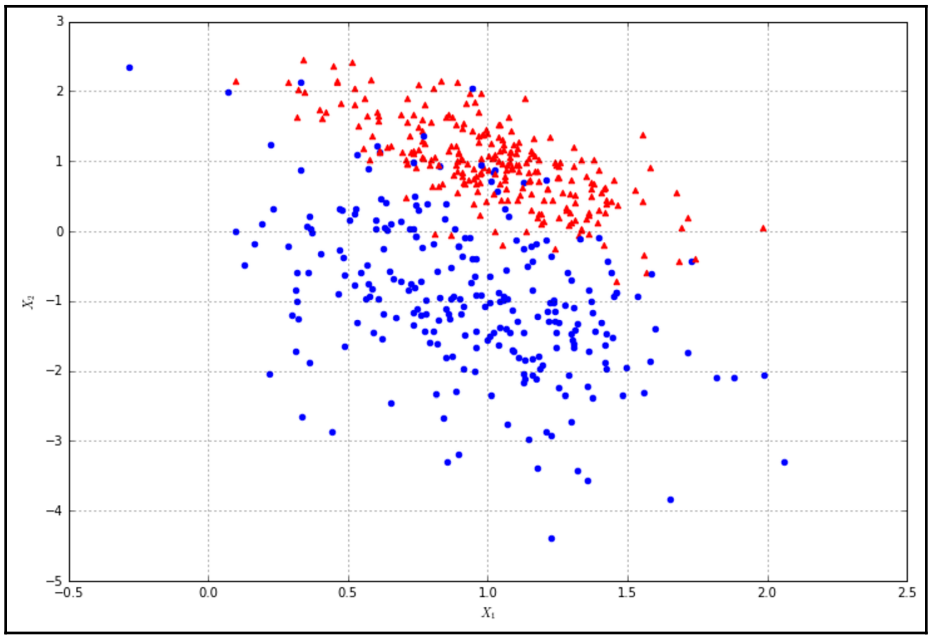

Class 0
Class 1
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
−40
−20
0
20
40
Xi

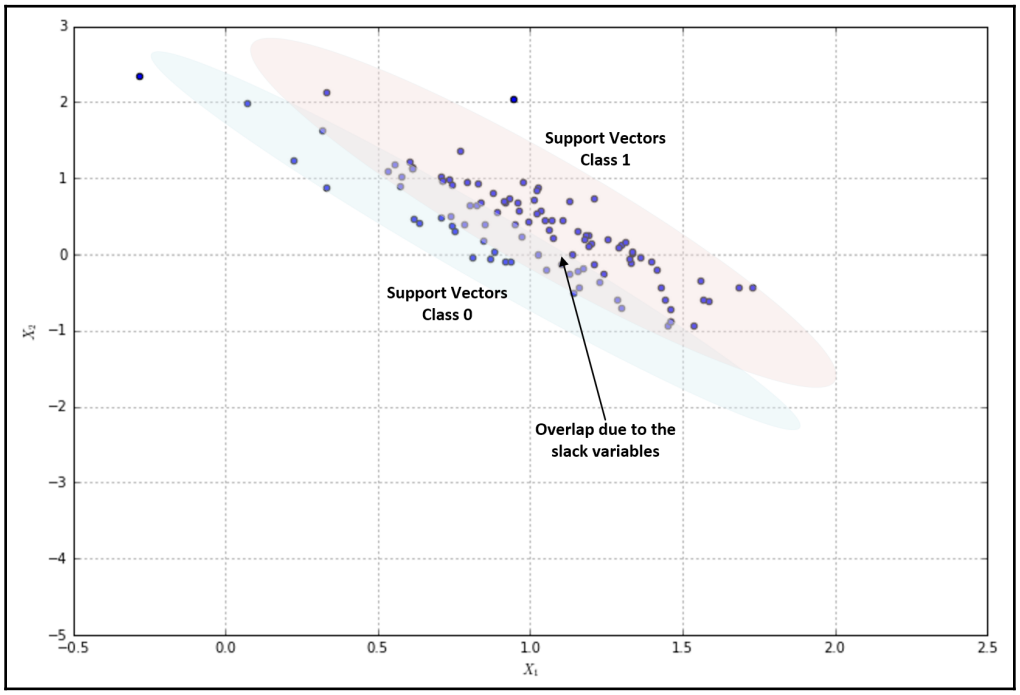
Chapter 07: Support Vector Machines

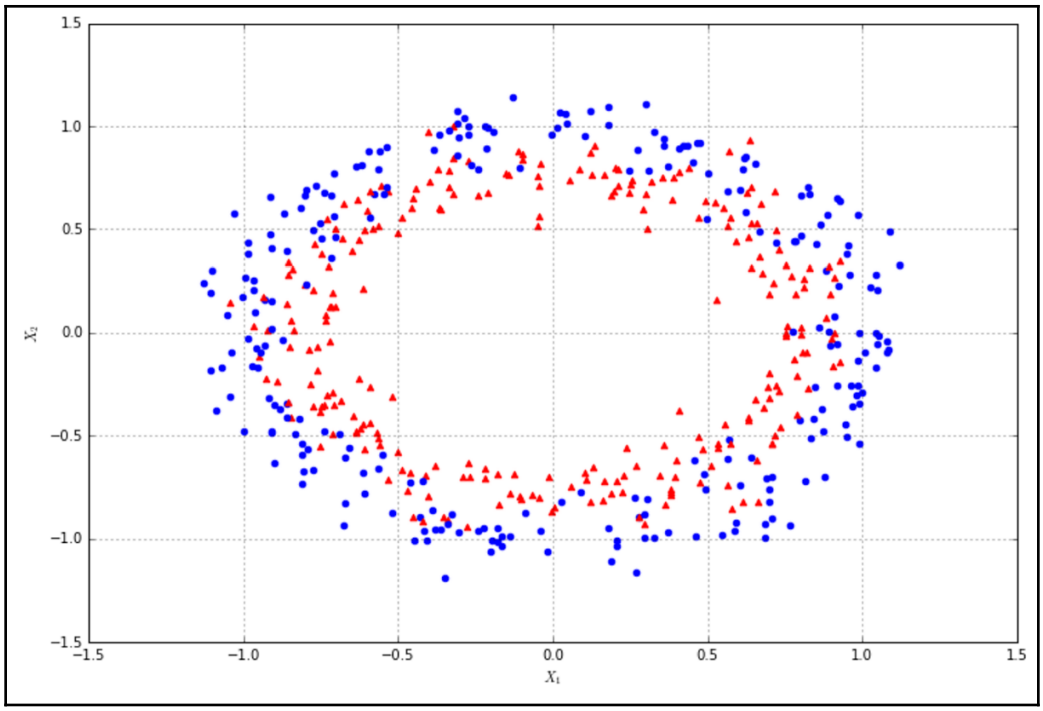


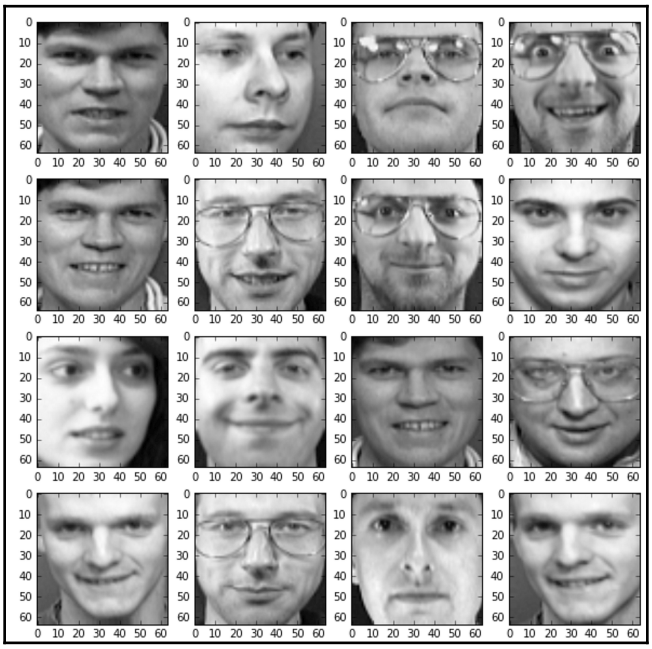

$\overline{w}^T\bar{x} + b = -1$
$\overline{w}^T\bar{x} + b = 1$
X_1
X_2

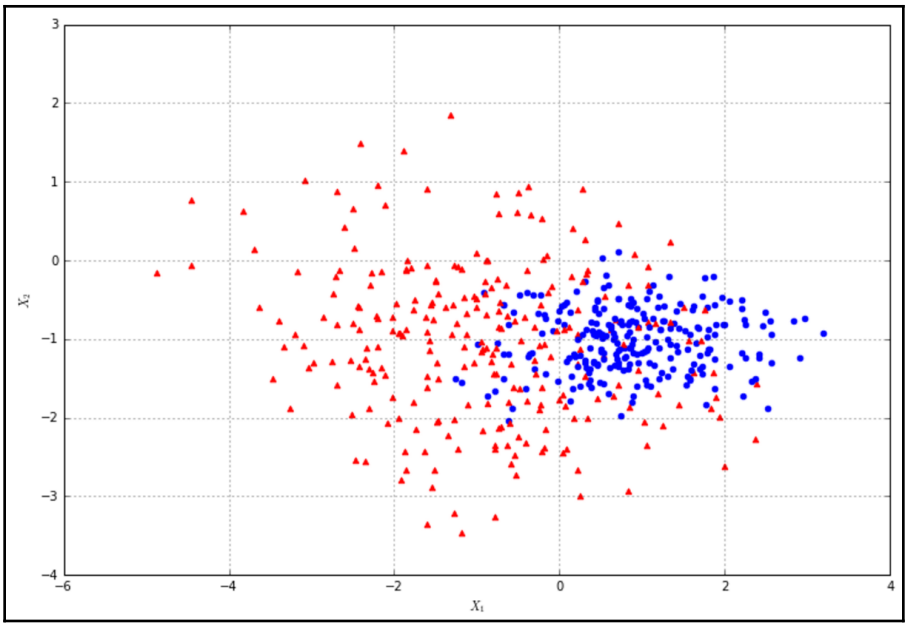

x1
d(x1,x2)
$\bar{w}^T\bar{x} + b = 1$
$\bar{w}^T\bar{x} + b = -1$
x2

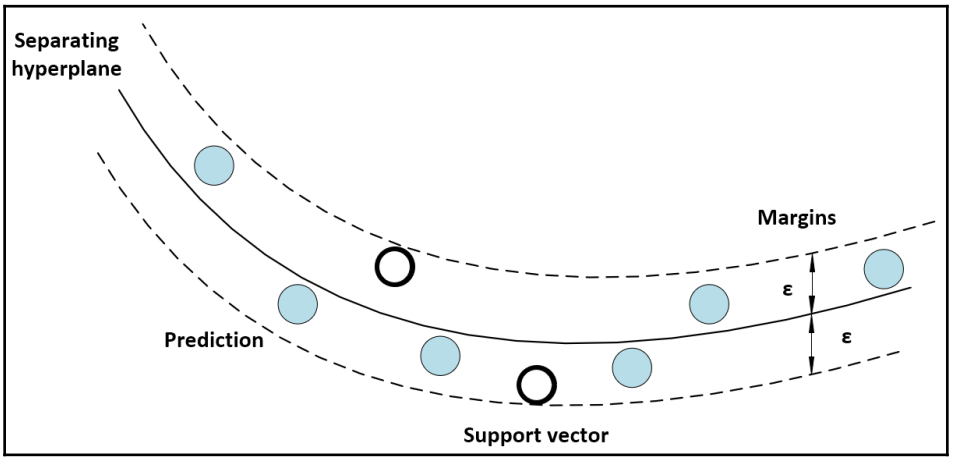

3
2
1
0
−1
−2
−3
−4
−5
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
X_2
X_1

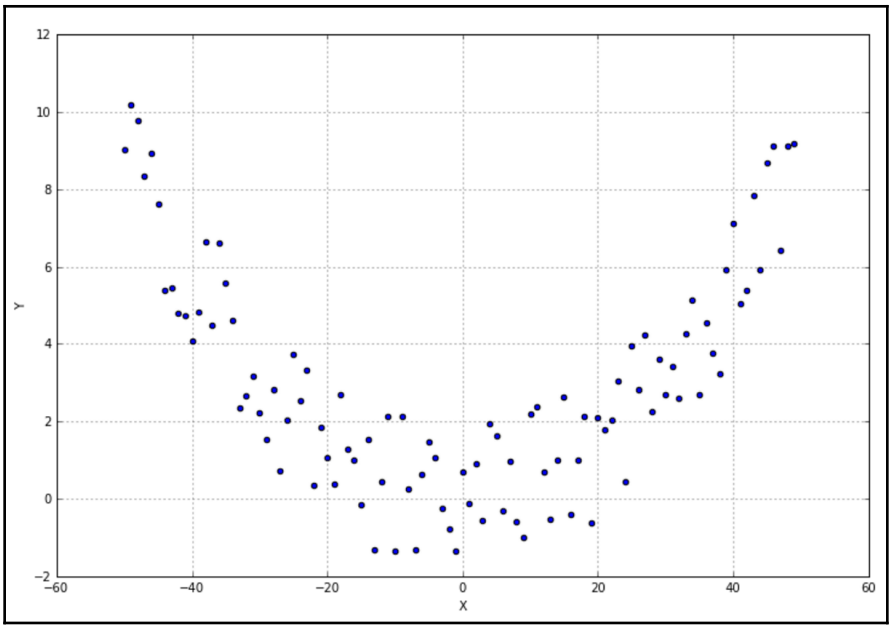

Support Vectors
Class 1
Support Vectors
Class 0
Overlap due to the
slack variables
3
2
1
0
−1
−2
−3
−4
−5
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
X_2
X_1

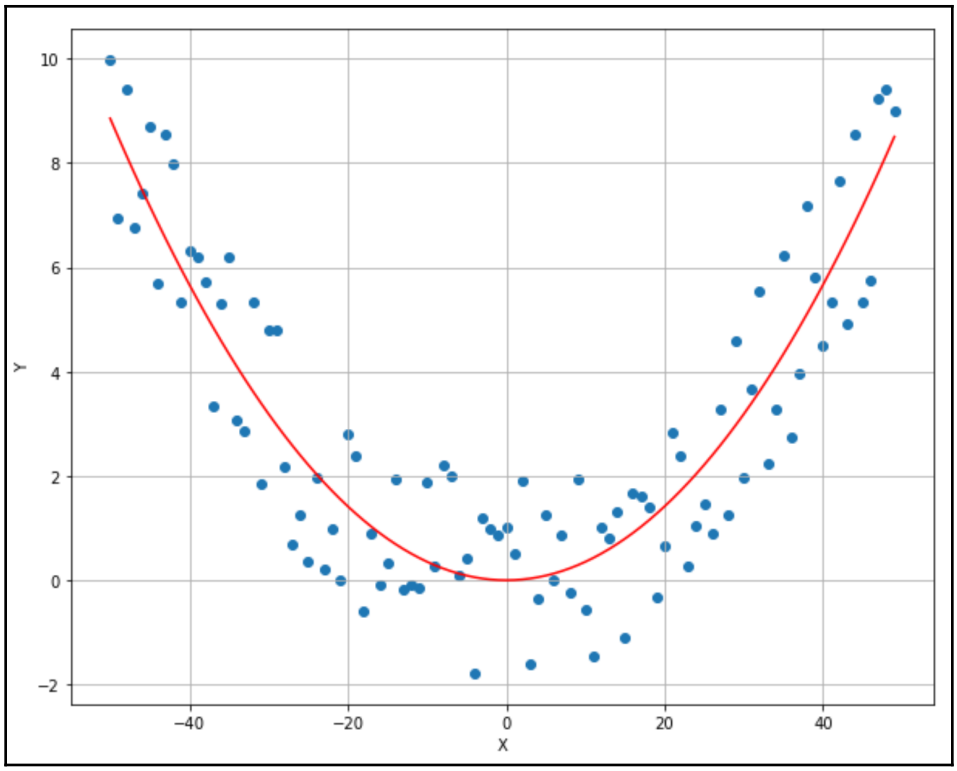

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
X_2
X_1



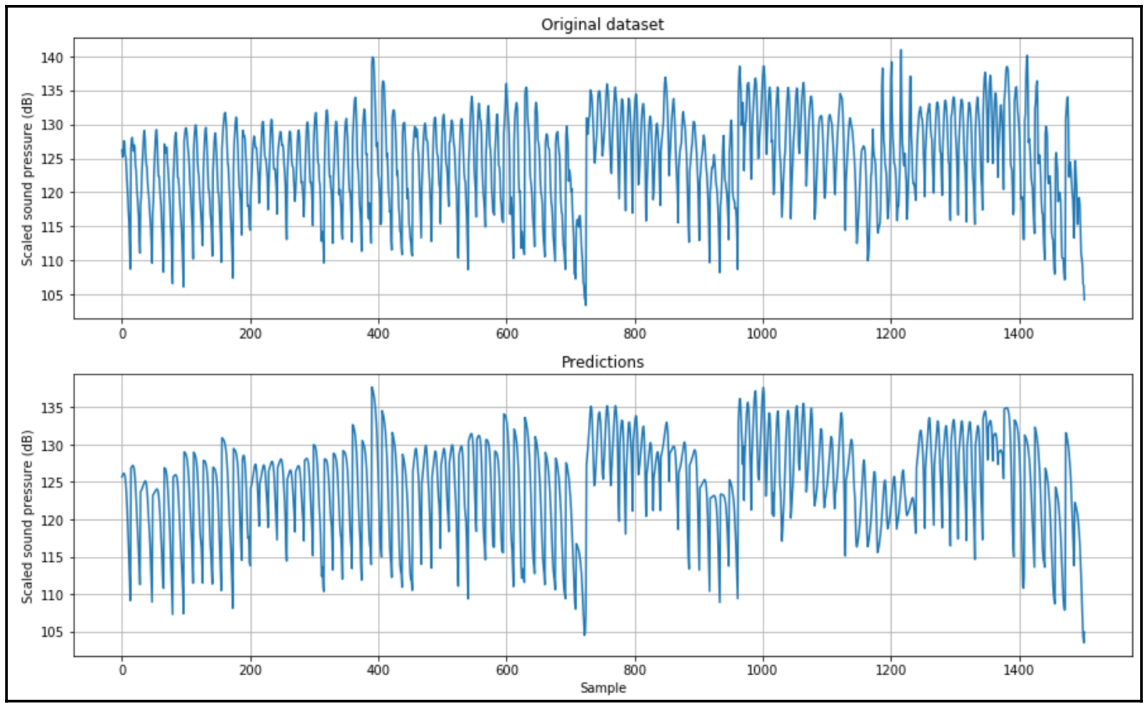

3
2
1
0
−1
−2
−3
−4
−6
−4
−2
0
2
4
X_2
X_1

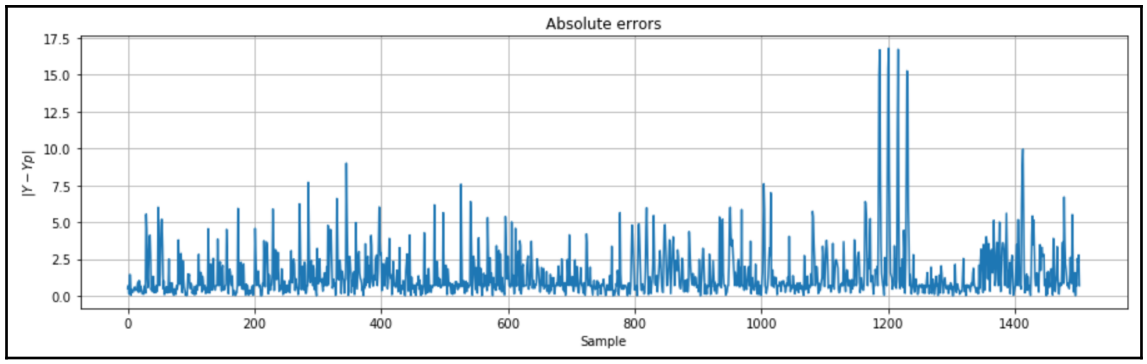

Separating hyperplane
Margins
ε
ε
Prediction
Support vector

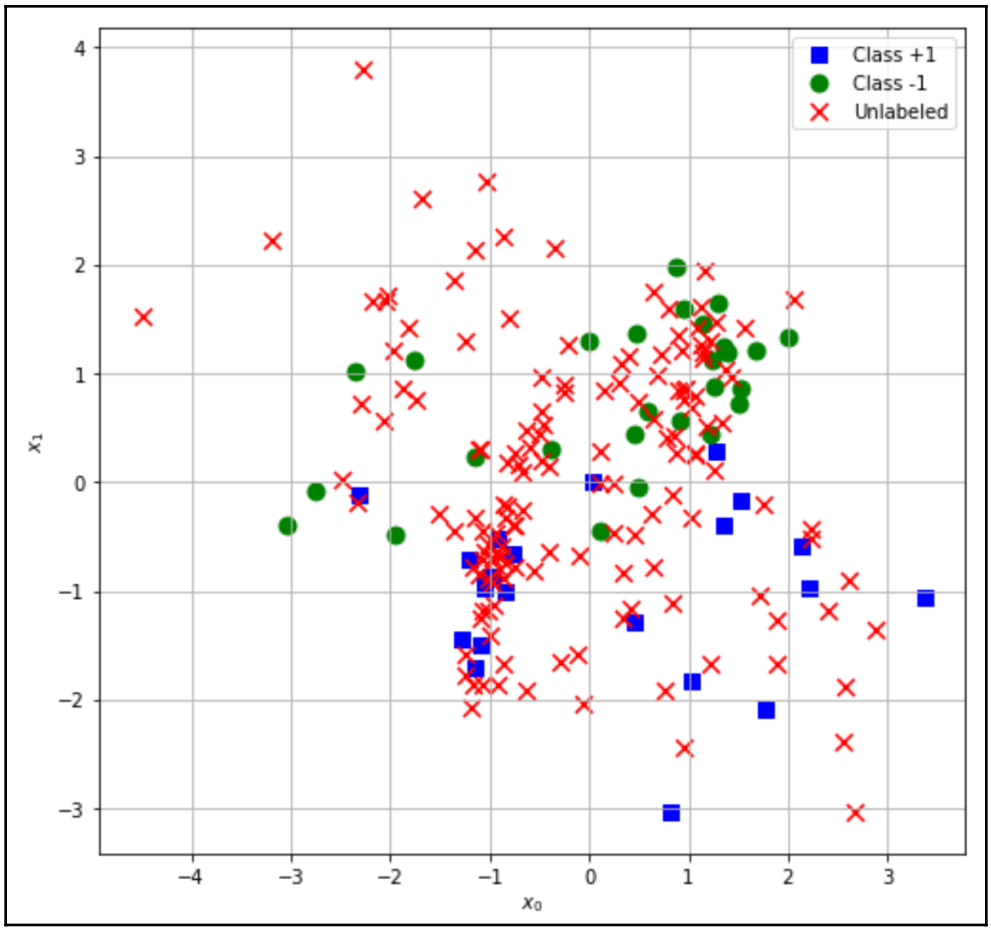

12
10
8
6
4
2
0
−2
Y
−60
−40
−20
0
20
40
60
X

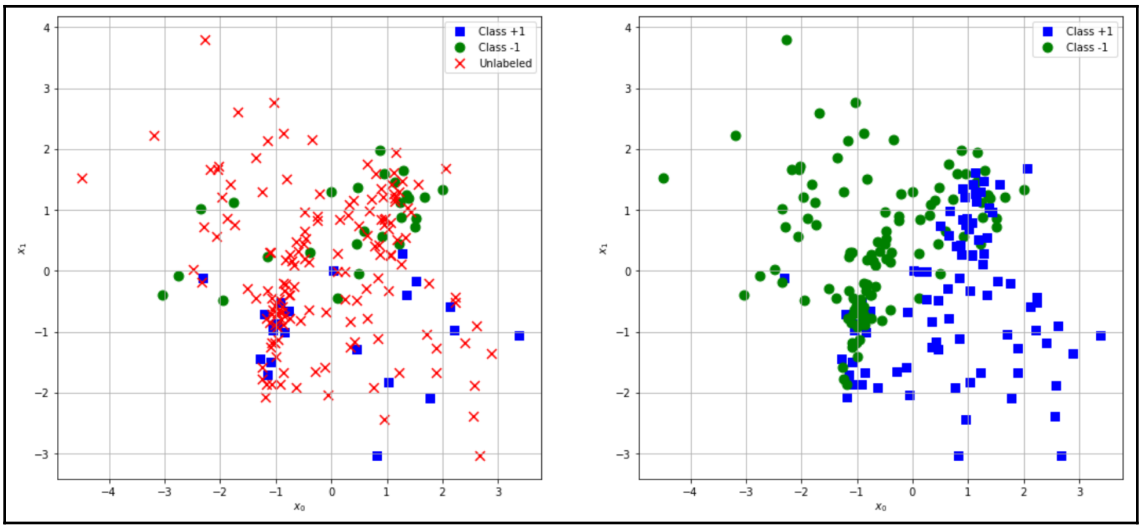


	0	1	2	3	4	5
count	1503.000000	1503.000000	1503.000000	1503.000000	1503.000000	1503.000000
mean	2886.380572	6.782302	0.136548	50.860745	0.011140	124.835943
std	3152.573137	5.918128	0.093541	15.572784	0.013150	6.898657
min	200.000000	0.000000	0.025400	31.700000	0.000401	103.380000
25%	800.000000	2.000000	0.050800	39.600000	0.002535	120.191000
50%	1600.000000	5.400000	0.101600	39.600000	0.004957	125.721000
75%	4000.000000	9.900000	0.228600	71.300000	0.015576	129.995500
max	20000.000000	22.200000	0.304800	71.300000	0.058411	140.987000

