Machine Learning

Supervised Learning

Data: X (n obs, p features), y (labels)

Regression, classification

Train/learn/fit f from data (model)

Score: for new x, get f(x)

Algos: LR, k-NN, DT, RF, GBM, NN/DL, SVM, NB…

Goal: max acc/min err new data

Metrics: MSE, AUC (ROC)

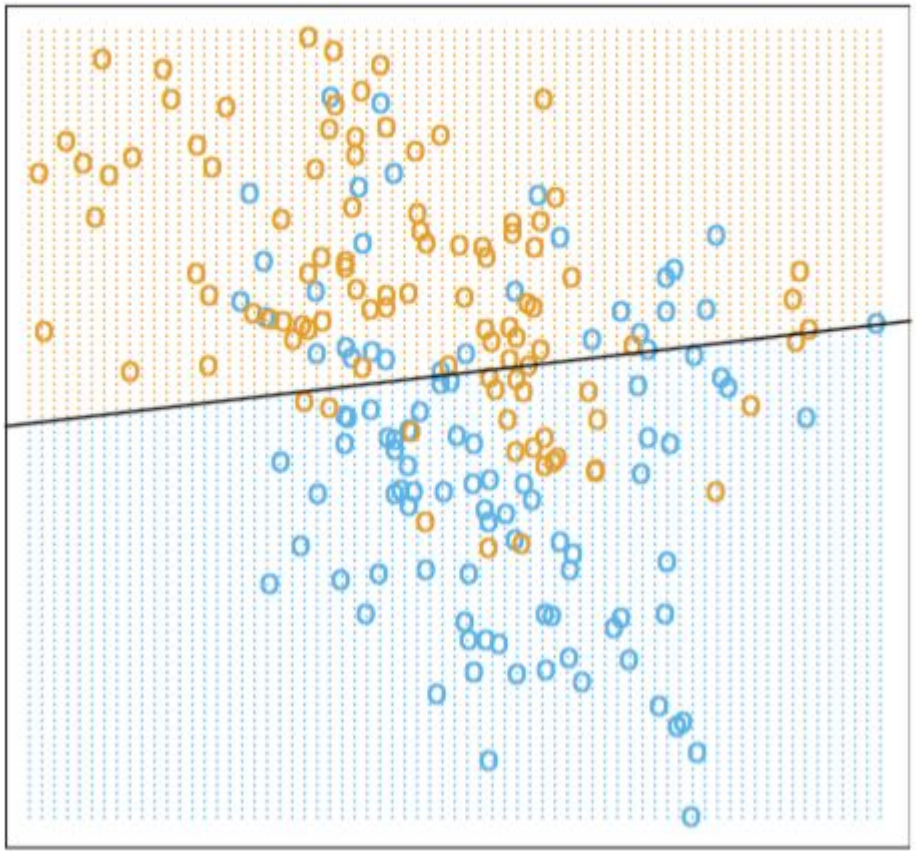
Bad: measure on train set. Need: test/CV