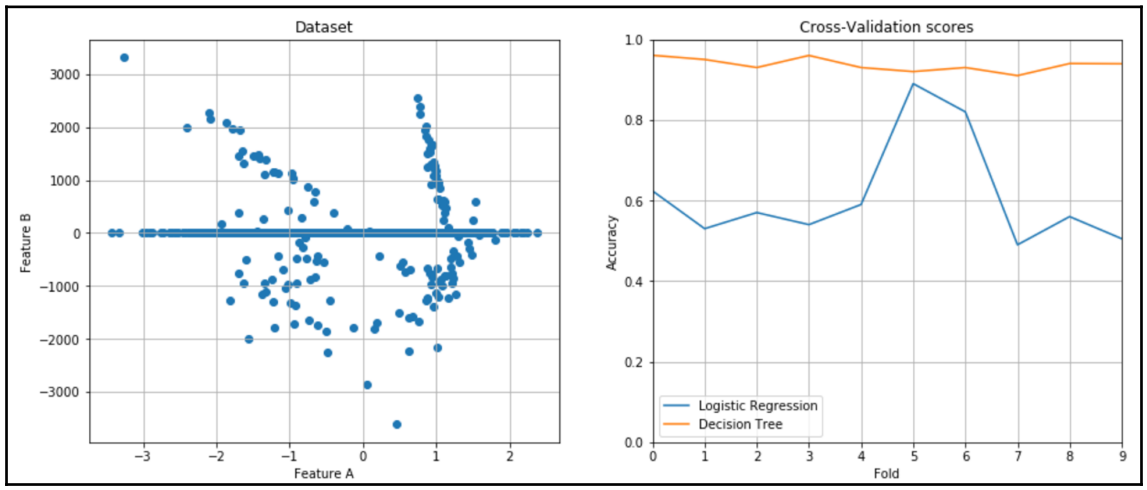

Original dataset
Scaled sound pressure (dB)
Predictions
Scaled sound pressure (dB)
Sample

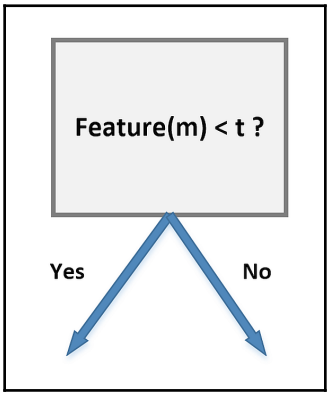

Absolute errors
|Y − Yp|
Sample

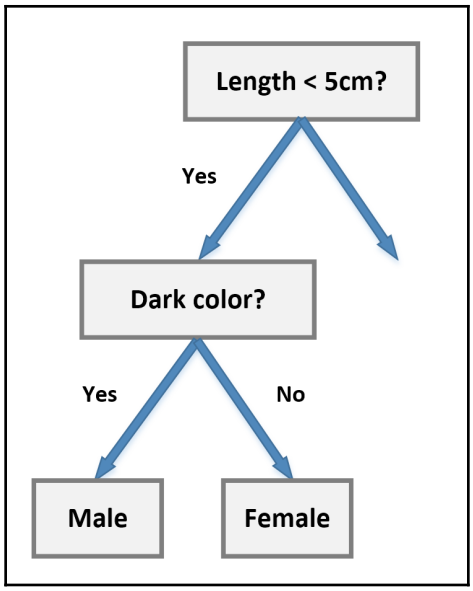

Class +1
Class -1
Unlabeled
x_0
x_1

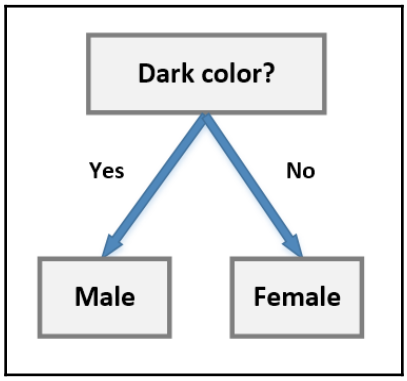

Class +1
Class -1
Unlabeled
x_0
x_1
Class +1
Class -1
x_0
x_1

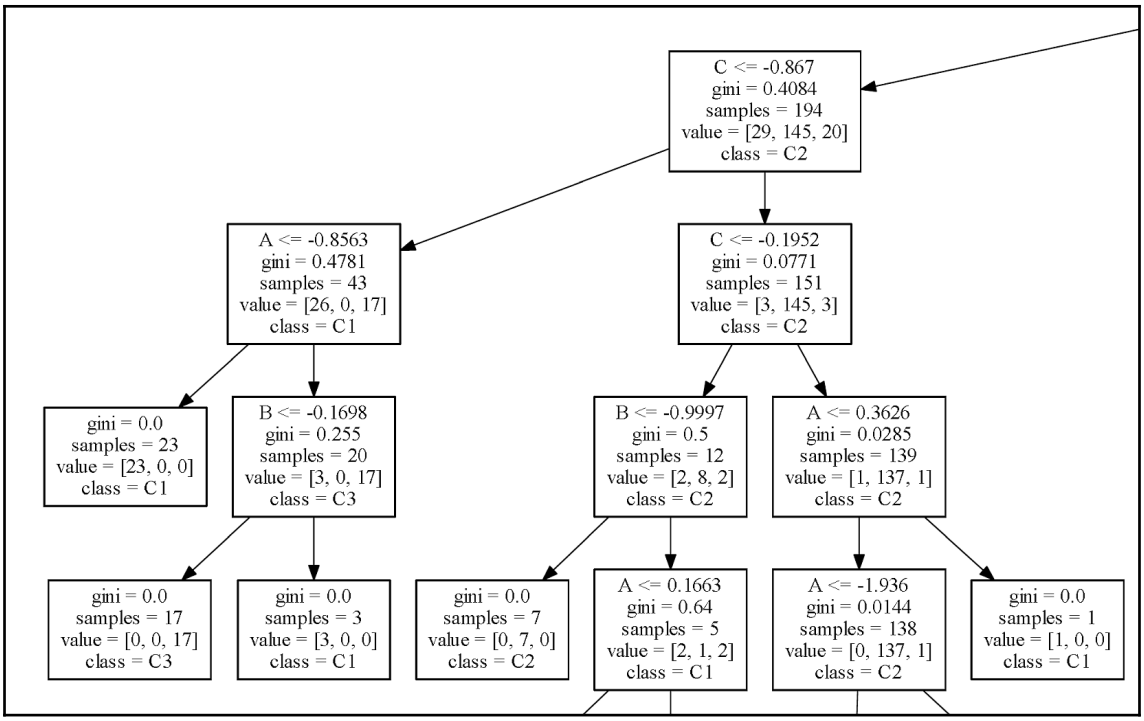
Chapter 08: Decision Trees and Ensemble Learning

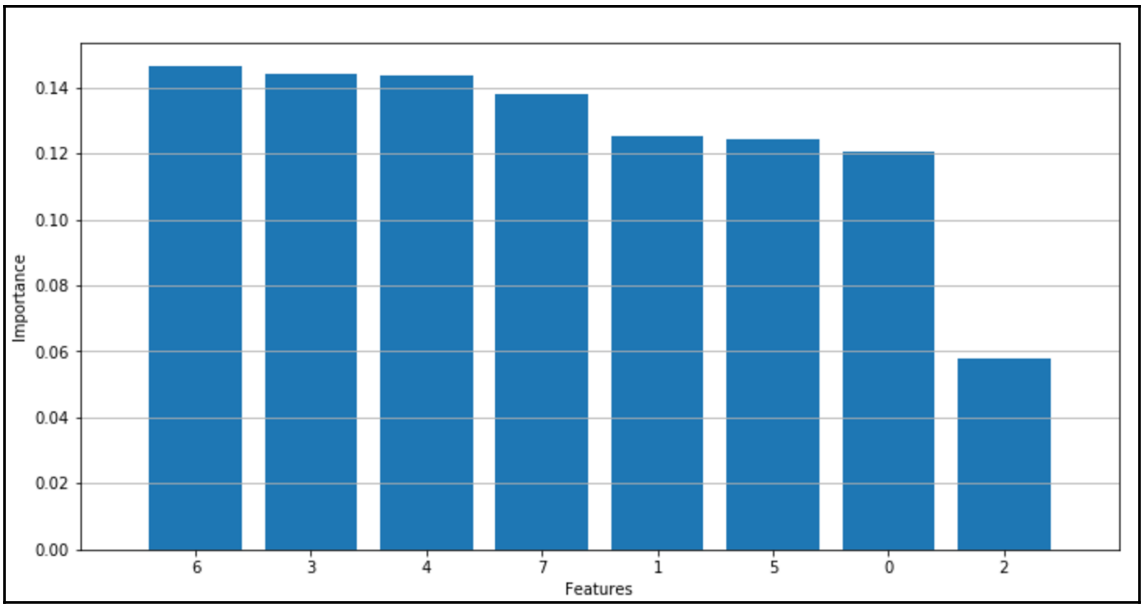


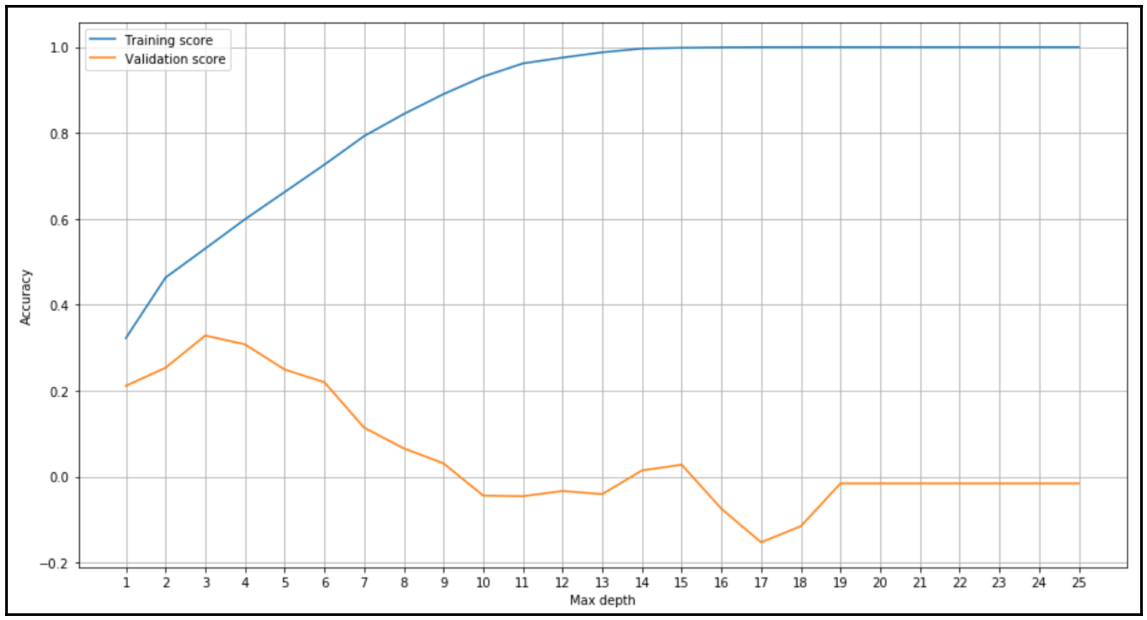


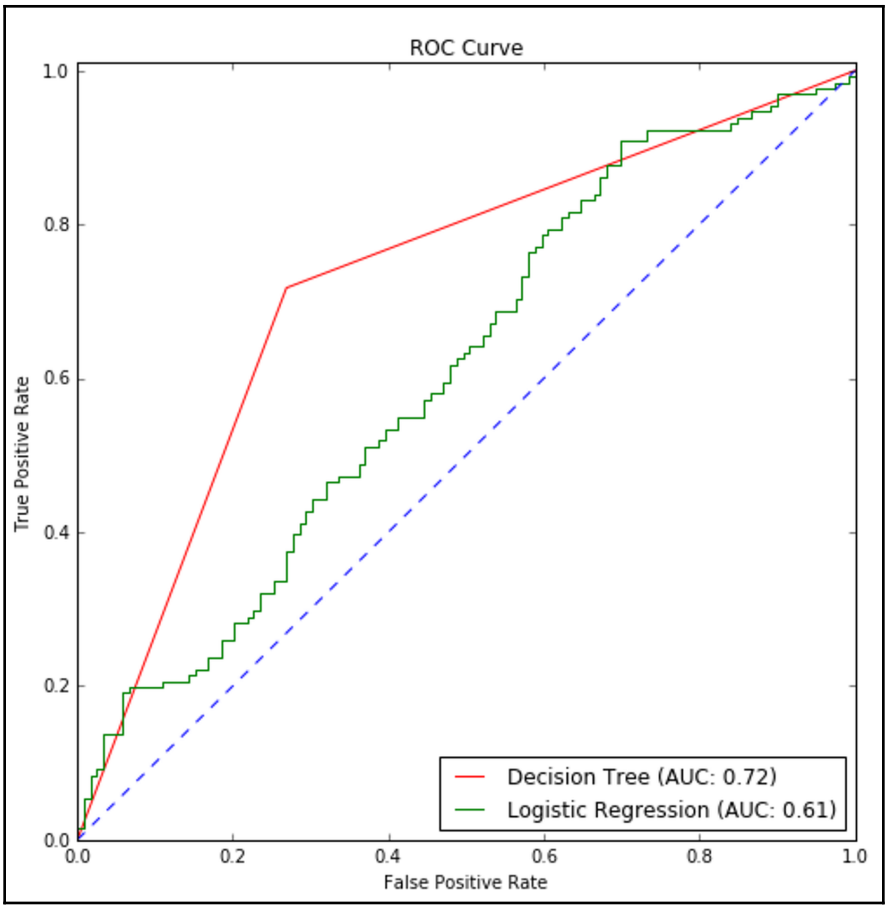

Length < 5cm?
Yes
Dark color?
Yes
No
Male
Female


Dark color?
Yes
No
Male
Female


C <= -0.867
gini = 0.4084
samples = 194
value = [29, 145, 20]
class = C2
A <= -0.8563
gini = 0.4781
samples = 43
value = [26, 0, 17]
class = C1
C <= -0.1952
gini = 0.0771
samples = 151
value = [3, 145, 3]
class = C2
gini = 0.0
samples = 23
value = [23, 0, 0]
class = C1
B <= -0.1698
gini = 0.255
samples = 20
value = [3, 0, 17]
class = C3
B <= -0.9997
gini = 0.5
samples = 12
value = [2, 8, 2]
class = C2
A <= 0.3626
gini = 0.0285
samples = 139
value = [1, 137, 1]
class = C2
gini = 0.0
samples = 17
value = [0, 0, 17]
class = C3
gini = 0.0
samples = 3
value = [3, 0, 0]
class = C1
gini = 0.0
samples = 7
value = [0, 7, 0]
class = C2
A <= 0.1663
gini = 0.64
samples = 5
value = [2, 1, 2]
class = C1
A <= -1.936
gini = 0.0144
samples = 138
value = [0, 137, 1]
class = C2
gini = 0.0
samples = 1
value = [1, 0, 0]
class = C1

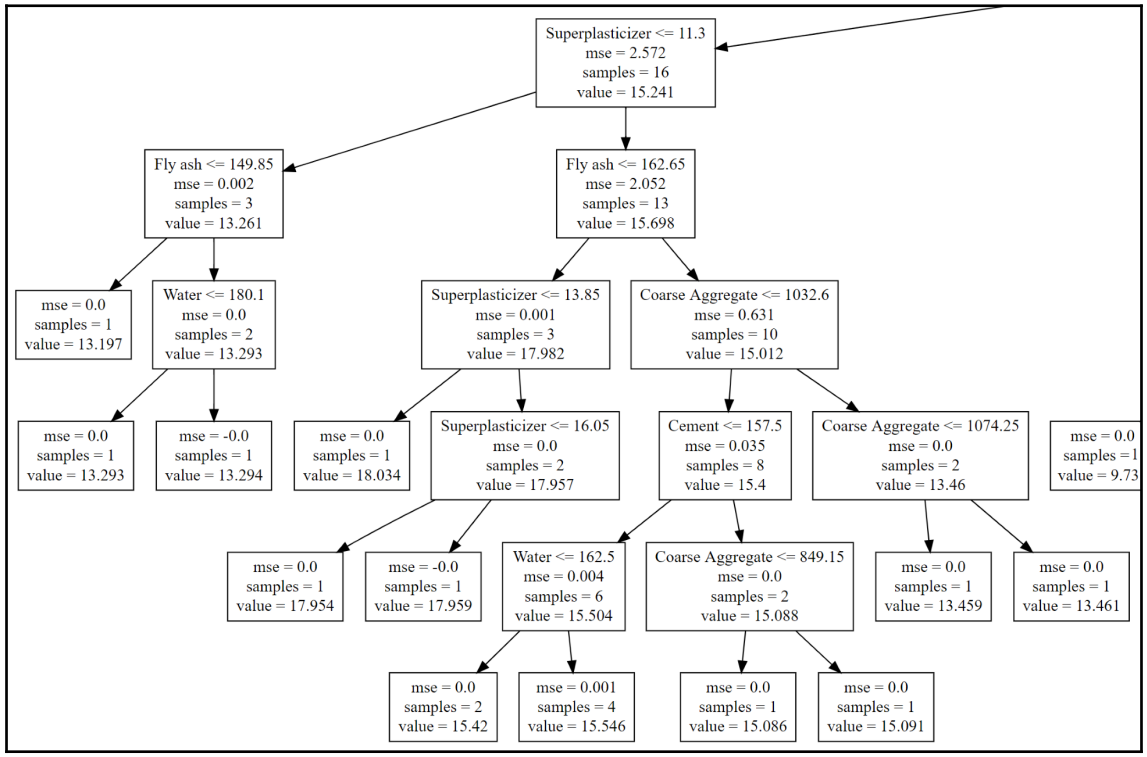

Importance
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
6
3
4
7
1
5
0
2
Features

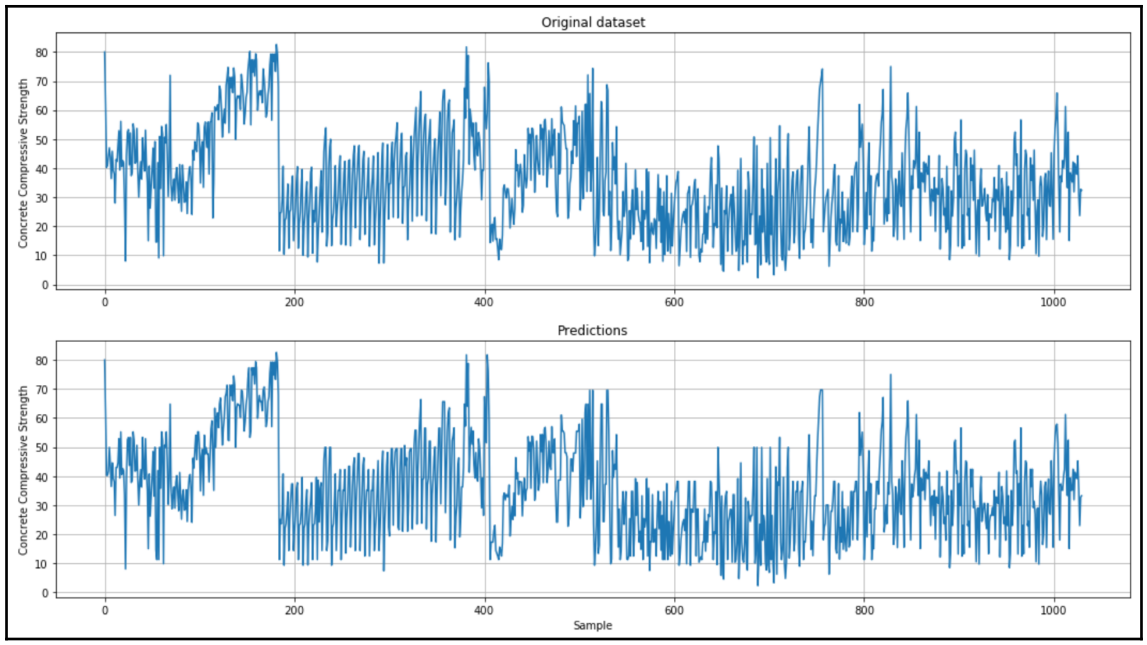

Training score
Validation score
Accuracy
Max depth
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

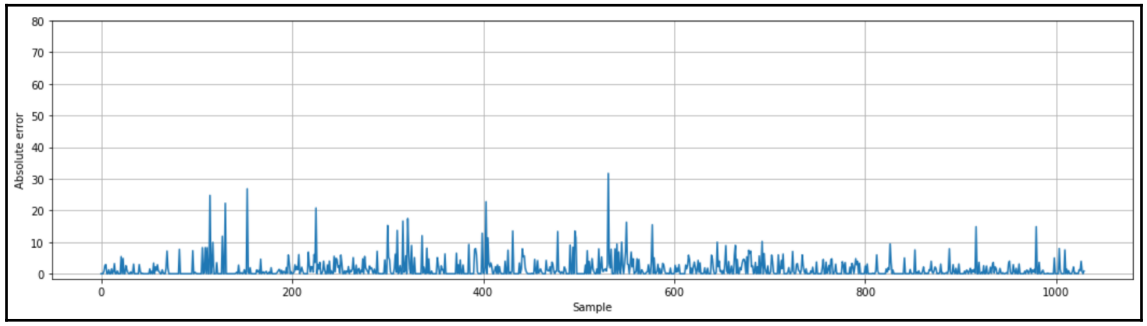


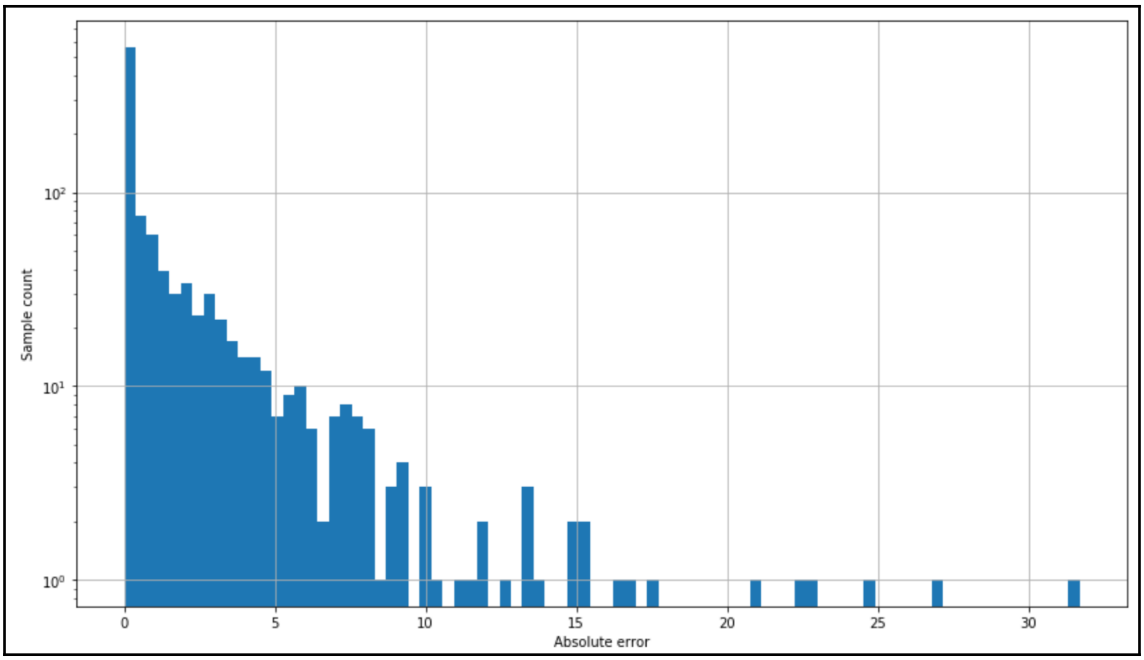
	Cement (component 1)(kg in a m^3 mixture)	Blast Furnace Slag (component 2)(kg in a m^3 mixture)	Fly Ash (component 3)(kg in a m^3 mixture)	Water (component 4)(kg in a m^3 mixture)	Superplasticizer (component 5) (kg in a m^3 mixture)	Coarse Aggregate (component 6)(kg in a m^3 mixture)	Fine Aggregate (component 7) (kg in a m^3 mixture)	Age (day)	Concrete compressive strength(MPa, megapascals)
0	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.986111
1	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.887366
2	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.269535
3	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.052780
4	198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.296075

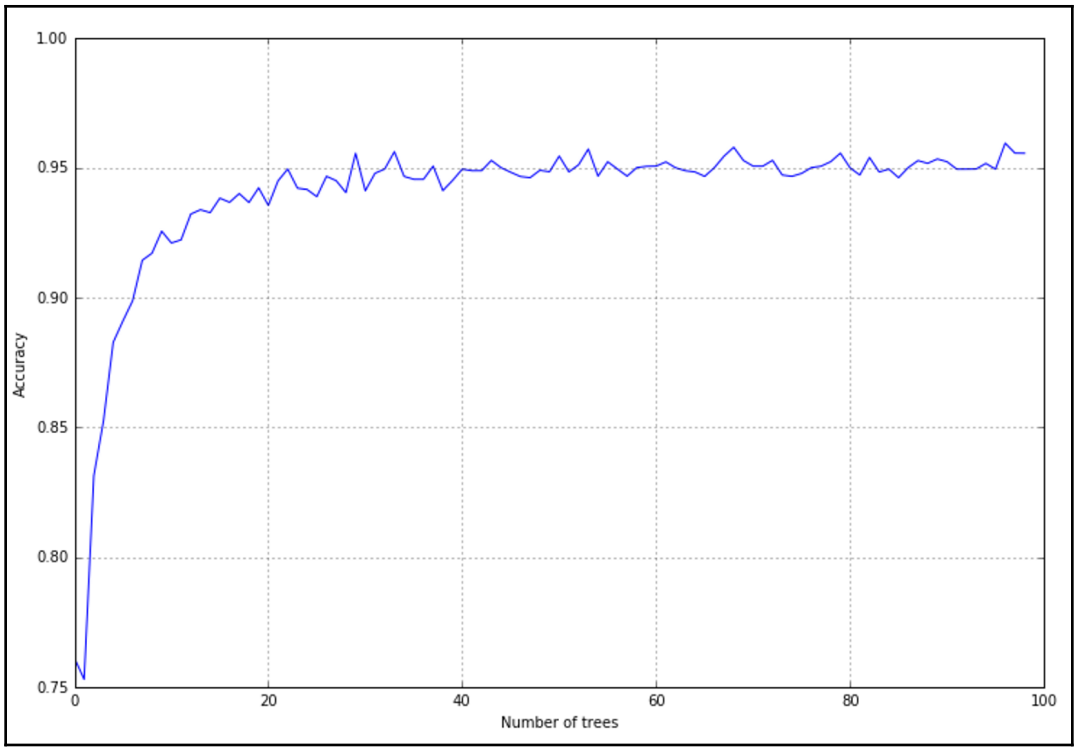
	Cement (component 1)(kg in a m^3 mixture)	Blast Furnace Slag (component 2) (kg in a m^3 mixture)	Fly Ash (component 3)(kg in a m^3 mixture)	Water (component 4)(kg in a m^3 mixture)	Superplasticizer (component 5) (kg in a m^3 mixture)	Coarse Aggregate (component 6) (kg in a m^3 mixture)	Fine Aggregate (component 7) (kg in a m^3 mixture)	Age (day)	Concrete compressive strength(MPa, megapascals)
count	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000
mean	281.165631	73.895485	54.187136	181.566359	6.203112	972.918592	773.578883	45.662136	35.817836
std	104.507142	86.279104	63.996469	21.355567	5.973492	77.753818	80.175427	63.169912	16.705679
min	102.000000	0.000000	0.000000	121.750000	0.000000	801.000000	594.000000	1.000000	2.331808
25%	192.375000	0.000000	0.000000	164.900000	0.000000	932.000000	730.950000	7.000000	23.707115
50%	272.900000	22.000000	0.000000	185.000000	6.350000	968.000000	779.510000	28.000000	34.442774
75%	350.000000	142.950000	118.270000	192.000000	10.160000	1029.400000	824.000000	56.000000	46.136287
max	540.000000	359.400000	200.100000	247.000000	32.200000	1145.000000	992.600000	365.000000	82.599225

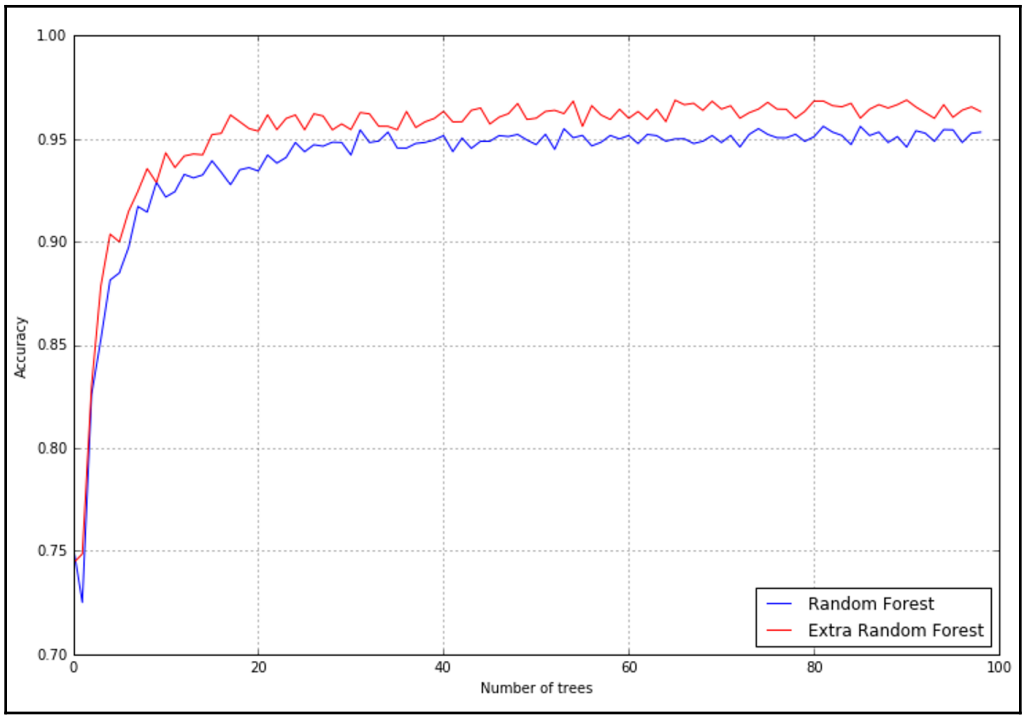


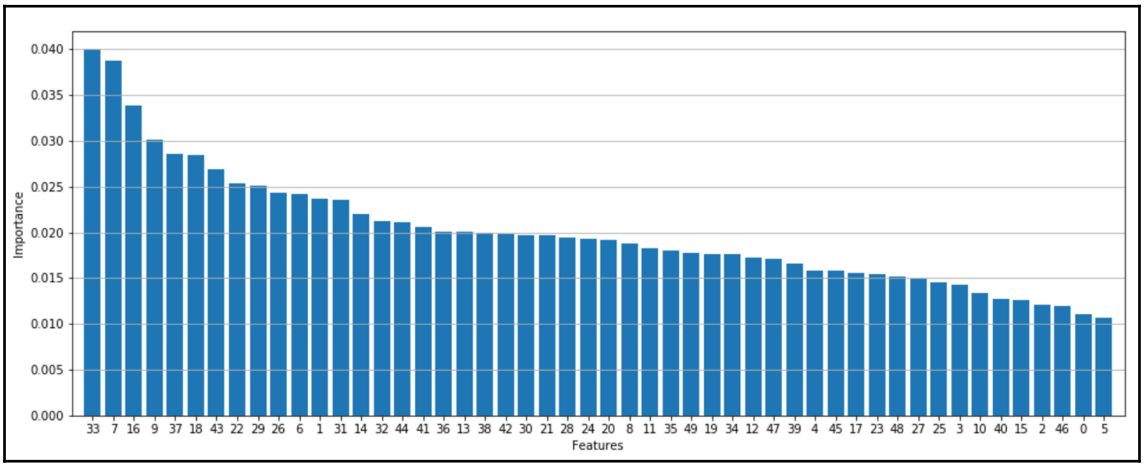

Original dataset
Concrete Compressive Strength
Predictions
Concrete Compressive Strength
Sample

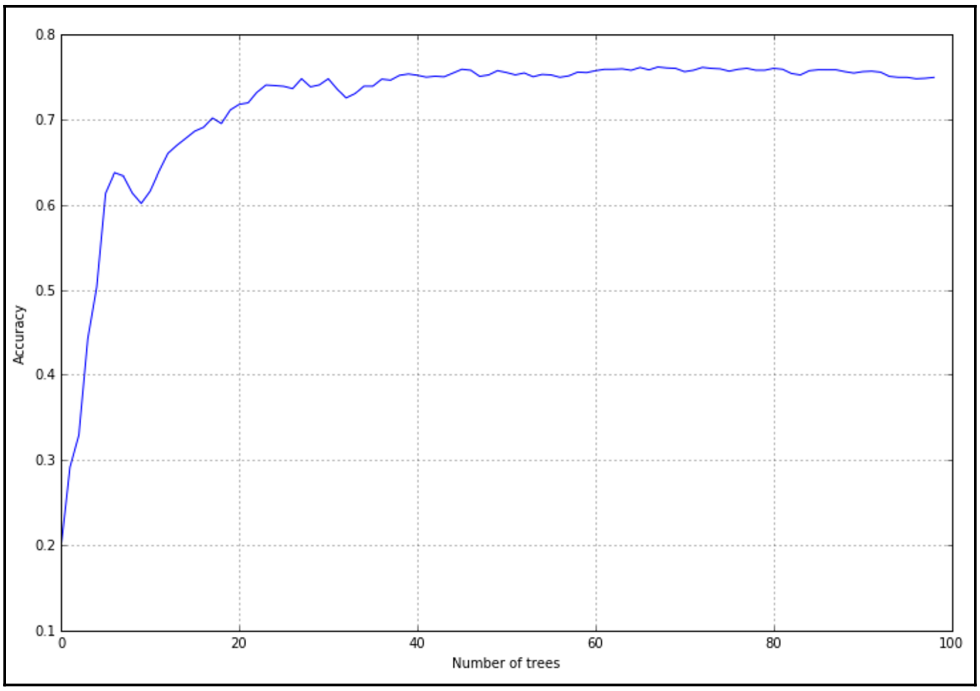

Absolute error
Sample

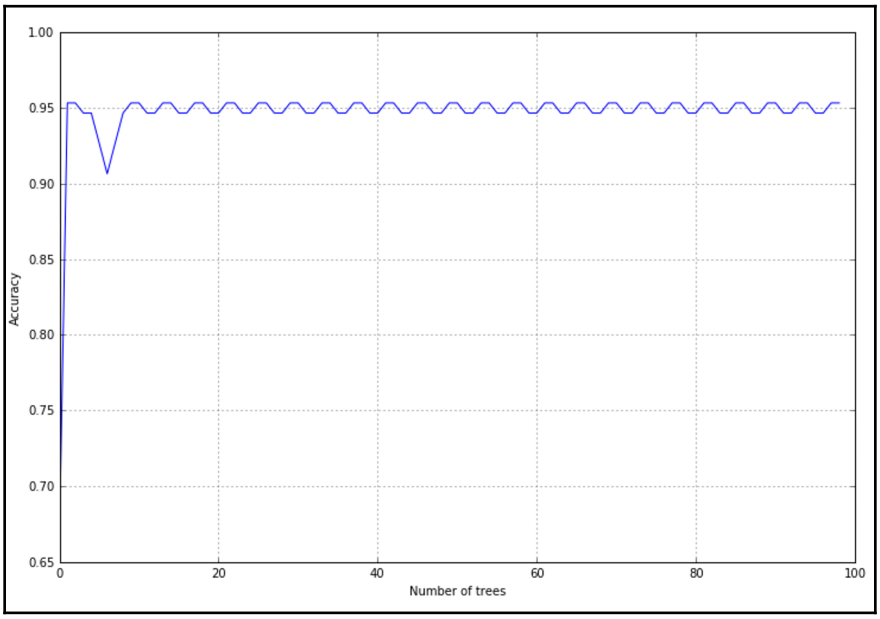

Sample count
10^2
10^1
10^0
0
5
10
15
20
25
30
Absolute error

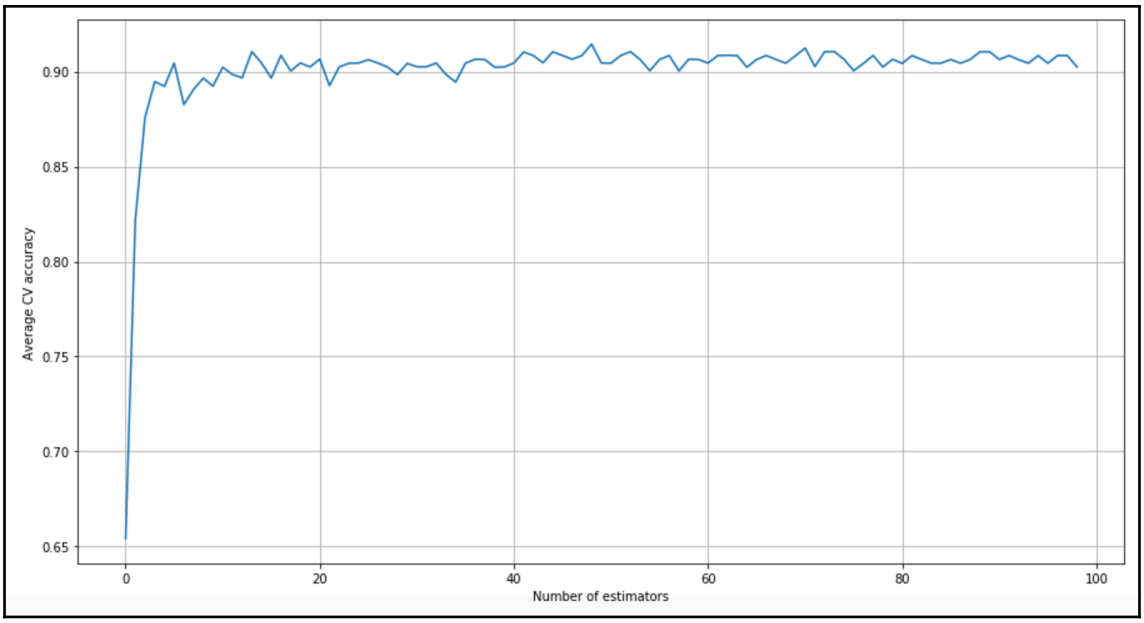

1.00
0.95
0.90
0.85
0.80
0.75
Accuracy
0
20
40
60
80
100
Number of trees

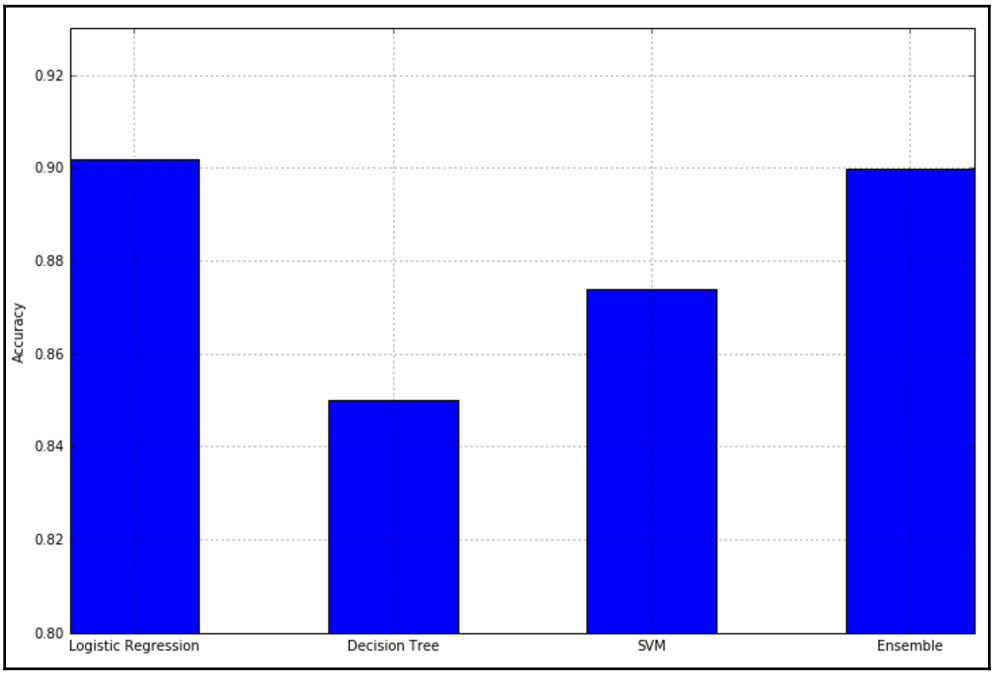

1.00
0.95
0.90
0.85
0.80
0.75
0.70
Accuracy
0
20
40
60
80
100
Number of trees
Random Forest
Extra Random Forest


0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Importance
33 7 16 9 37 18 43 22 29 26 6 1 31 14 32 44 41 36 13 38 42 30 21 28 24 20 8 11 35 49 19 34 12 47 39 4 45 17 23 48 27 25 3 10 40 15 2 46 0 5
Features

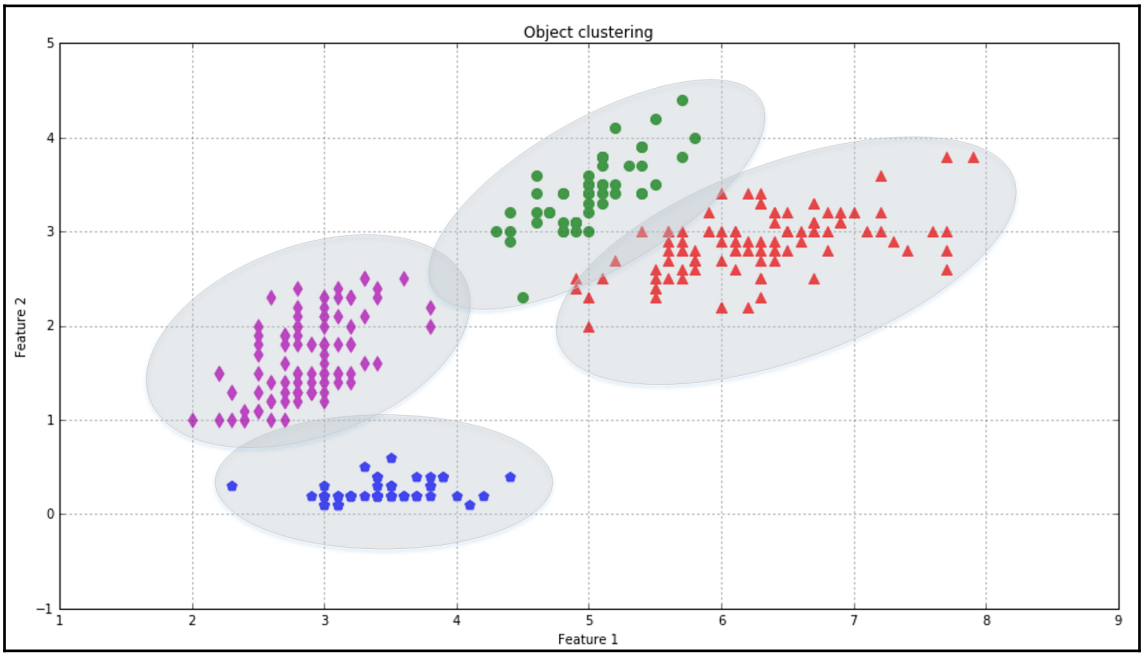

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Accuracy
0
20
40
60
80
100
Number of trees

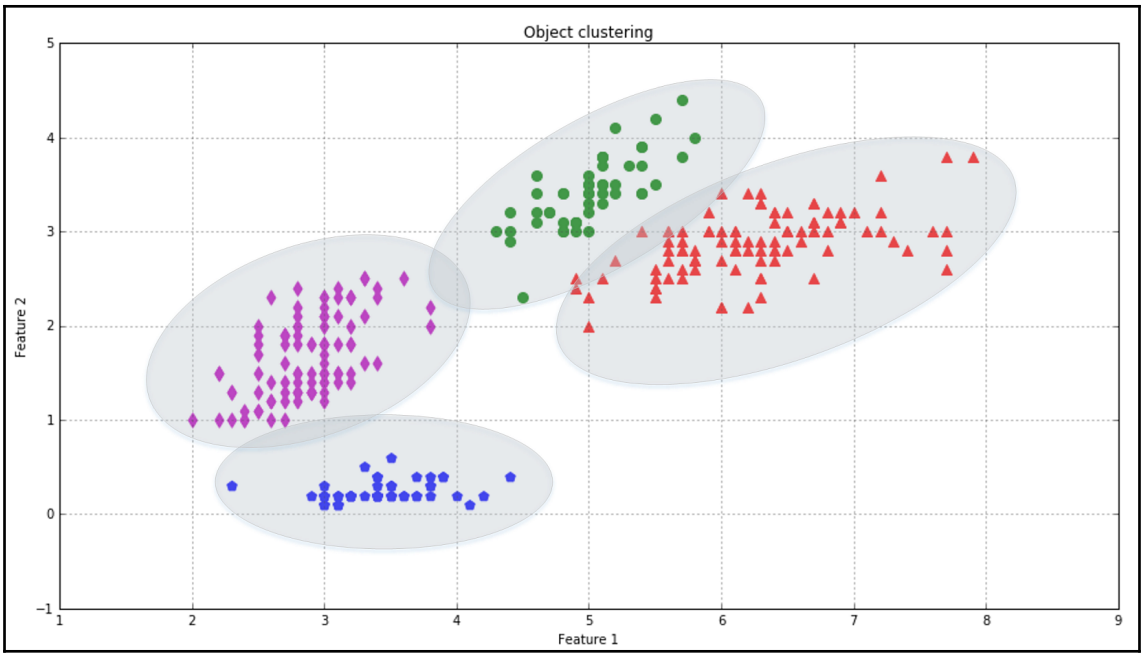

1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
Accuracy
0
20
40
60
80
100
Number of trees

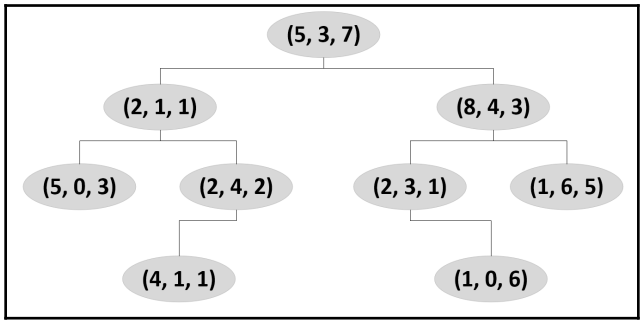

Average CV accuracy
0.90
0.85
0.80
0.75
0.70
0.65
0
20
40
60
80
100
Number of estimators

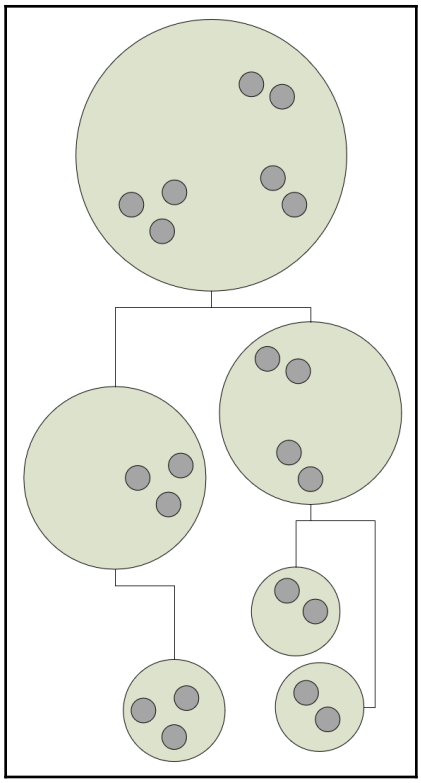

Accuracy
0.92
0.90
0.88
0.86
0.84
0.82
0.80
Logistic Regression
Decision Tree
SVM
Ensemble

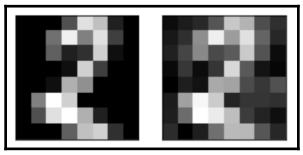
Chapter 09: Clustering Fundamentals

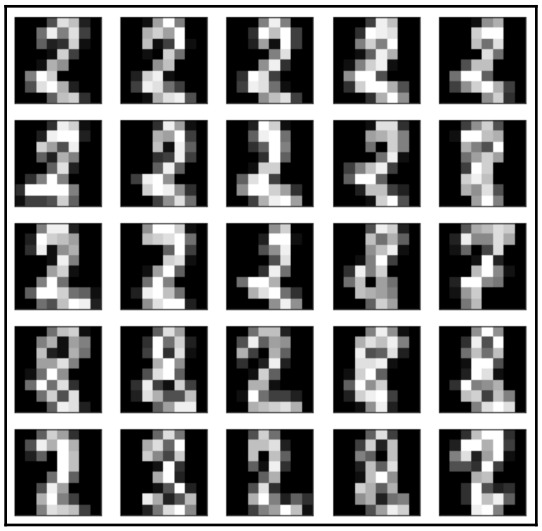


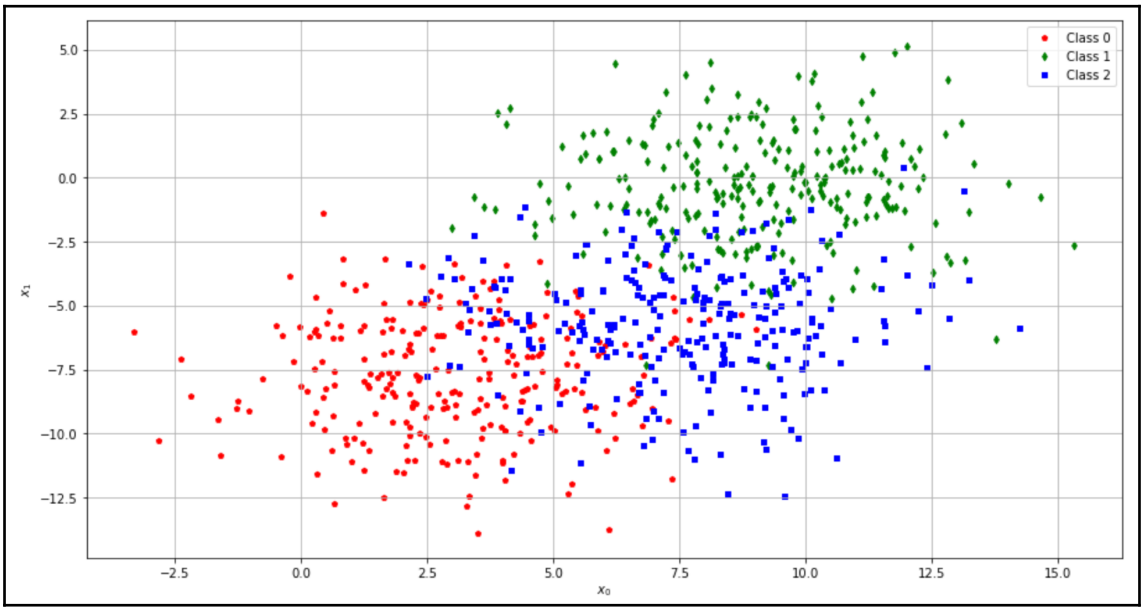

Object clustering
Feature 2
Feature 1
5
4
3
2
1
0
−1
1
2
3
4
5
6
7
8
9

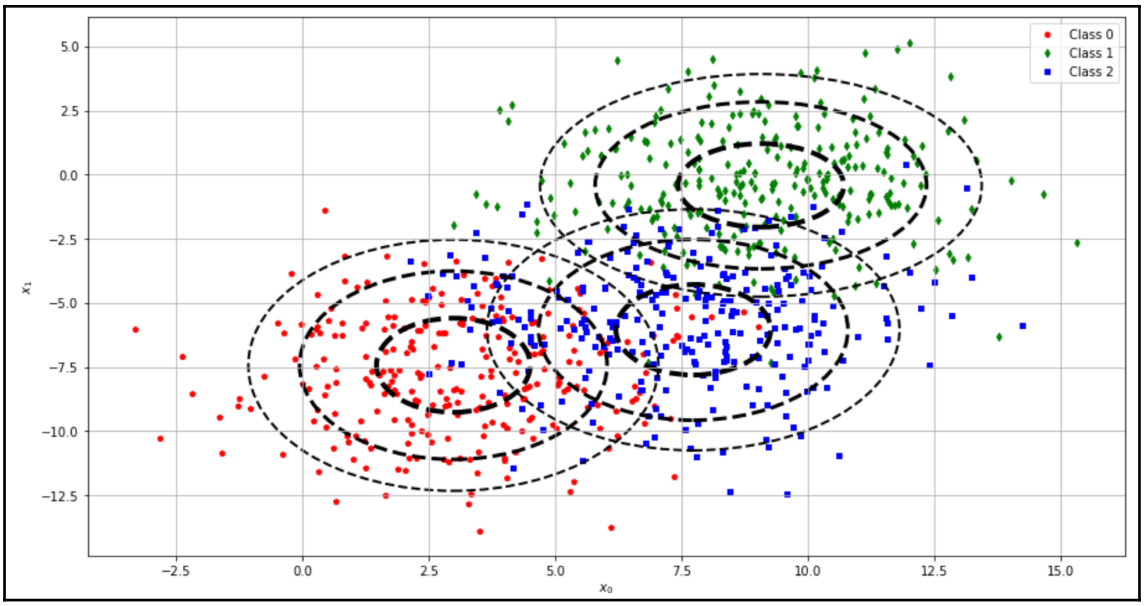

(5, 3, 7)
(2, 1, 1)
(8, 4, 3)
(5, 0, 3)
(2, 4, 2)
(2, 3, 1)
(1, 6, 5)
(4, 1, 1)
(1, 0, 6)

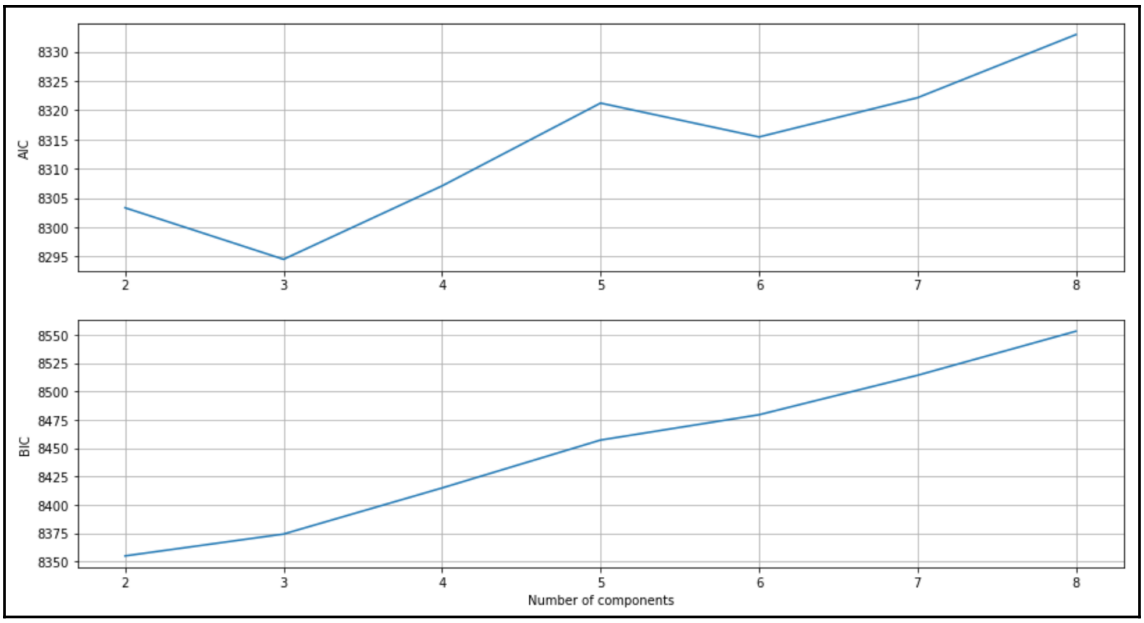


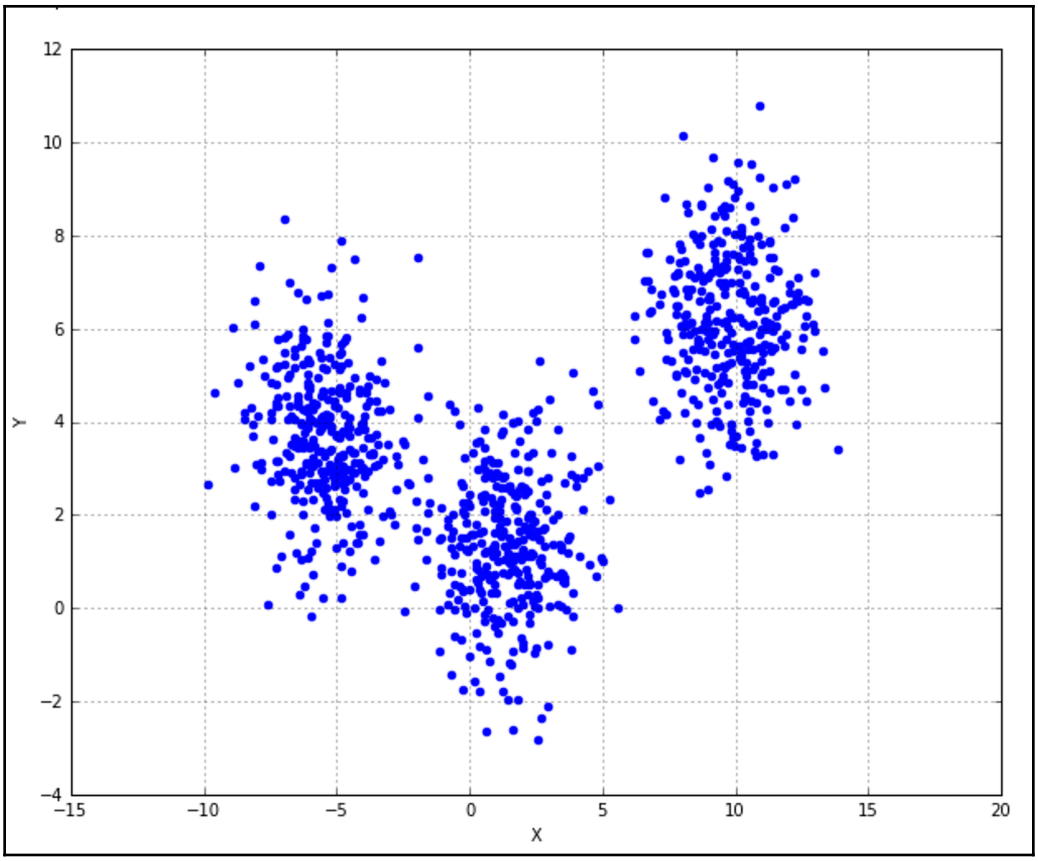


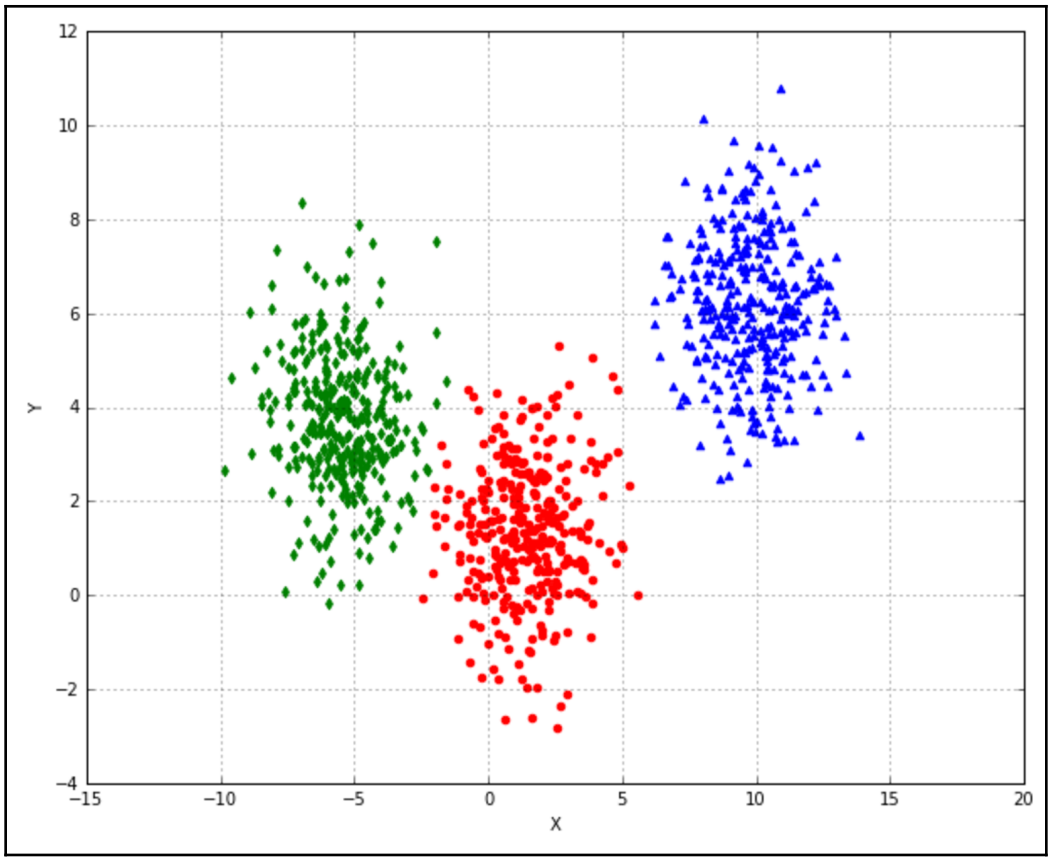


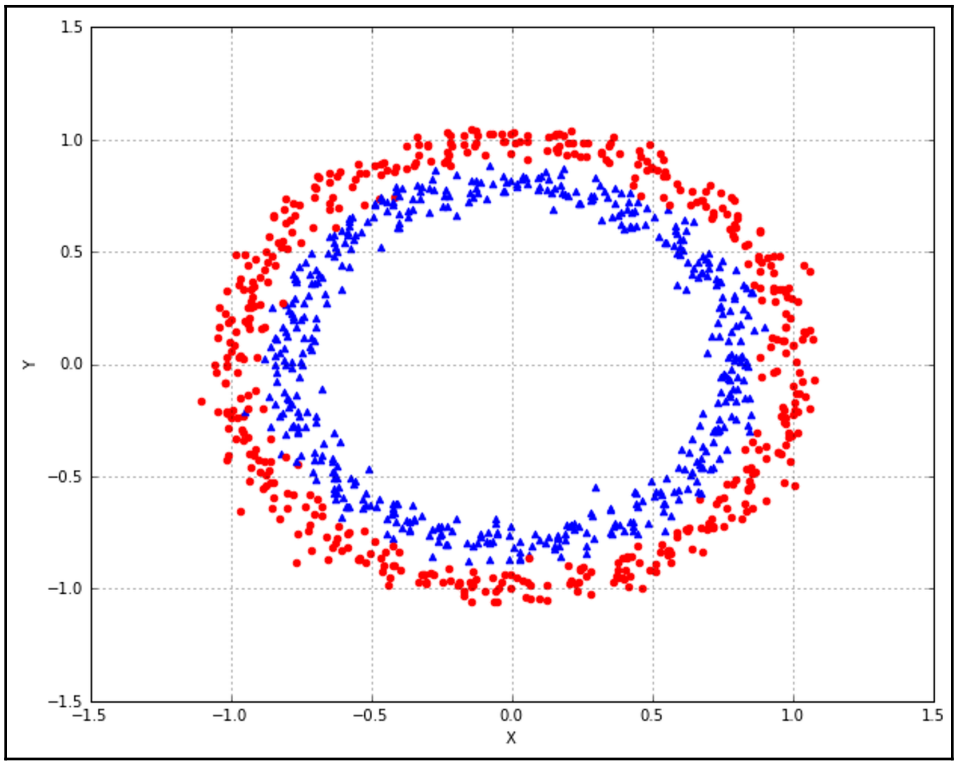

Class 0
Class 1
Class 2
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
x_1
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
x_0

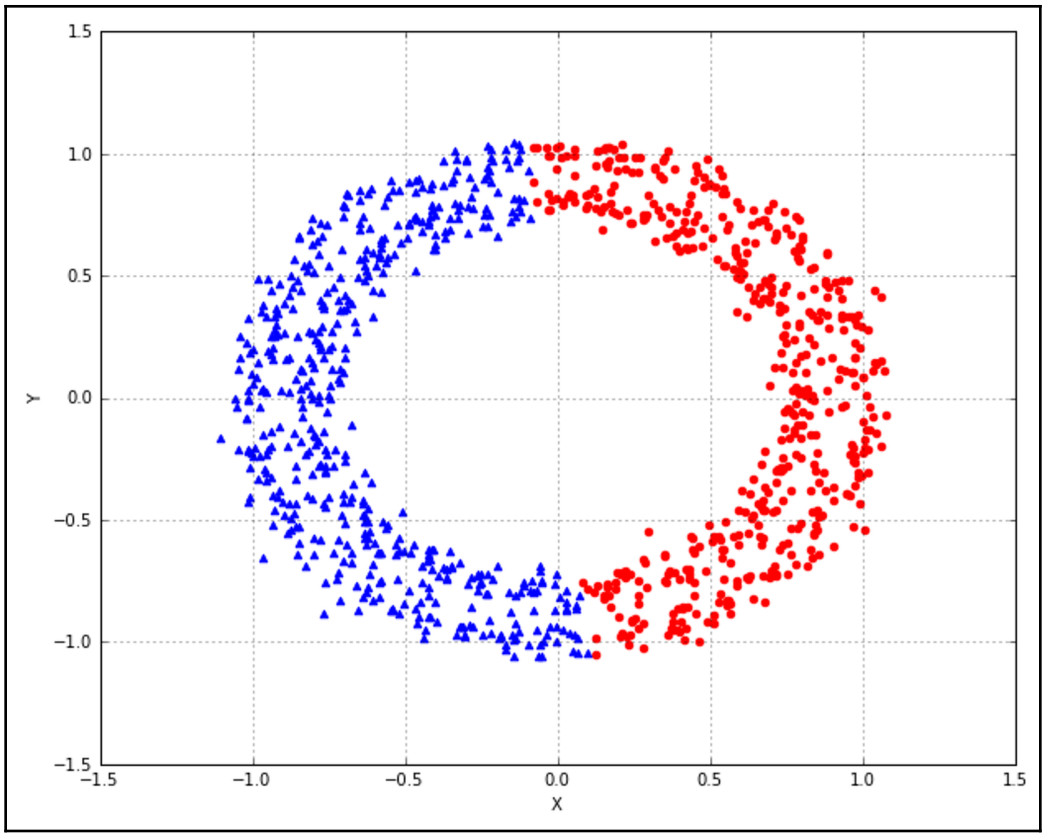

Class 0
Class 1
Class 2
x_1
x_0

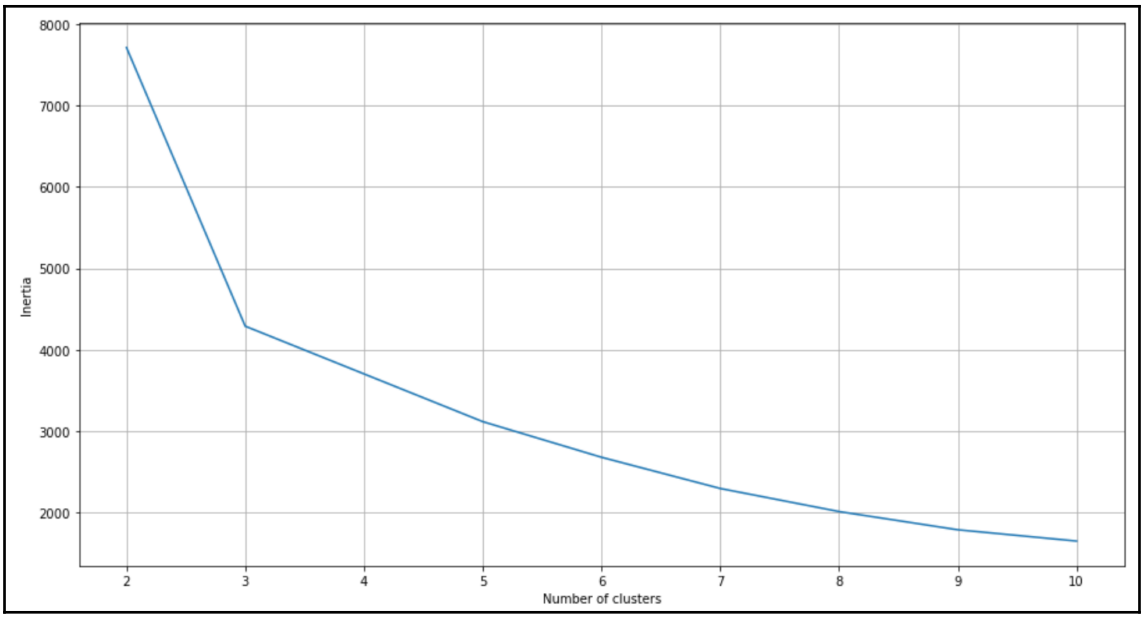

AIC
BIC
Number of components

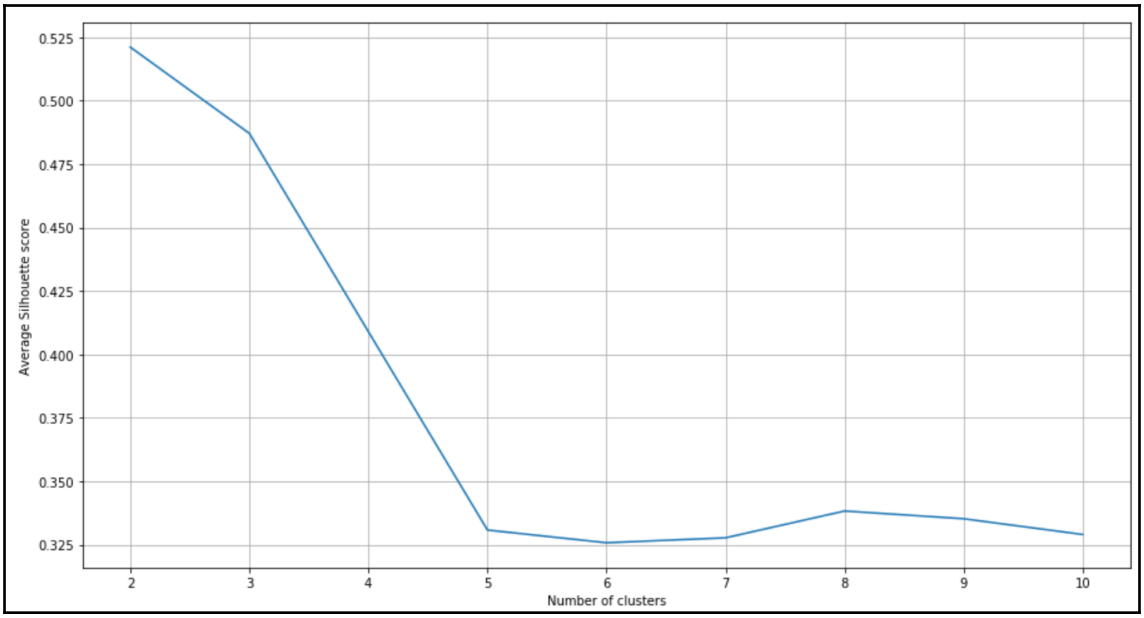

12
10
8
6
4
2
0
−2
−4
Y
−15
−10
−5
0
5
10
15
20
X

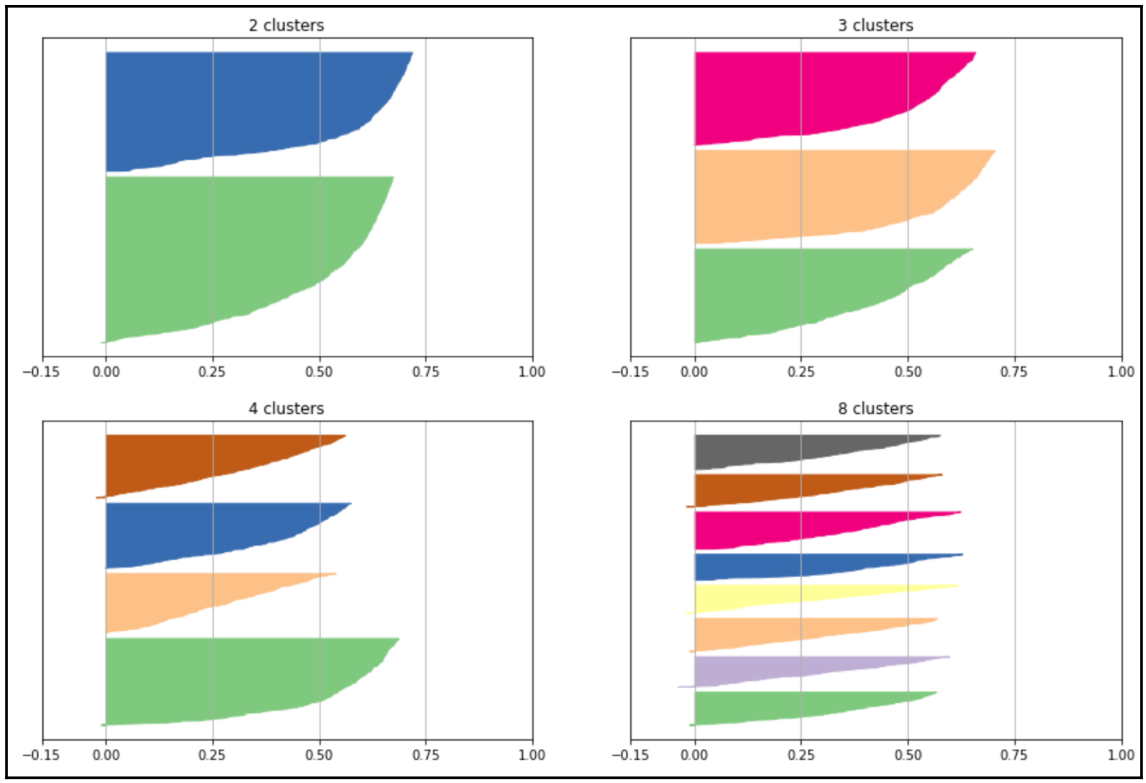

12
10
8
6
4
2
0
−2
−4
Y
−15
−10
−5
0
5
10
15
20
X

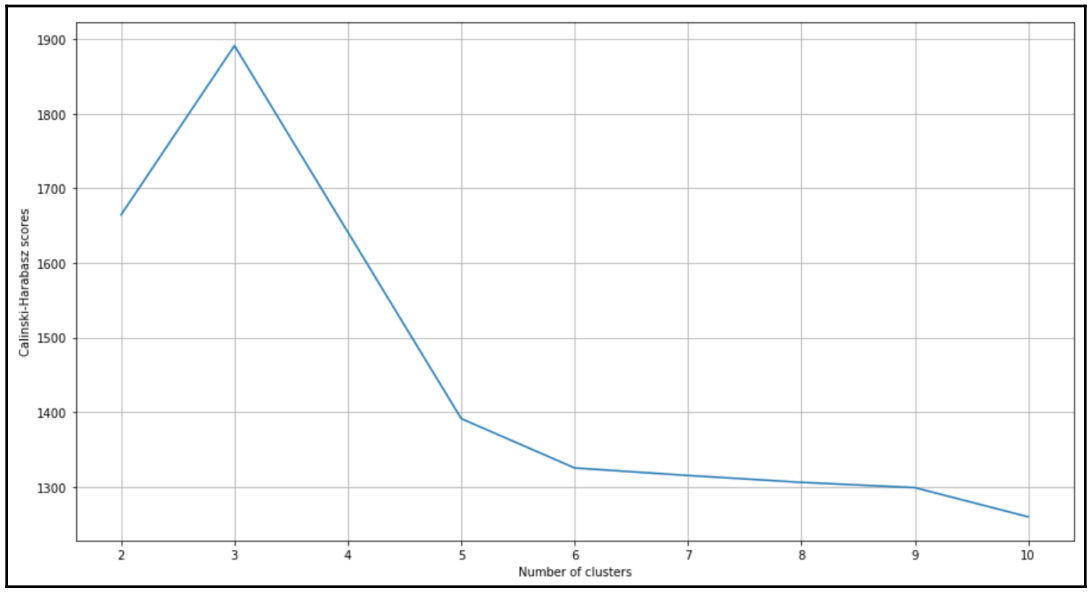

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
X

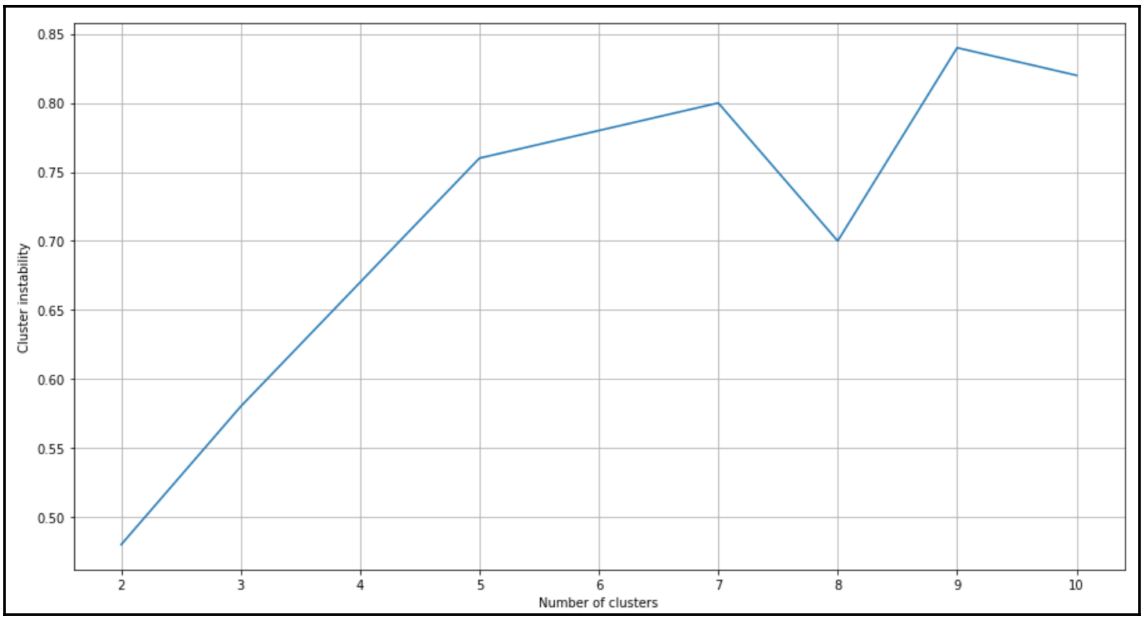

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
X


8000
7000
6000
5000
4000
3000
2000
Inertia
2
3
4
5
6
7
8
9
10
Number of clusters

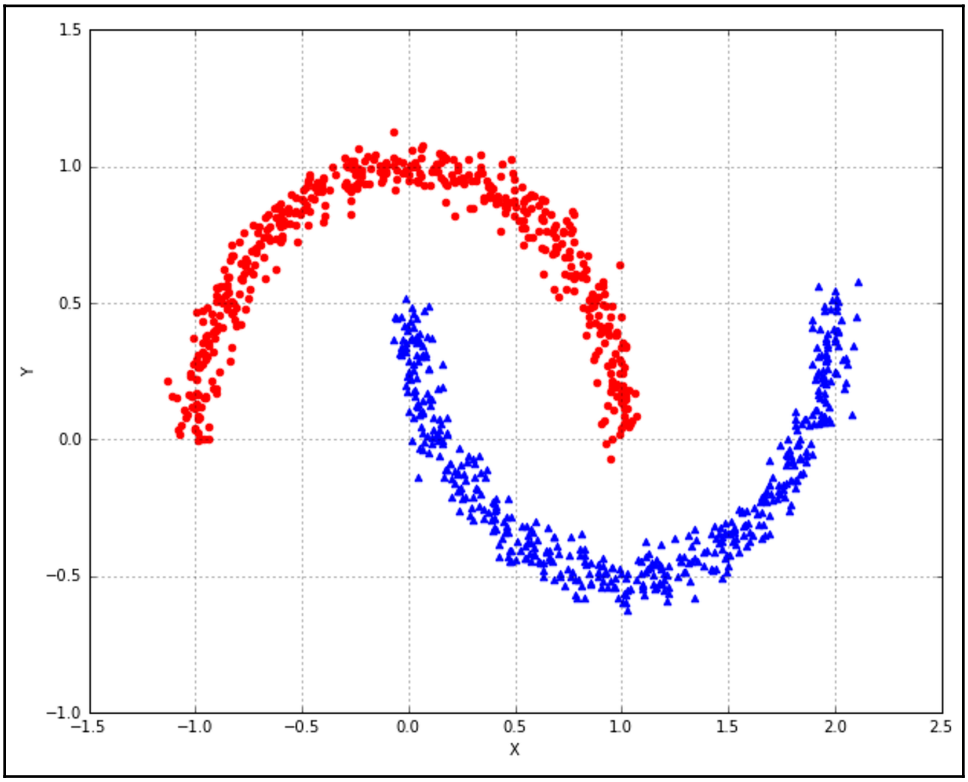

0.525
0.500
0.475
0.450
0.425
0.400
0.375
0.350
0.325
Average Silhouette score
2
3
4
5
6
7
8
9
10
Number of clusters

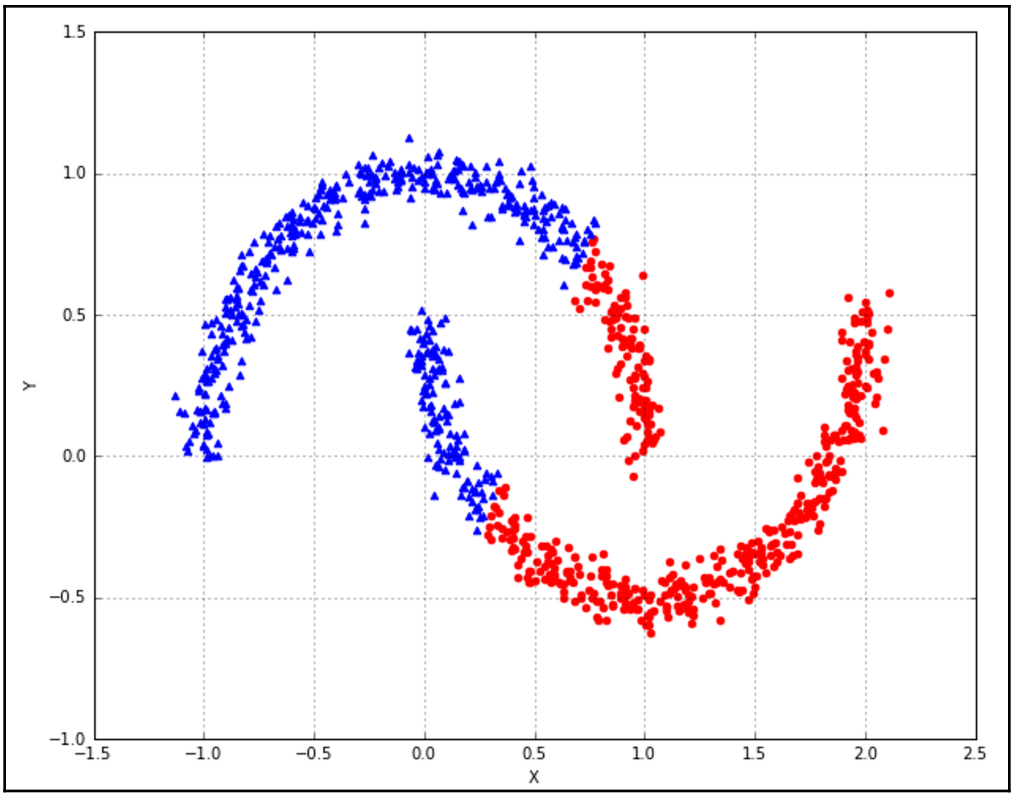

2 clusters
3 clusters
4 clusters
8 clusters
−0.15
0.00
0.25
0.50
0.75
1.00

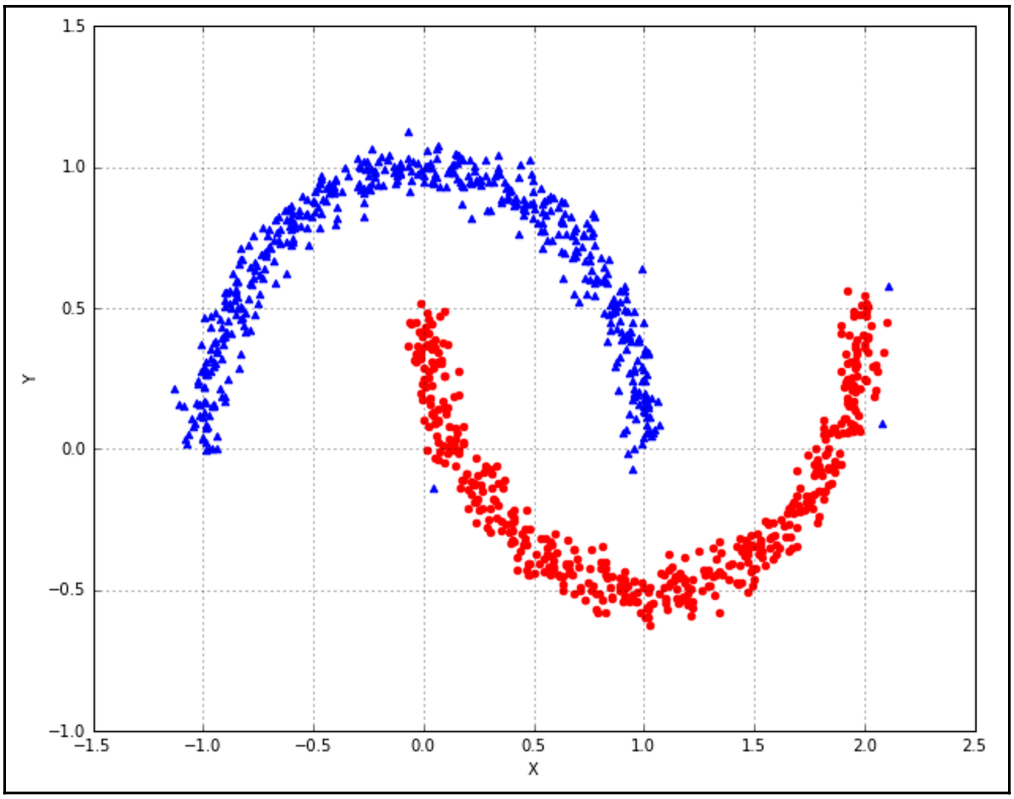

Calinski-Harabasz scores
Number of clusters
1900
1800
1700
1600
1500
1400
1300
2
3
4
5
6
7
8
9
10

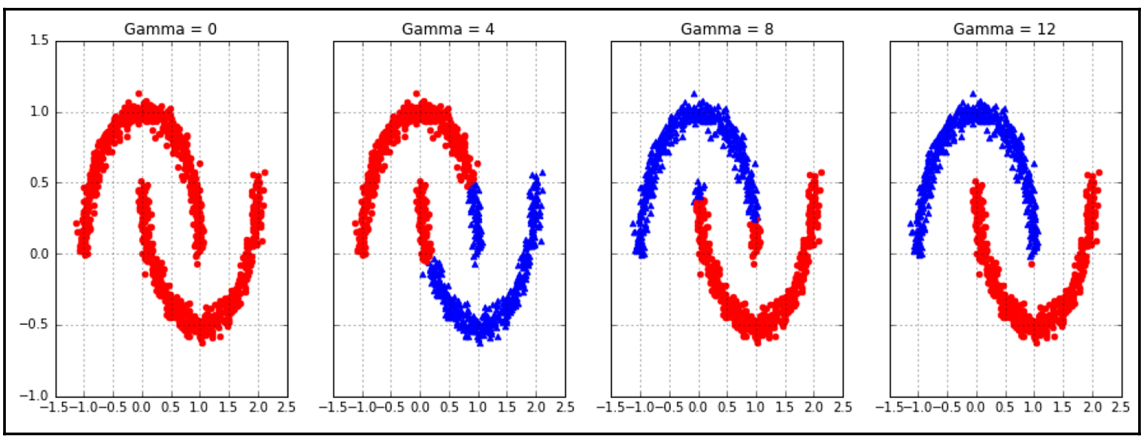

0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
Cluster instability
2
3
4
5
6
7
8
9
10
Number of clusters

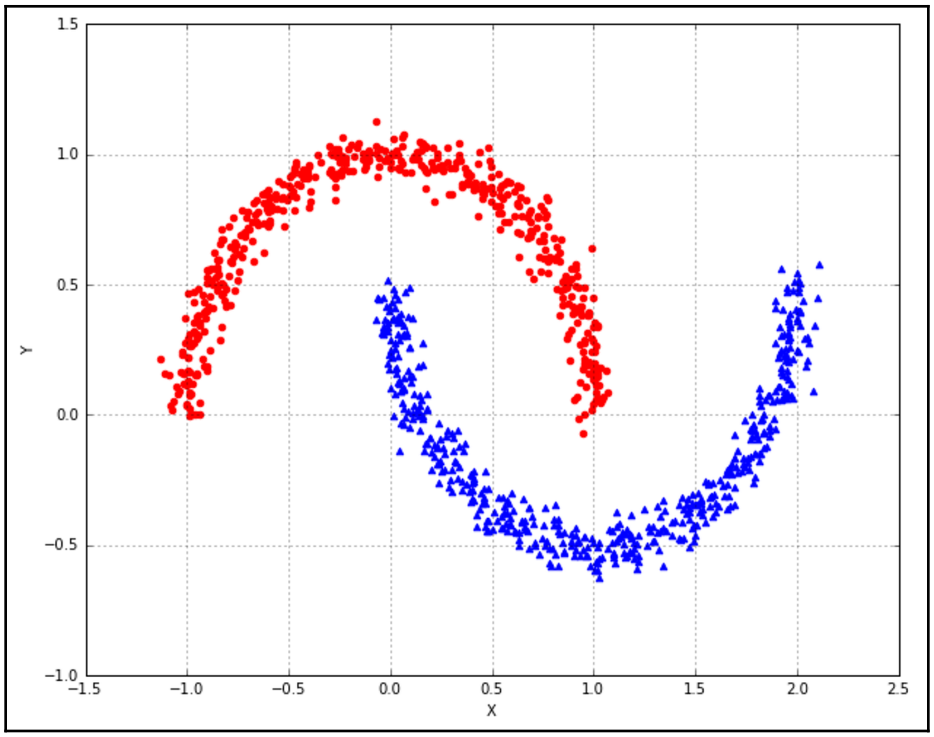
Chapter 10: Advanced Clustering

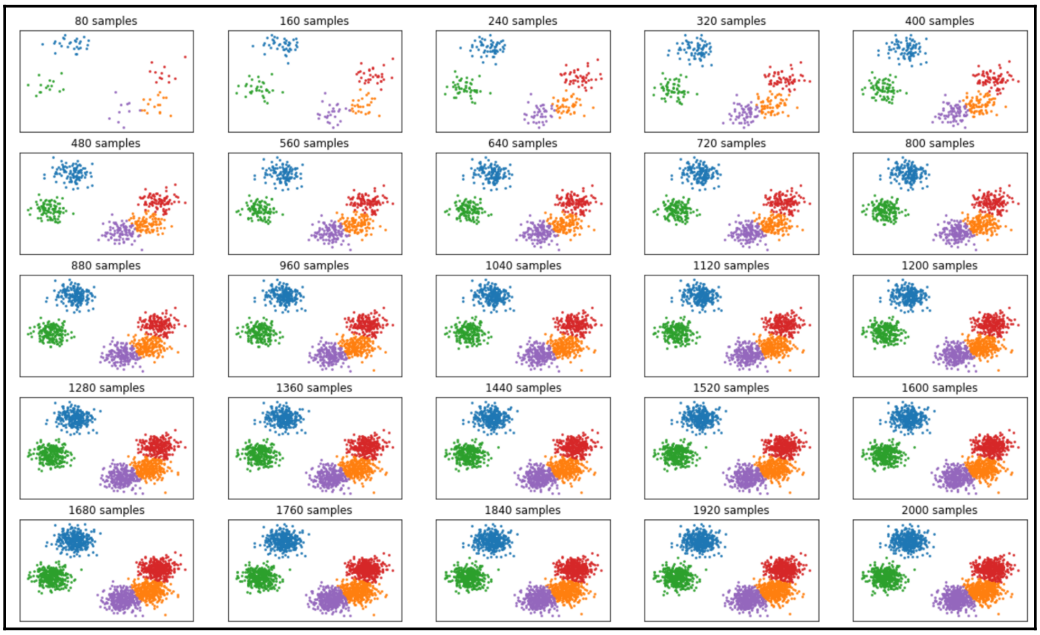


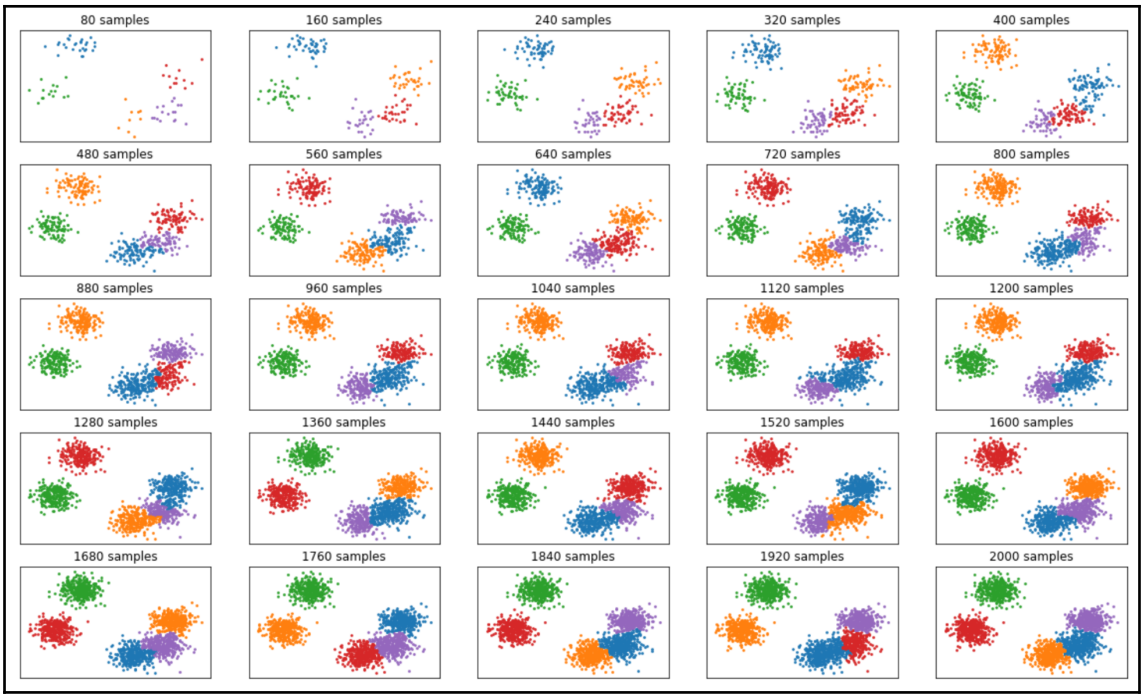

1.5
1.0
0.5
0.0
−0.5
−1.0
Y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
X

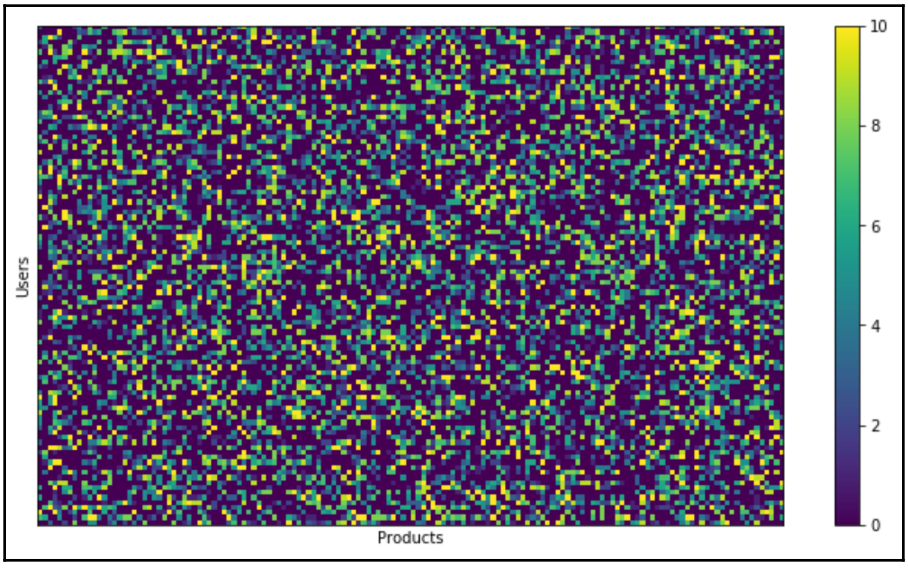

1.5
1.0
0.5
0.0
−0.5
−1.0
Y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
X

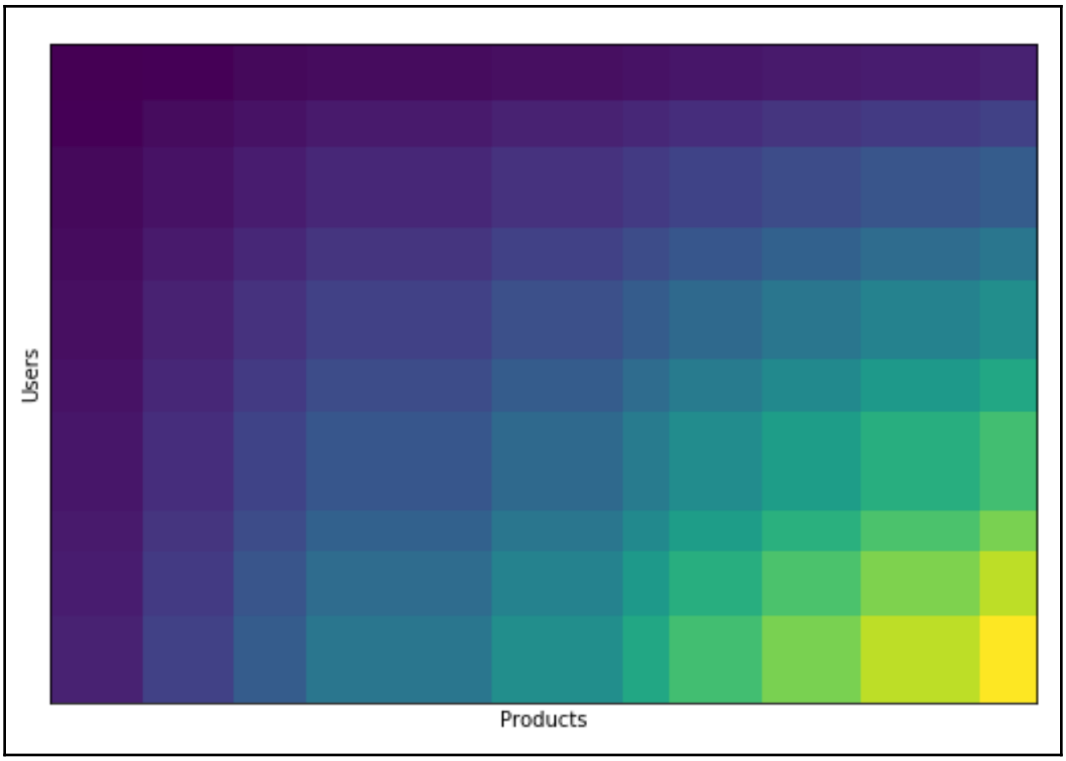

Gamma = 0
Gamma = 4
Gamma = 8
Gamma = 12


1.5
1.0
0.5
0.0
−0.5
−1.0
Y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
X

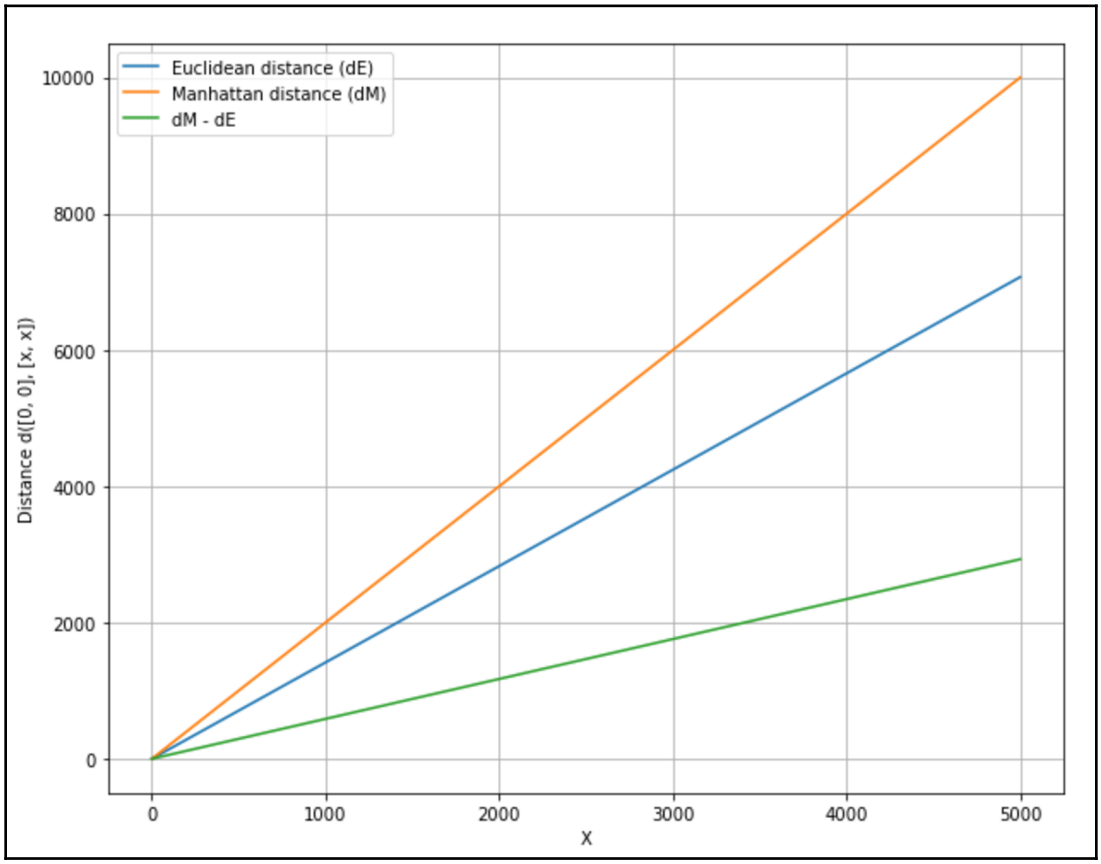

80 samples
160 samples
240 samples
320 samples
400 samples
480 samples
560 samples
640 samples
720 samples
800 samples
880 samples
960 samples
1040 samples
1120 samples
1200 samples
1280 samples
1360 samples
1440 samples
1520 samples
1600 samples
1680 samples
1760 samples
1840 samples
1920 samples
2000 samples

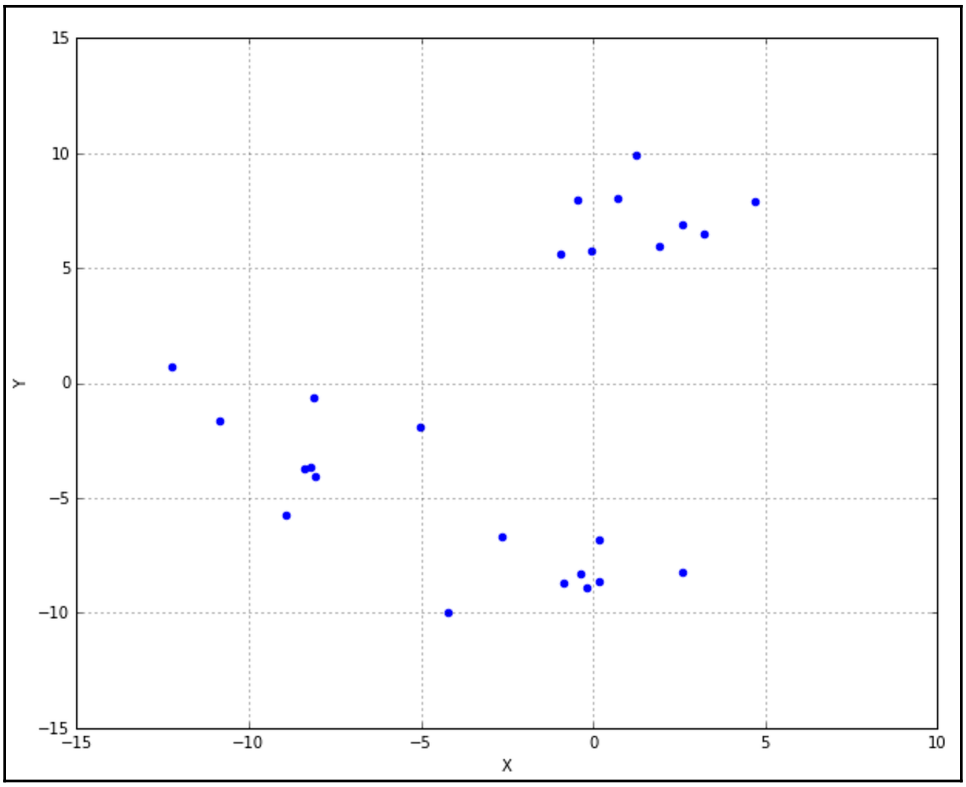

80 samples
160 samples
240 samples
320 samples
400 samples
480 samples
560 samples
640 samples
720 samples
800 samples
880 samples
960 samples
1040 samples
1120 samples
1200 samples
1280 samples
1360 samples
1440 samples
1520 samples
1600 samples
1680 samples
1760 samples
1840 samples
1920 samples
2000 samples

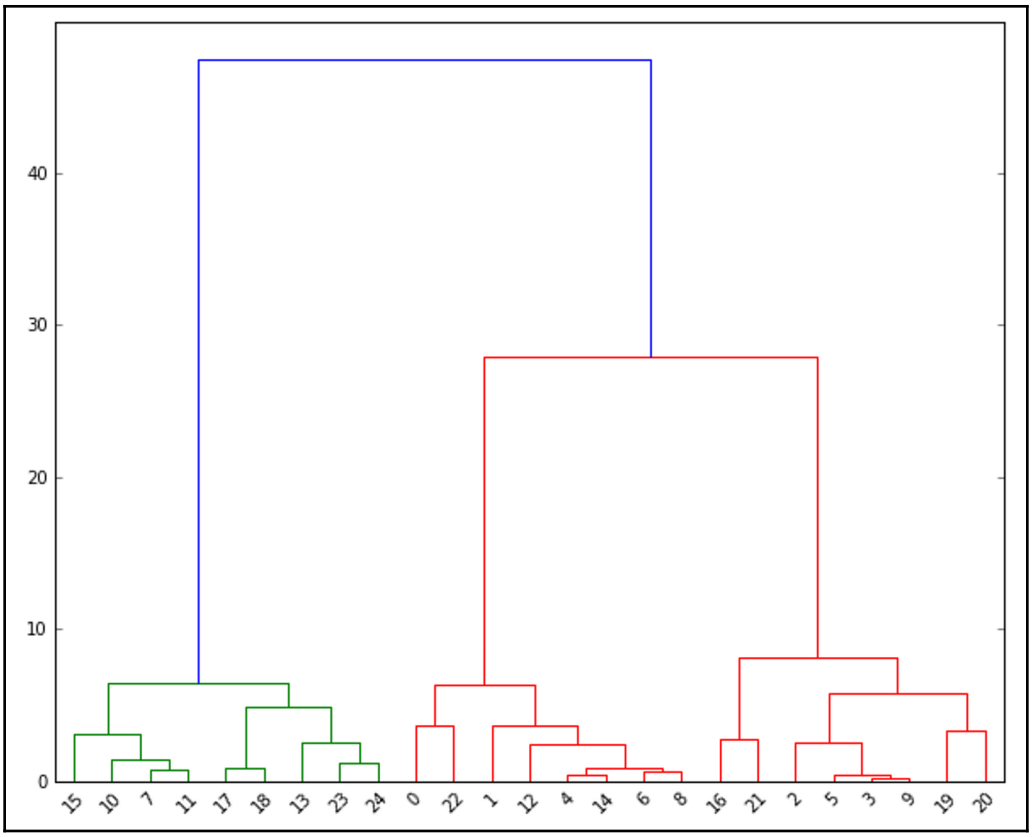

Users
Products
10
8
6
4
2
0

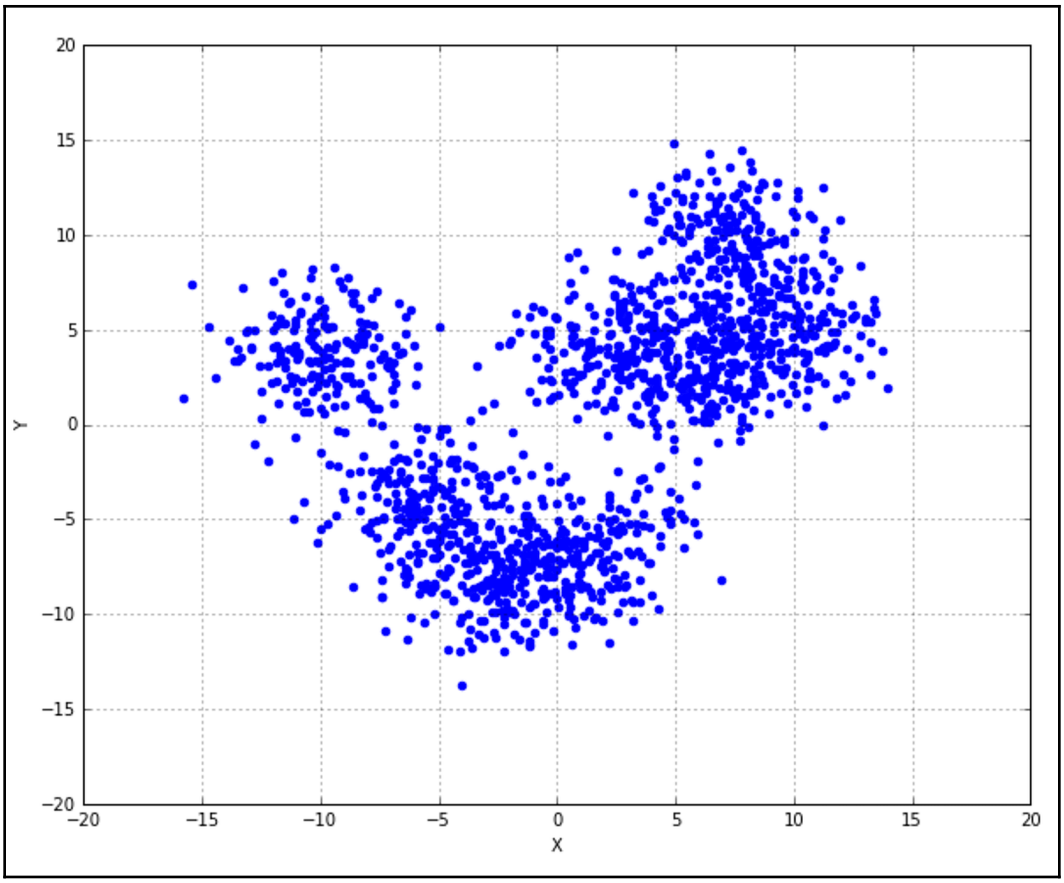

Users
Products

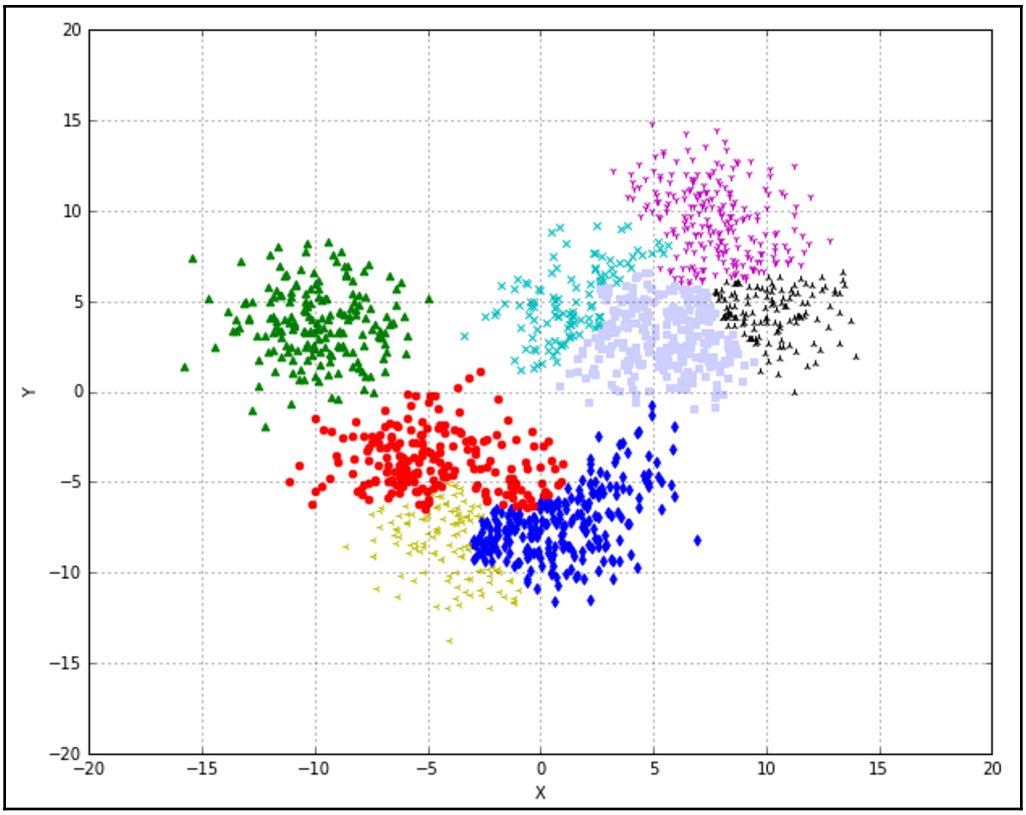
Chapter 11: Hierarchical Clustering

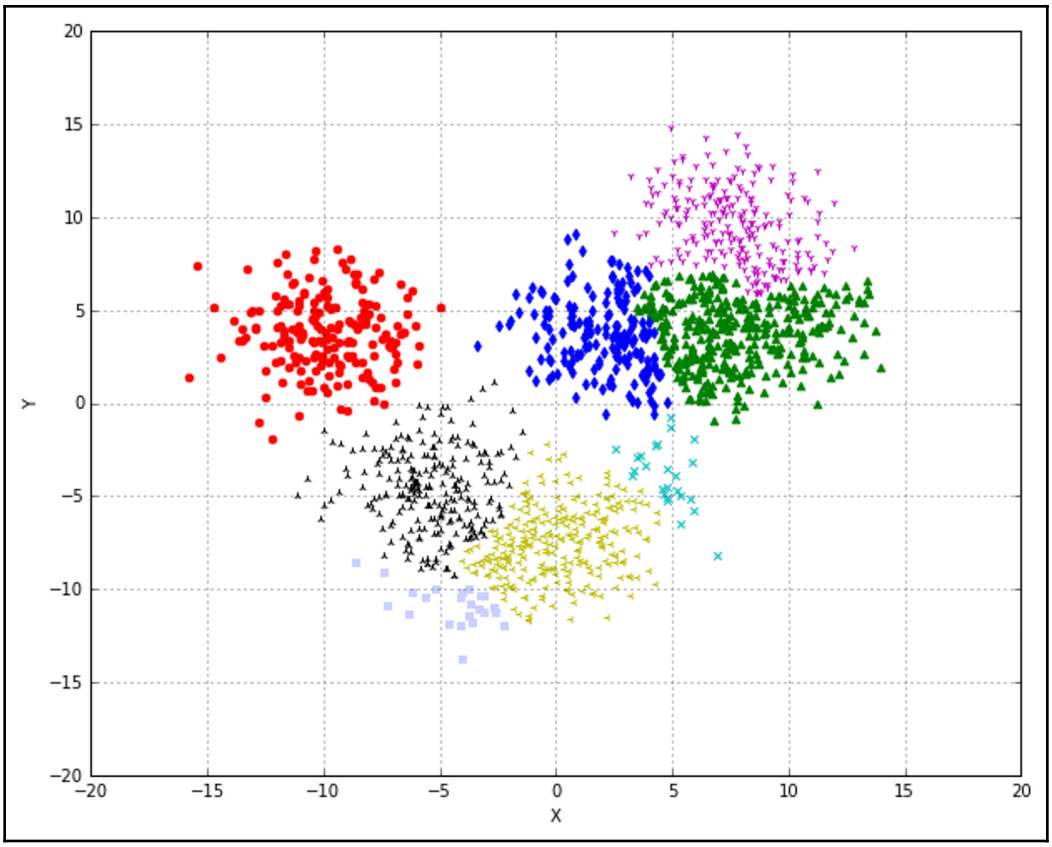


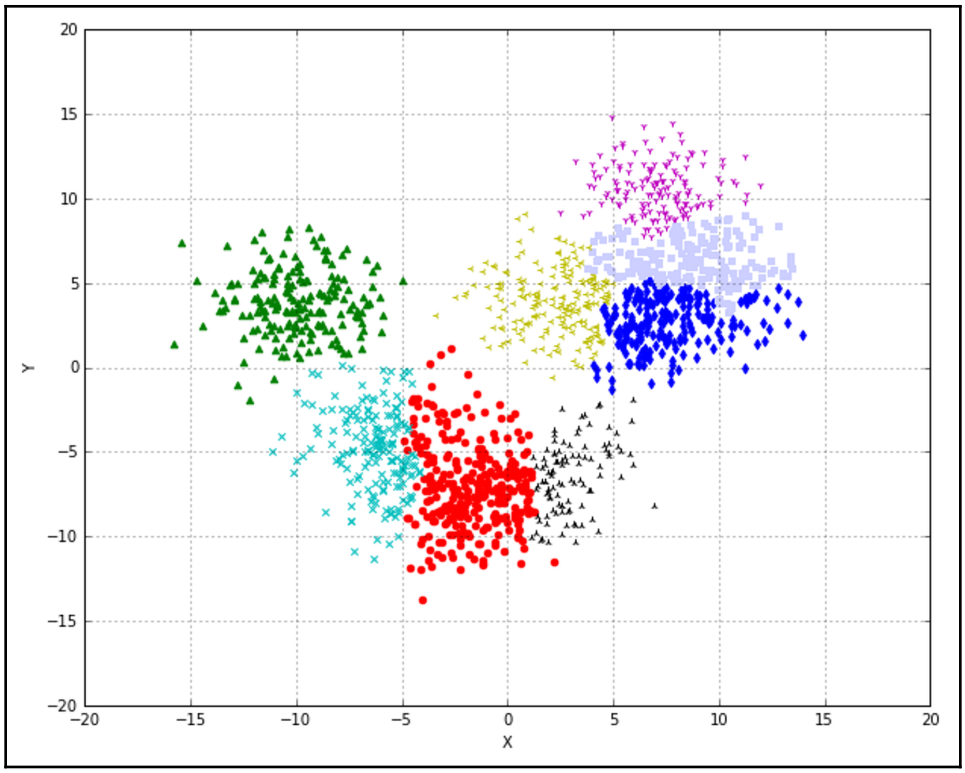

15
10
5
0
−5
−10
−15
Y
−15
−10
−5
0
5
10
X

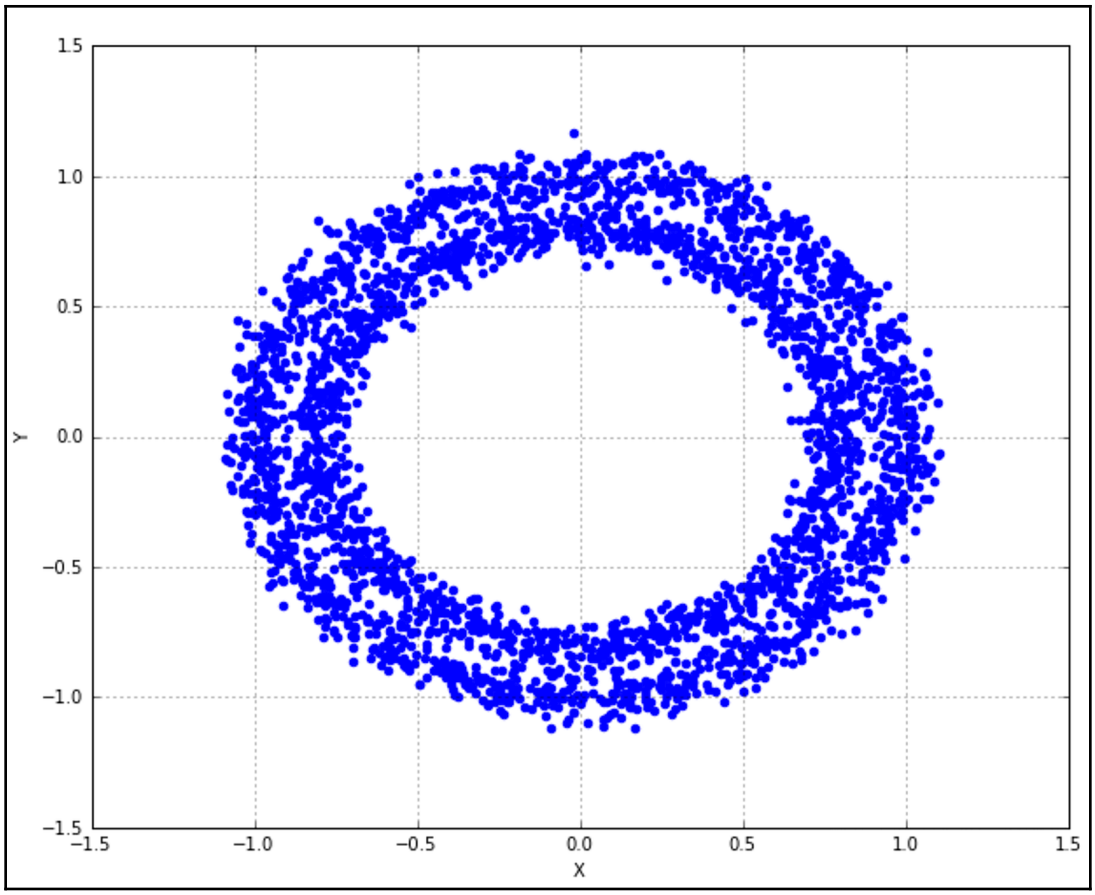

40
30
20
10
0
15
10
7
11
17
18
13
23
24
0
22
1
12
4
14
6
8
16
21
2
5
3
9
19
20

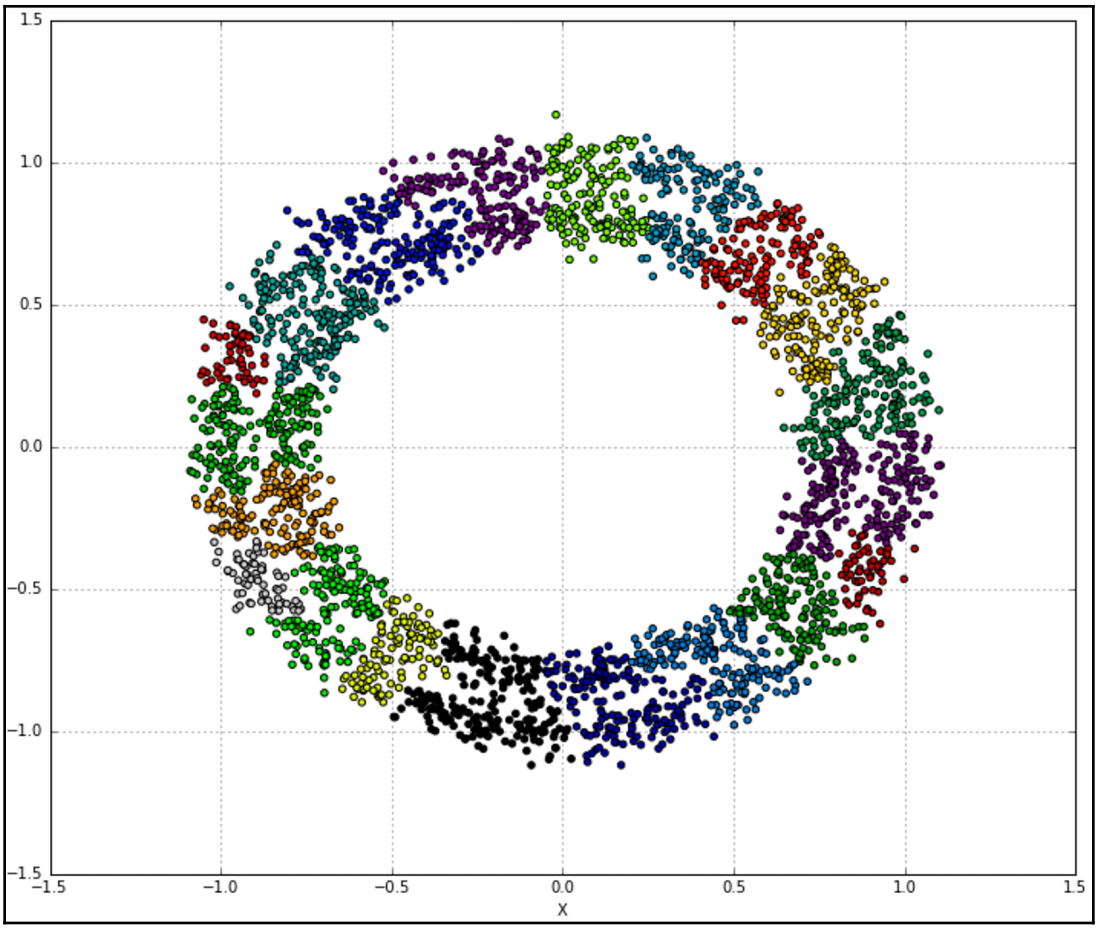

20
15
10
5
0
−5
−10
−15
−20
Y
−20
−15
−10
−5
0
5
10
15
20
X

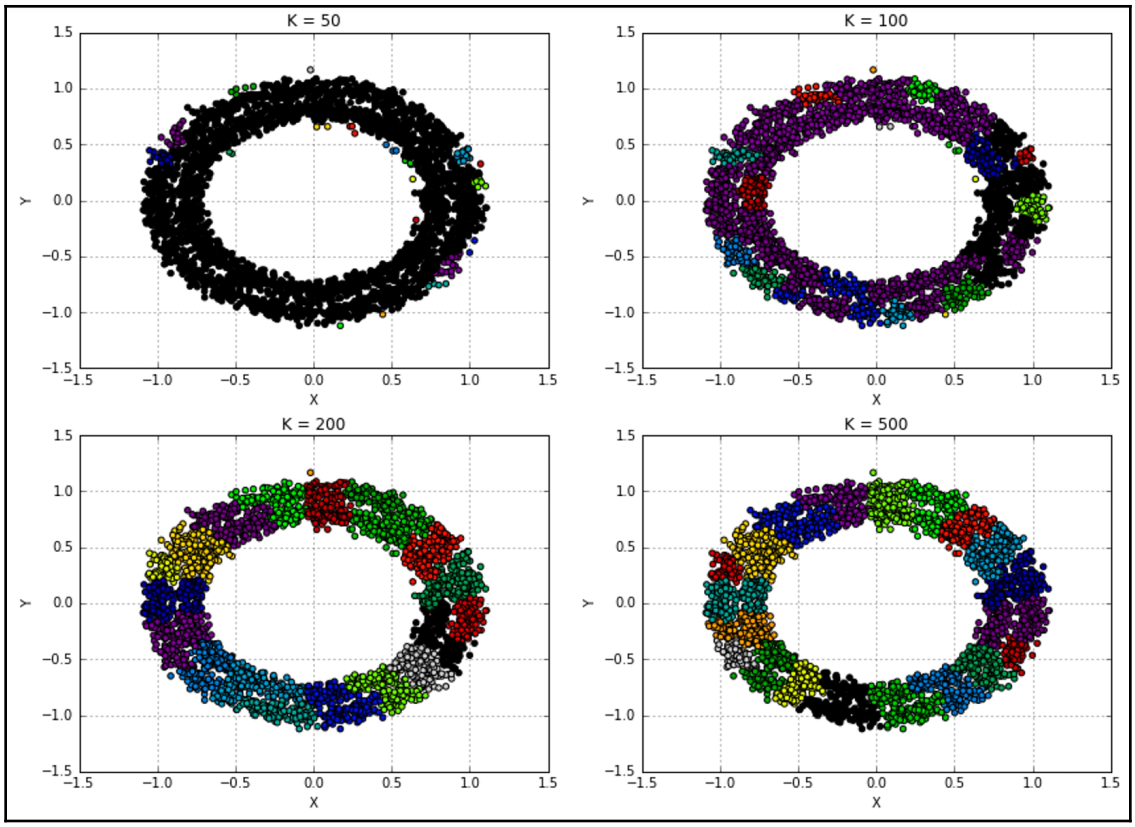

20
15
10
5
0
−5
−10
−15
−20
Y
−20
−15
−10
−5
0
5
10
15
20
X


20
15
10
5
0
−5
−10
−15
−20
Y
−20
−15
−10
−5
0
5
10
15
20
X

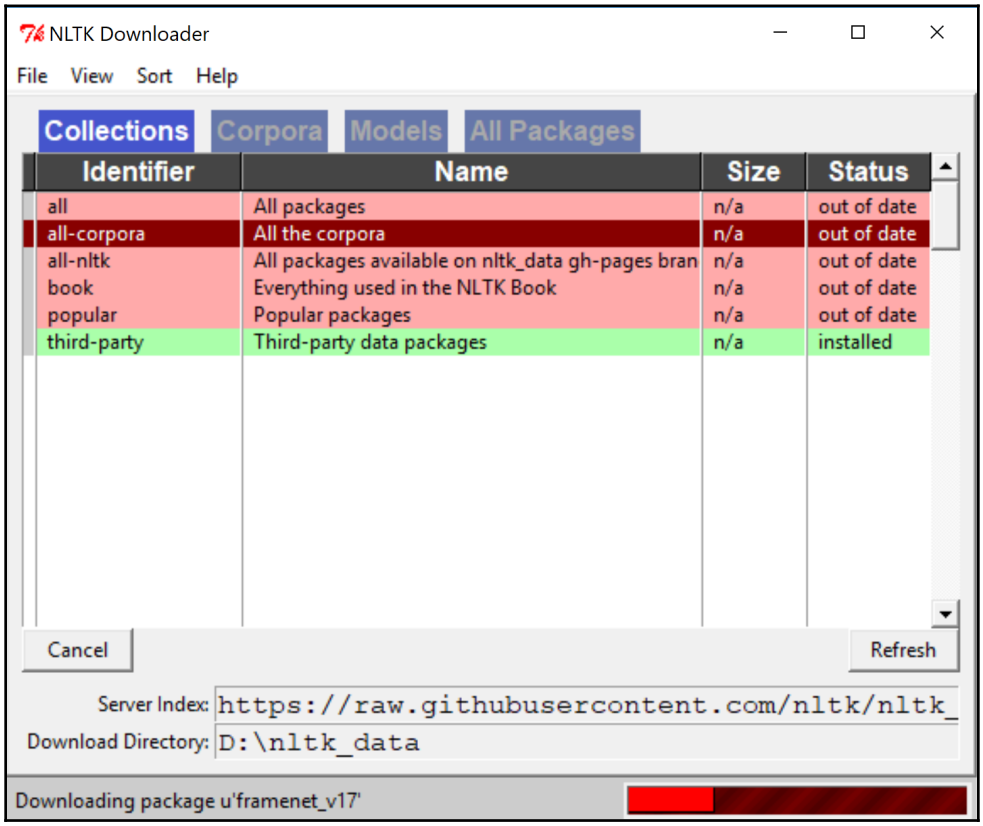

20
15
10
5
0
−5
−10
−15
−20
Y
−20
−15
−10
−5
0
5
10
15
20
X

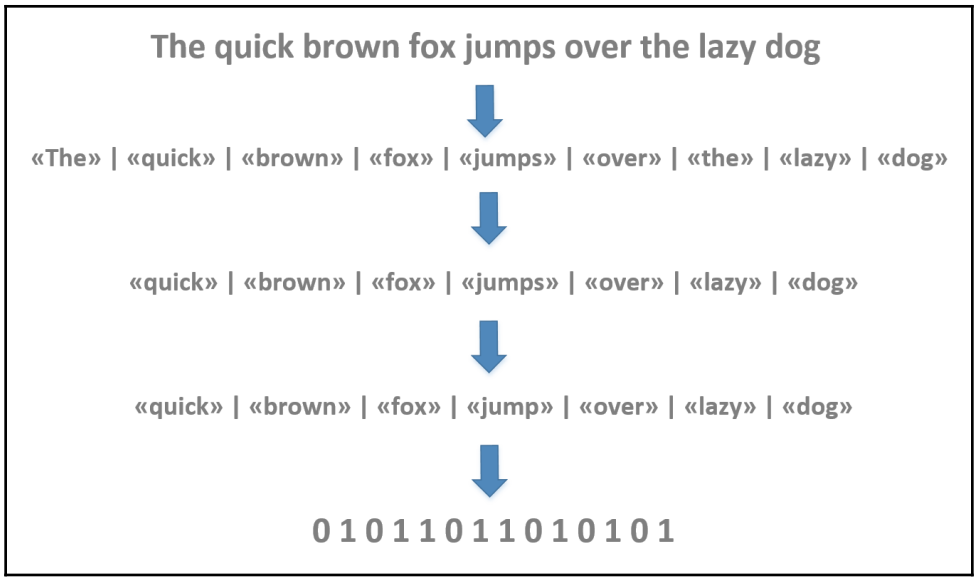

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
X


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
X

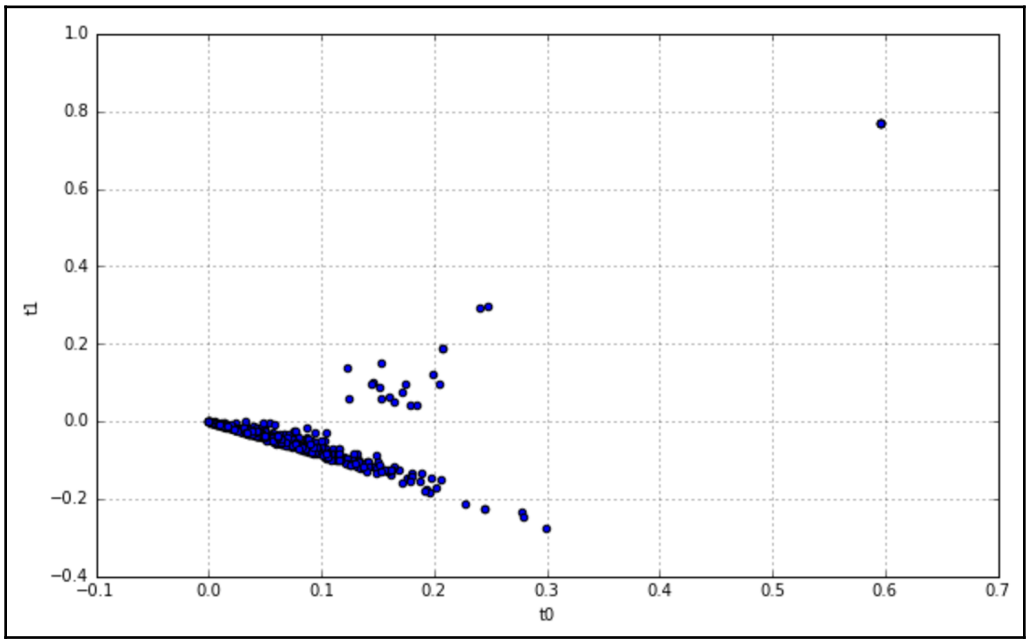

K = 50
K = 100
K = 200
K = 500
X
Y
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

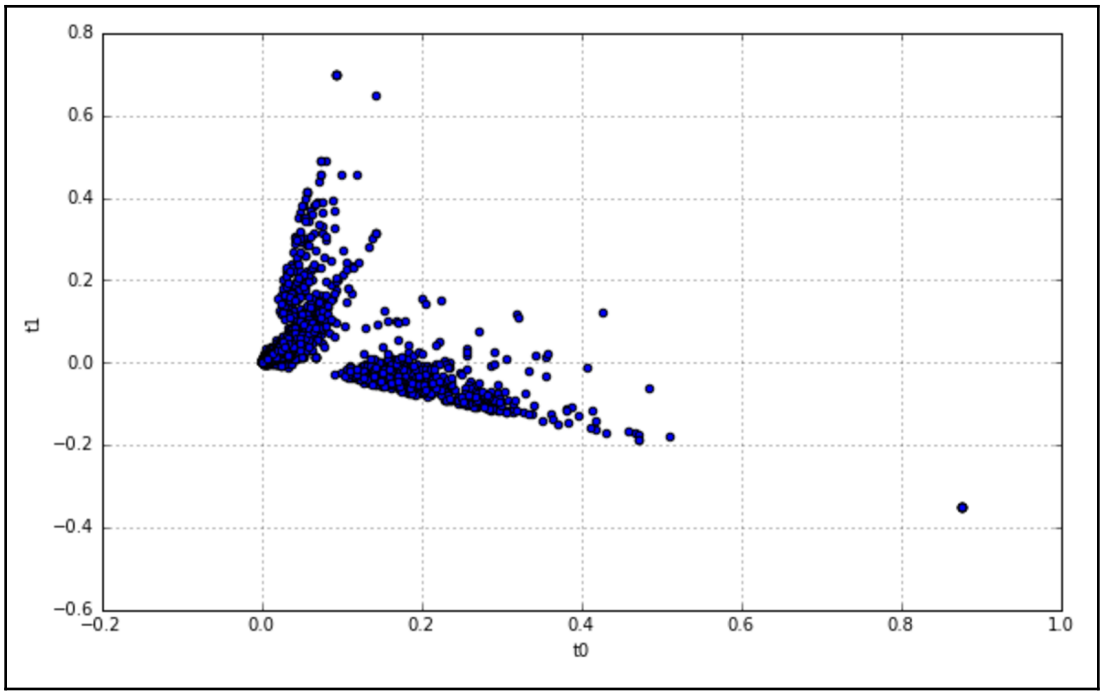
Chapter 13: Introducing Natural Language Processing

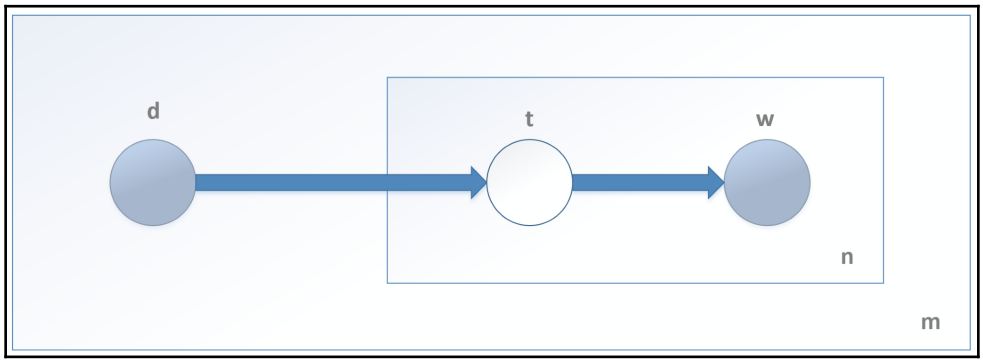


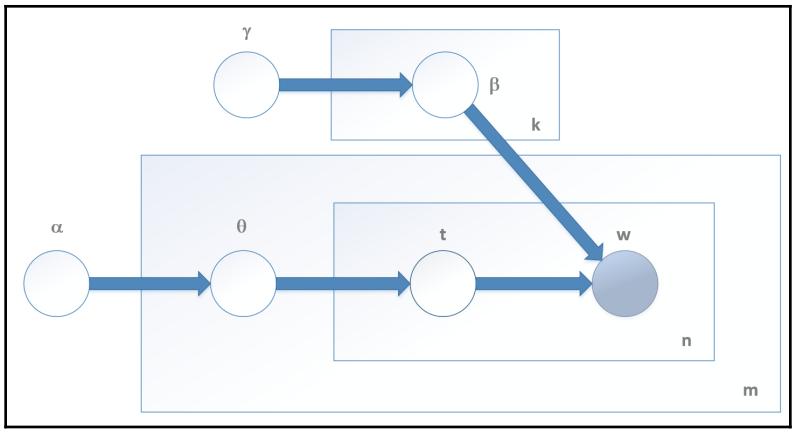

The quick brown fox jumps over the lazy dog
«The» | «quick» | «brown» | «fox» | «jumps» | «over» | «the» | «lazy» | «dog»
«quick» | «brown» | «fox» | «jumps» | «over» | «lazy» | «dog»
«quick» | «brown» | «fox» | «jump» | «over» | «lazy» | «dog»
0 1 0 1 1 0 1 1 0 1 0 1 0 1

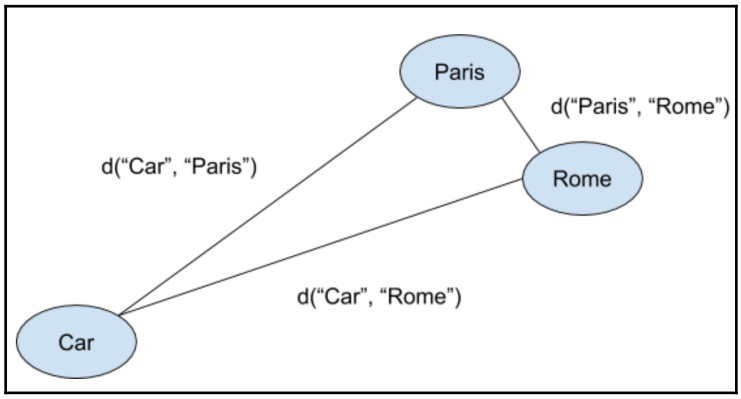
Chapter 14: Topic Modeling and Sentiment Analysis in NLP

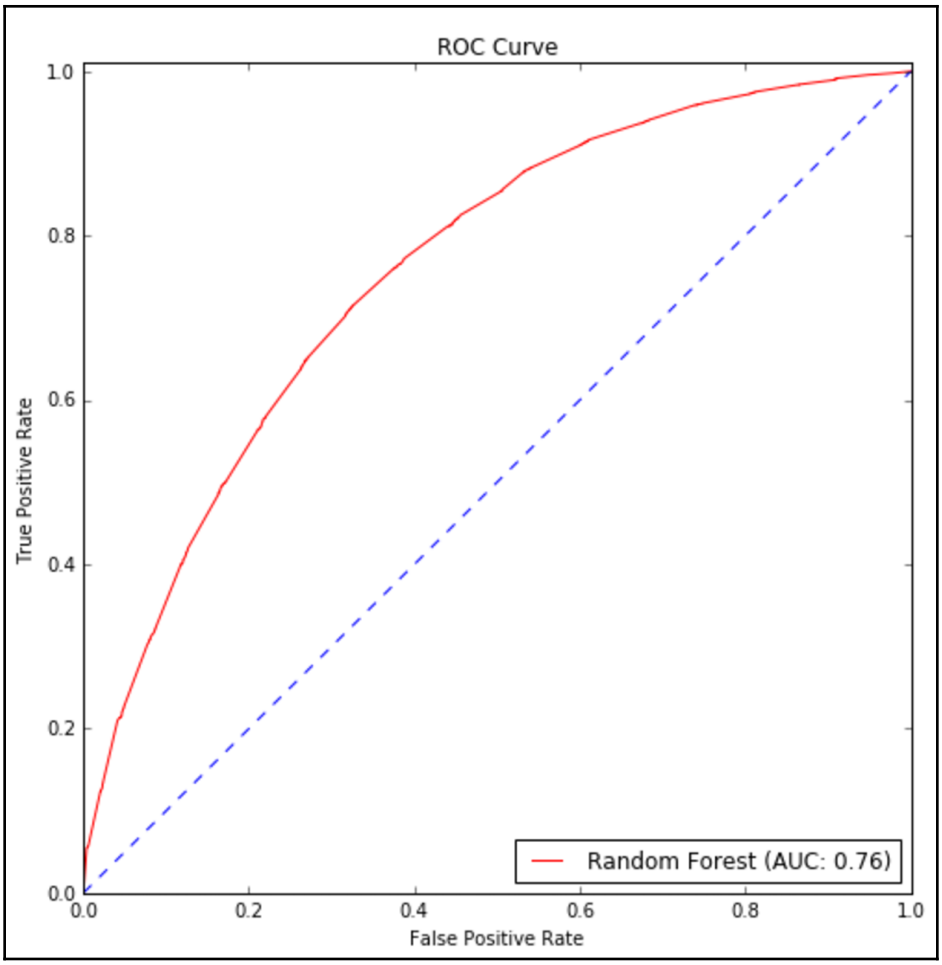



0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
t1
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
t0

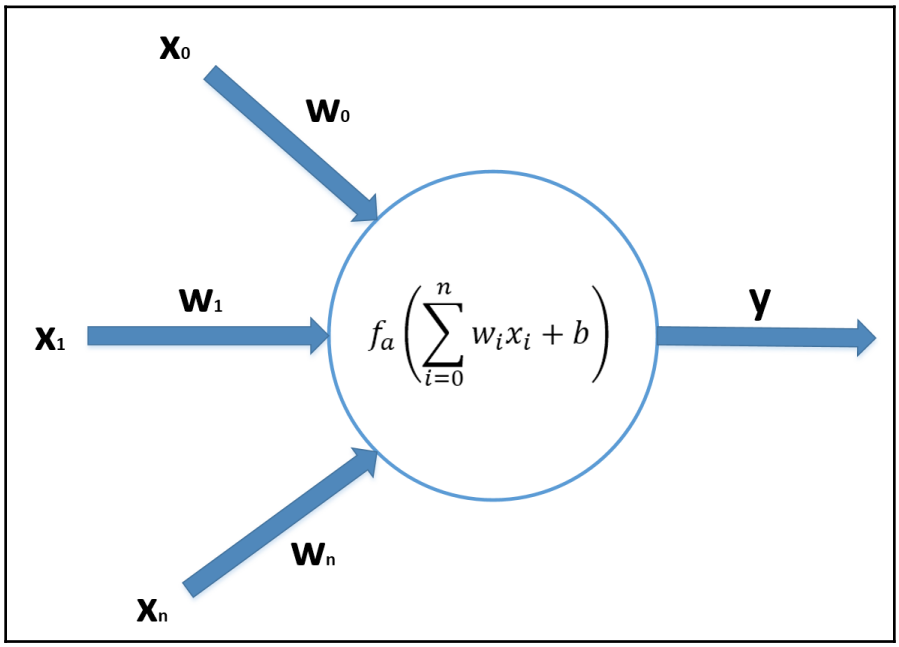

d
t
w
n
m

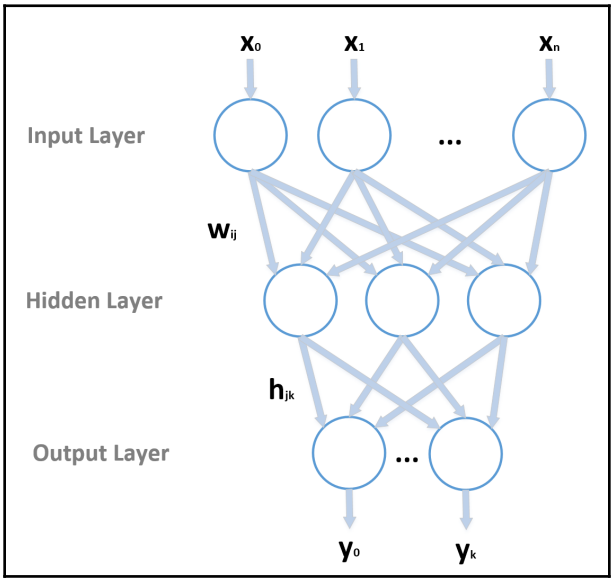

γ
β
k
α
θ
t
w
n
m

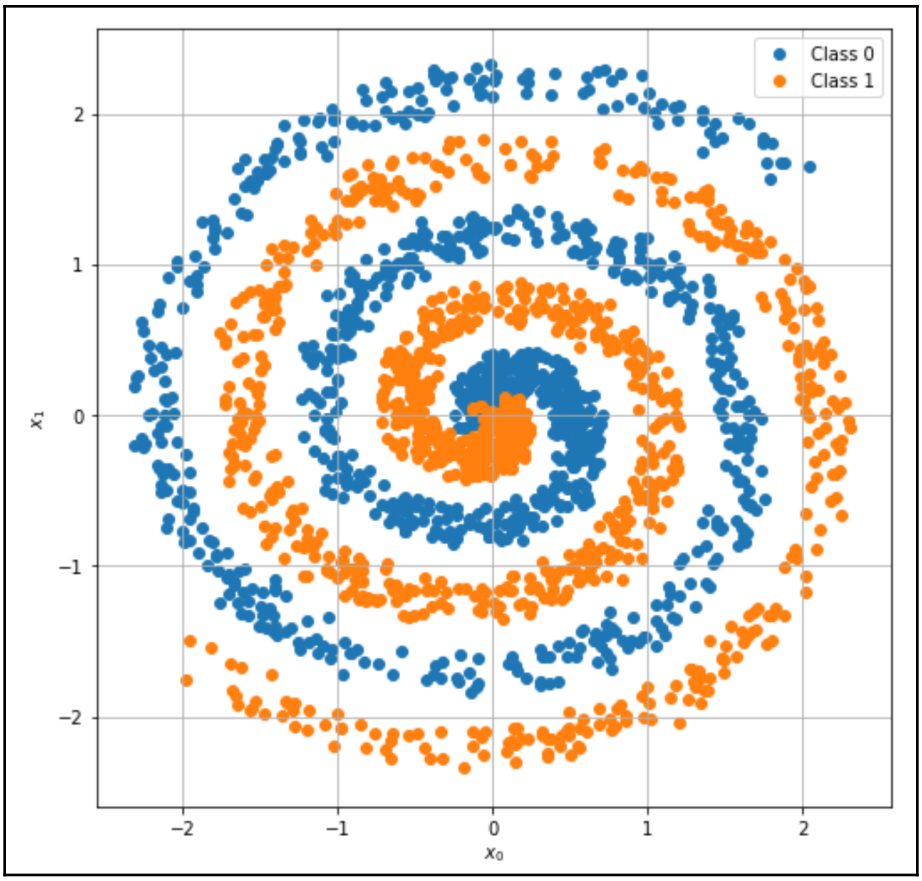

Paris
d("Paris", "Rome")
d("Car", "Paris")
Rome
d("Car", "Rome")
Car

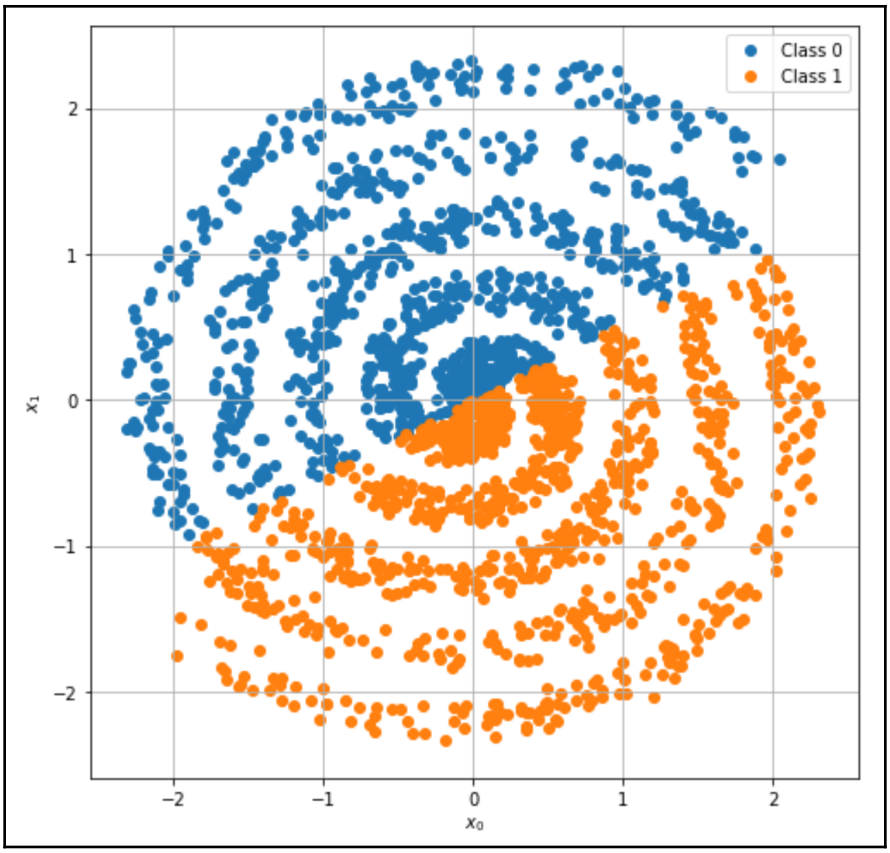

ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest (AUC: 0.76)

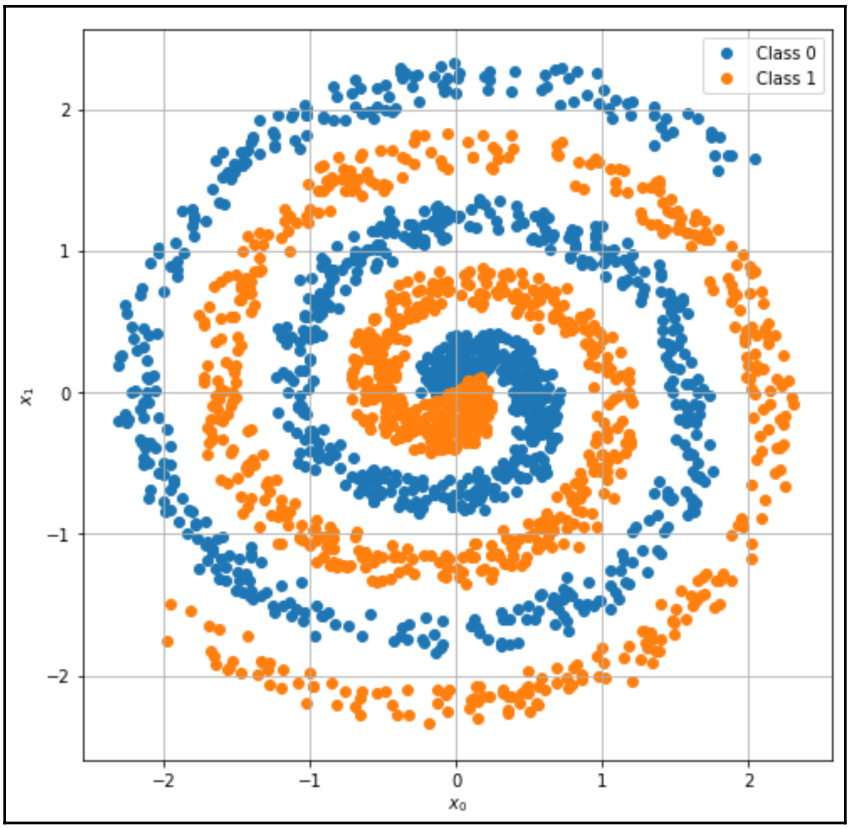
Chapter 15: Introducing Neural Networks

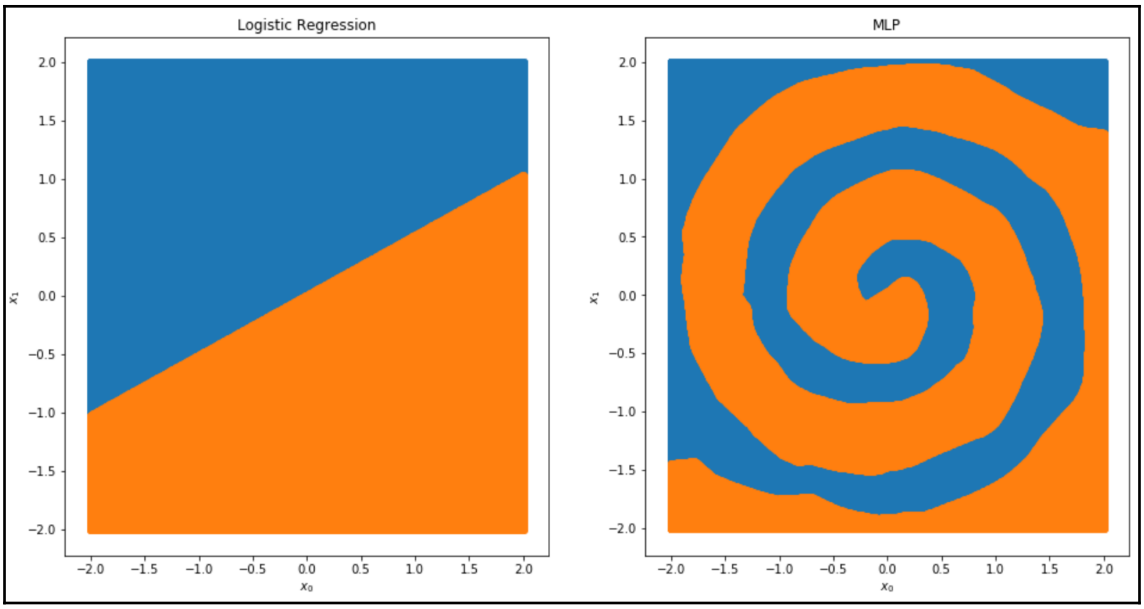




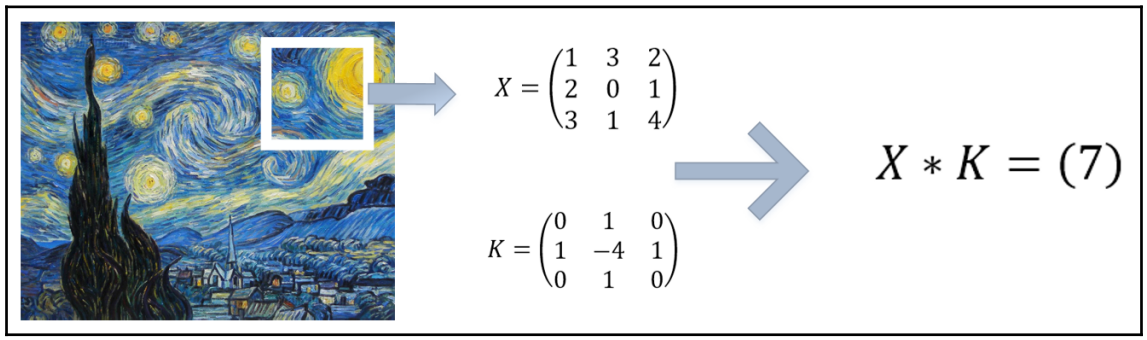

Class 0
Class 1
2
1
0
−1
−2
x_1
−2
−1
0
1
2
x_0

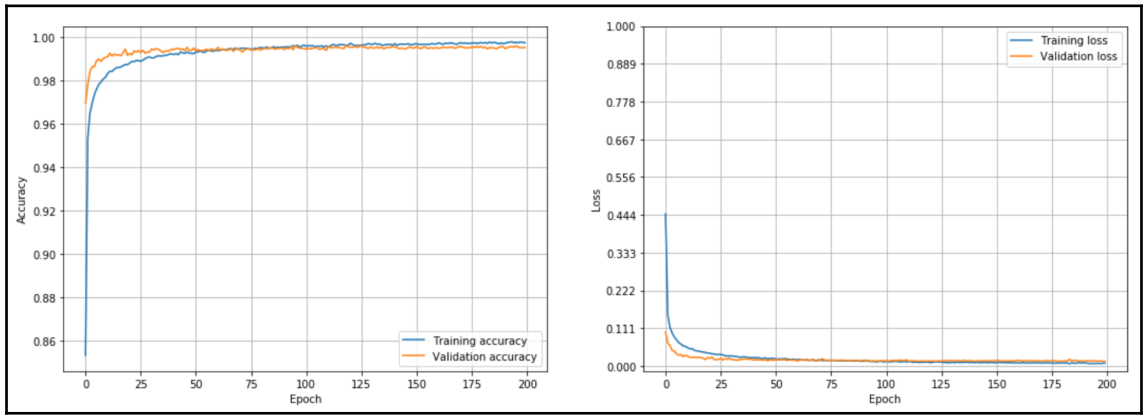

Class 0
Class 1
2
1
0
−1
−2
x_1
−2
−1
0
1
2
x_0

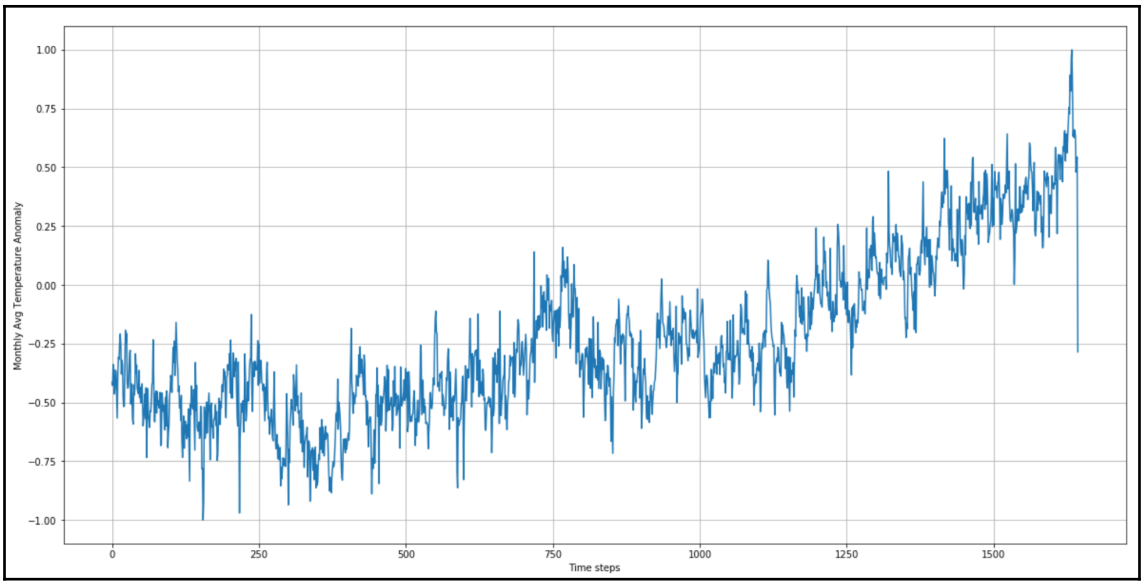

Class 0
Class 1
2
1
0
−1
−2
x_1
−2
−1
0
1
2
x_0

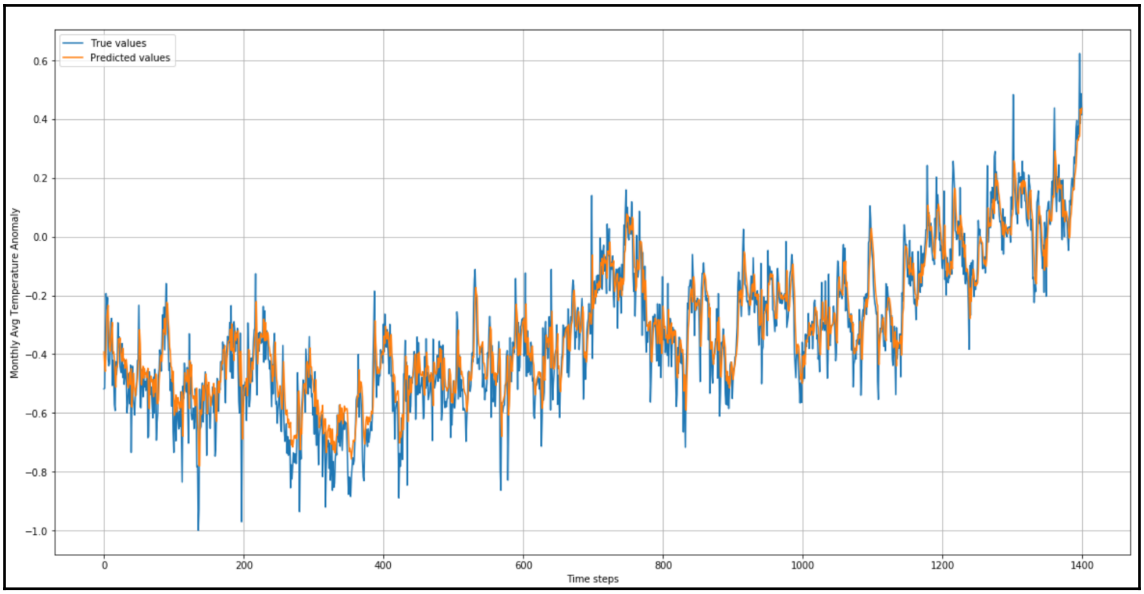

Logistic Regression
MLP
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
x_1
x_0

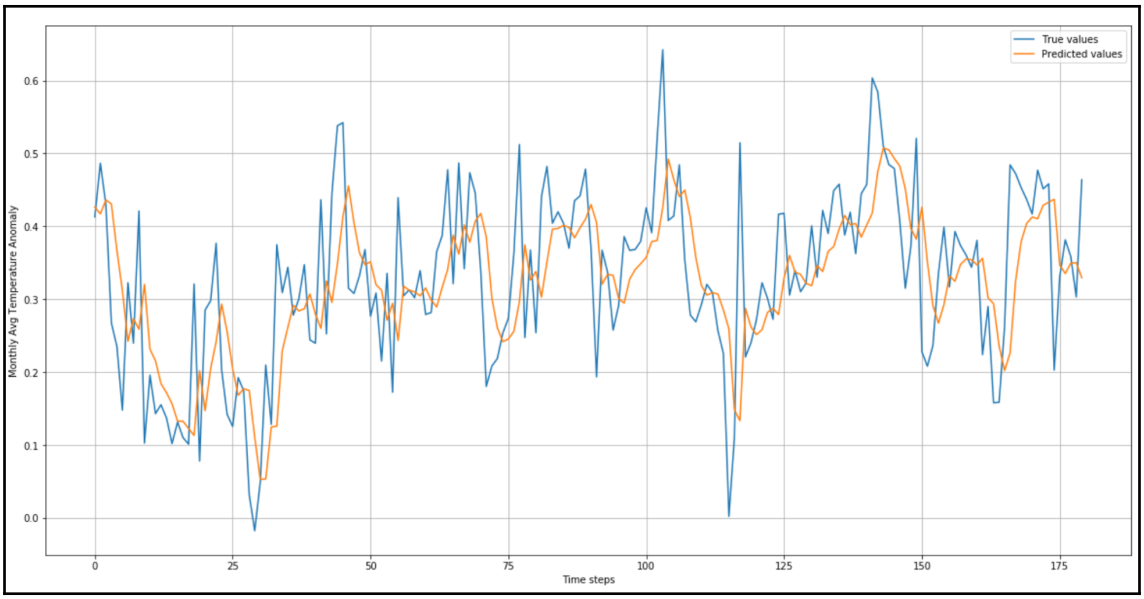
Chapter 16: Advanced Deep Learning Models

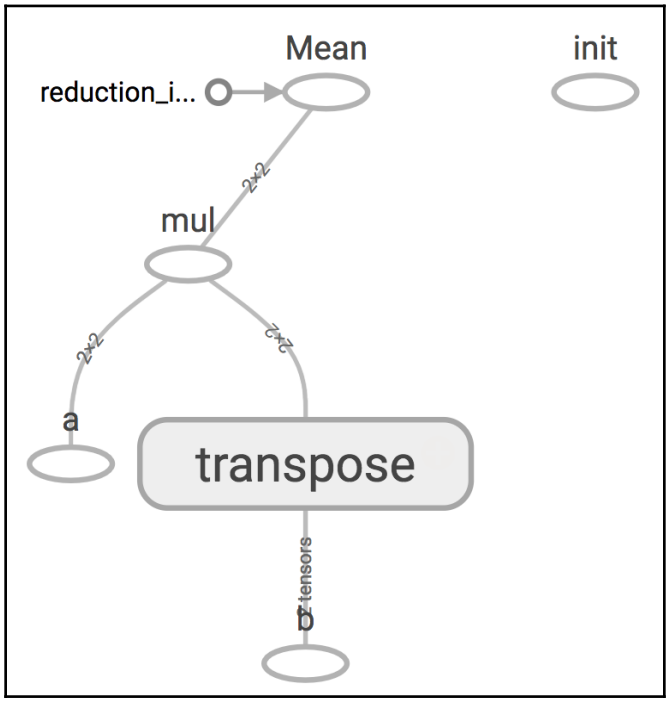


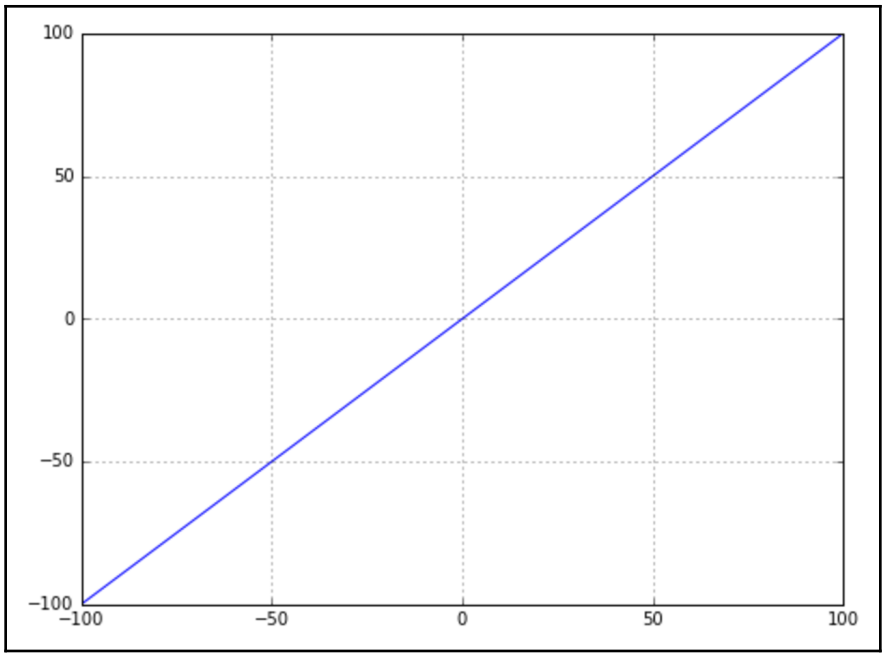


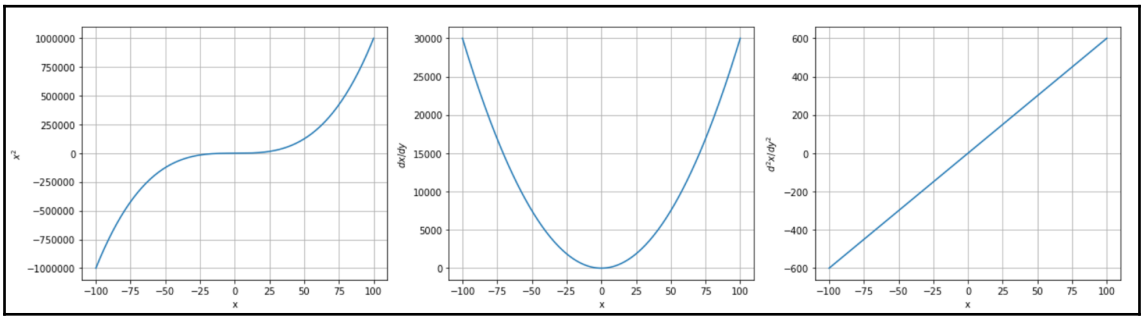

Monthly Avg Temperature Anomaly
Time steps
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
250
500
750
1000
1250
1500

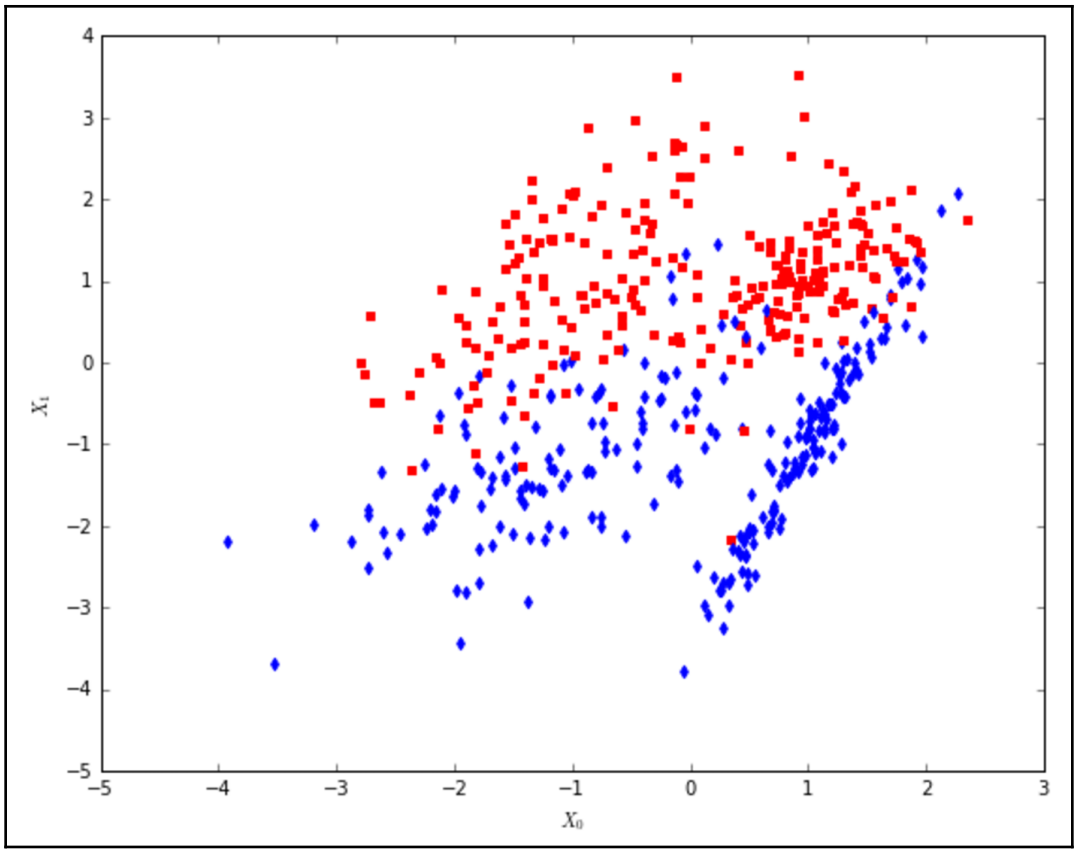

True values
Predicted values
Monthly Avg Temperature Anomaly
Time steps
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
0
200
400
600
800
1000
1200
1400

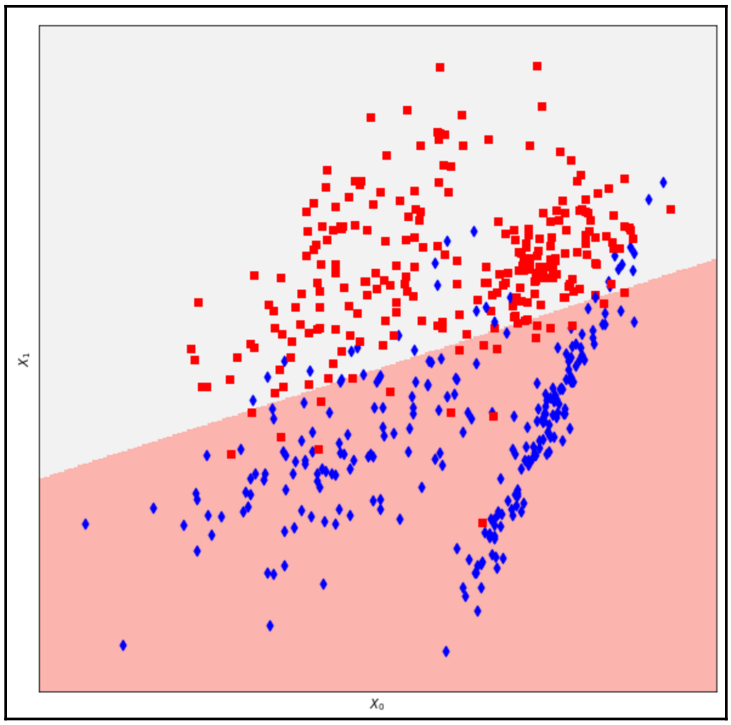

True values
Predicted values
Monthly Avg Temperature Anomaly
Time steps
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
25
50
75
100
125
150
175

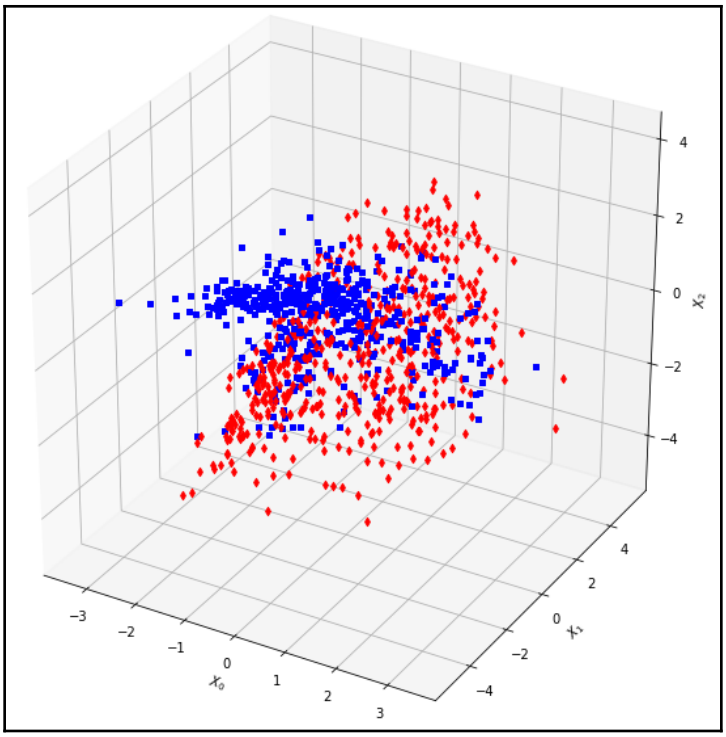

Mean
init
reduction_i...
2×2
mul
2×2
2×2
a
transpose
2 tensors


100
50
0
−50
−100
−100
−50
0
50
100

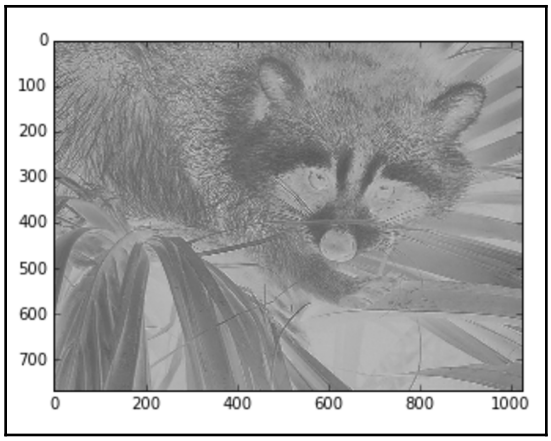

1000000
750000
500000
250000
0
−250000
−500000
−750000
−1000000
x^3
−100 −75 −50 −25 0 25 50 75 100
x
30000
25000
20000
15000
10000
5000
0
dx/dy
−100 −75 −50 −25 0 25 50 75 100
x
600
400
200
0
−200
−400
−600
d^2x/dy^2
−100 −75 −50 −25 0 25 50 75 100
x


4
3
2
1
0
−1
−2
−3
−4
−5
X_1
−5
−4
−3
−2
−1
0
1
2
3
X_0

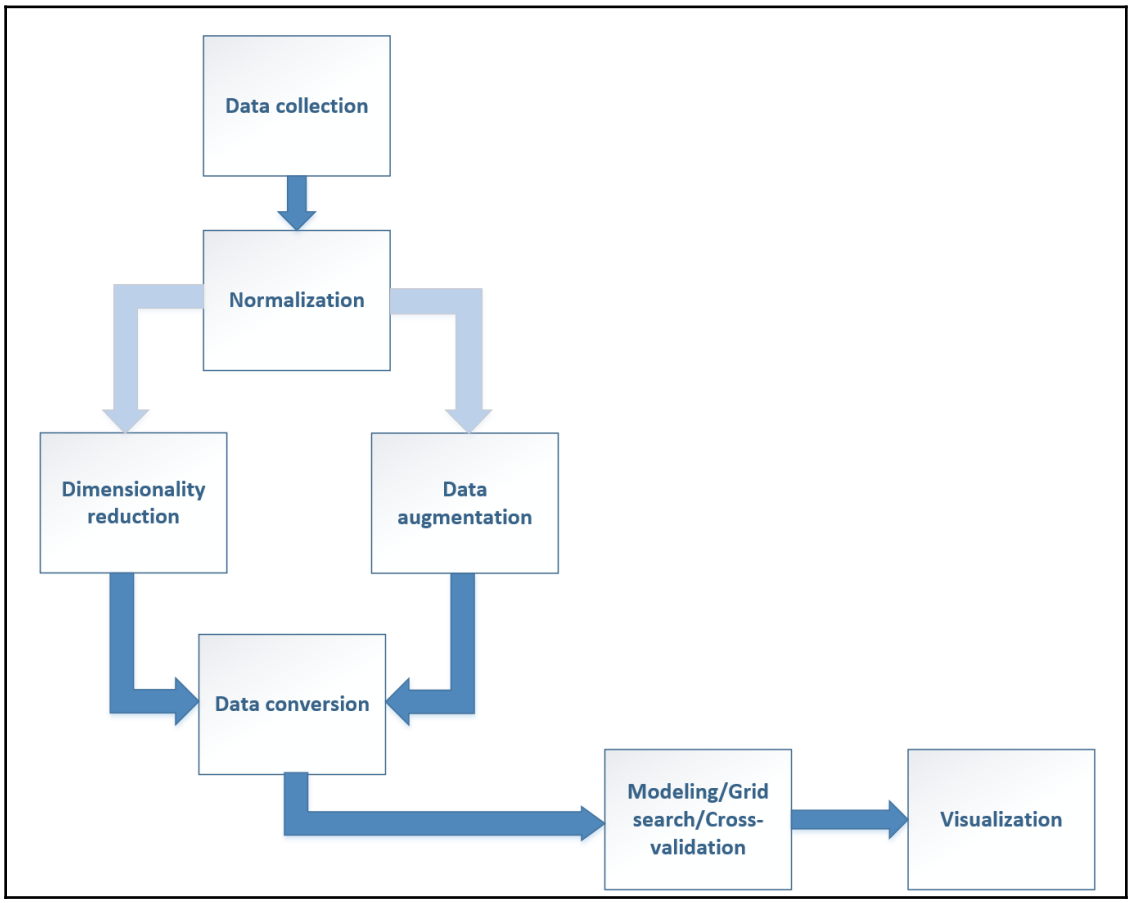

x_1
x_0

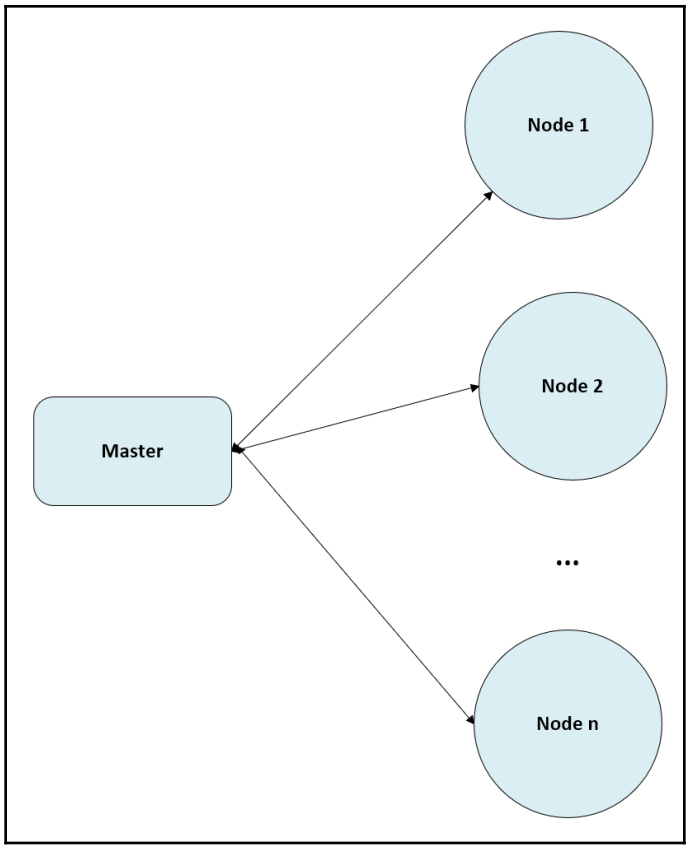

4
2
0
−2
−4
X_2
−3
−2
−1
0
1
2
3
X_0
4
2
0
−2
−4
X_1


0
100
200
300
400
500
600
700
0
200
400
600
800
1000


0
100
200
300
400
500
600
700
0
200
400
600
800
1000

Chapter 17: Creating a Machine Learning Architecture




Node 1
Node 2
Master
...
Node n