Hyperparams, model capacity, overfitting

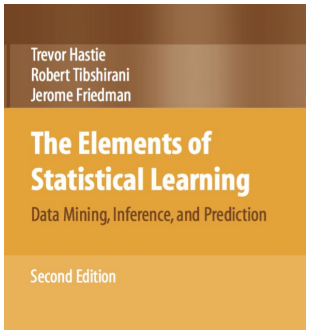
Regularization

Model selection

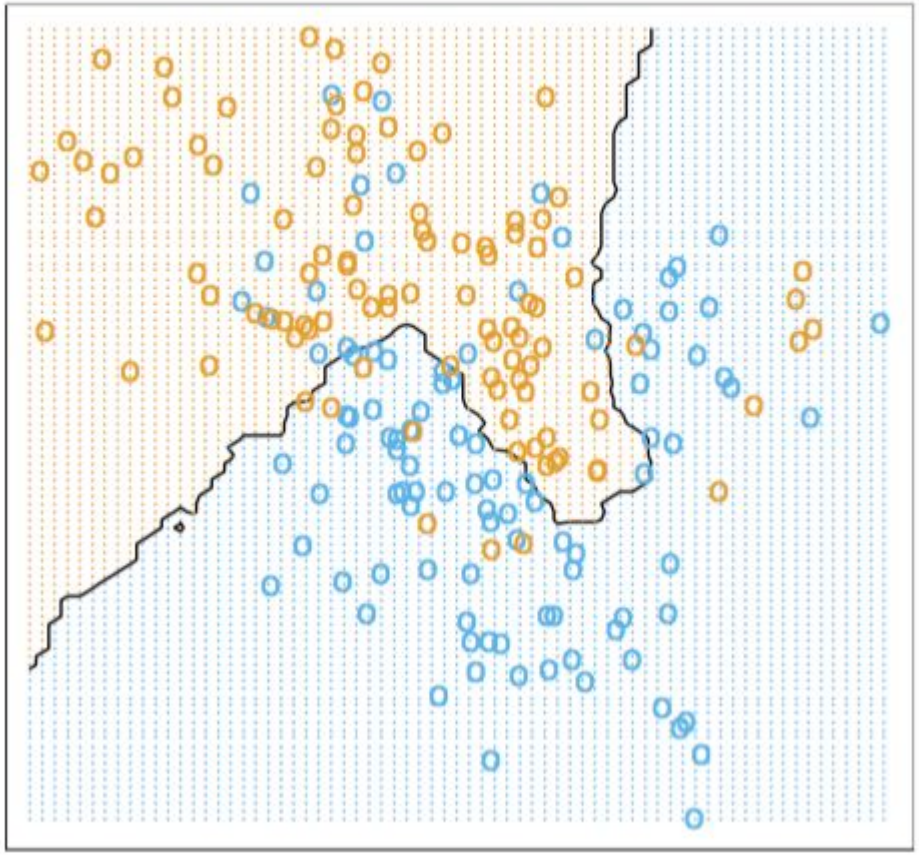
Linear Regression of 0/1 Response

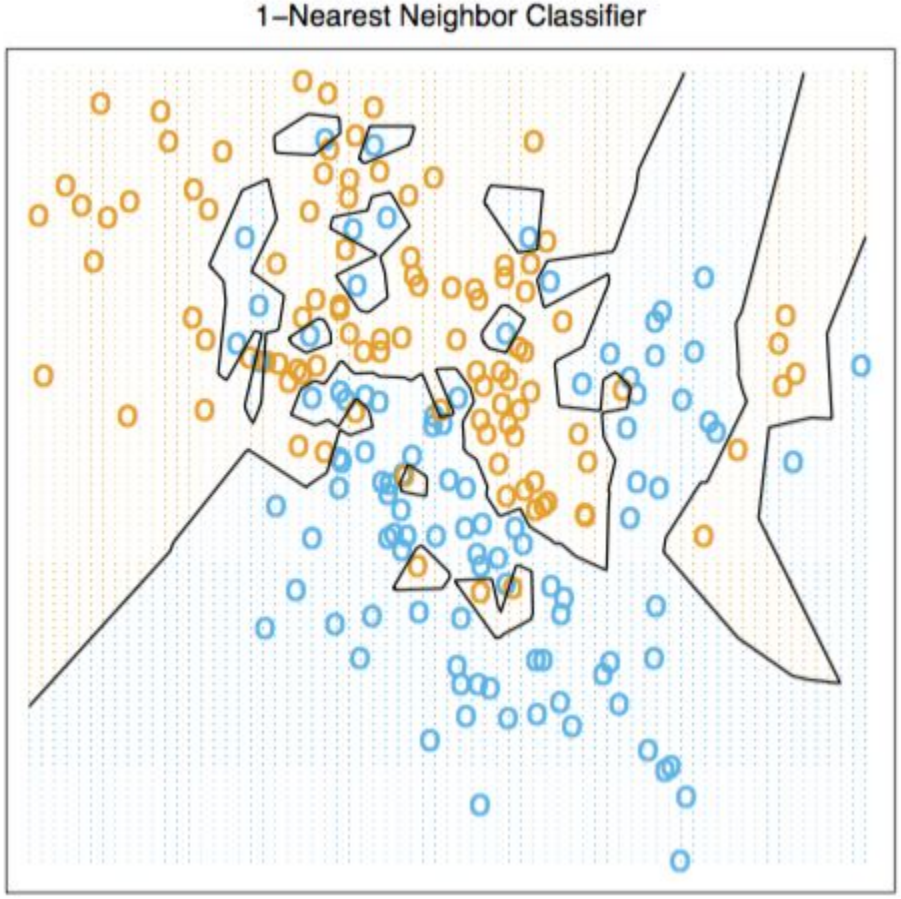


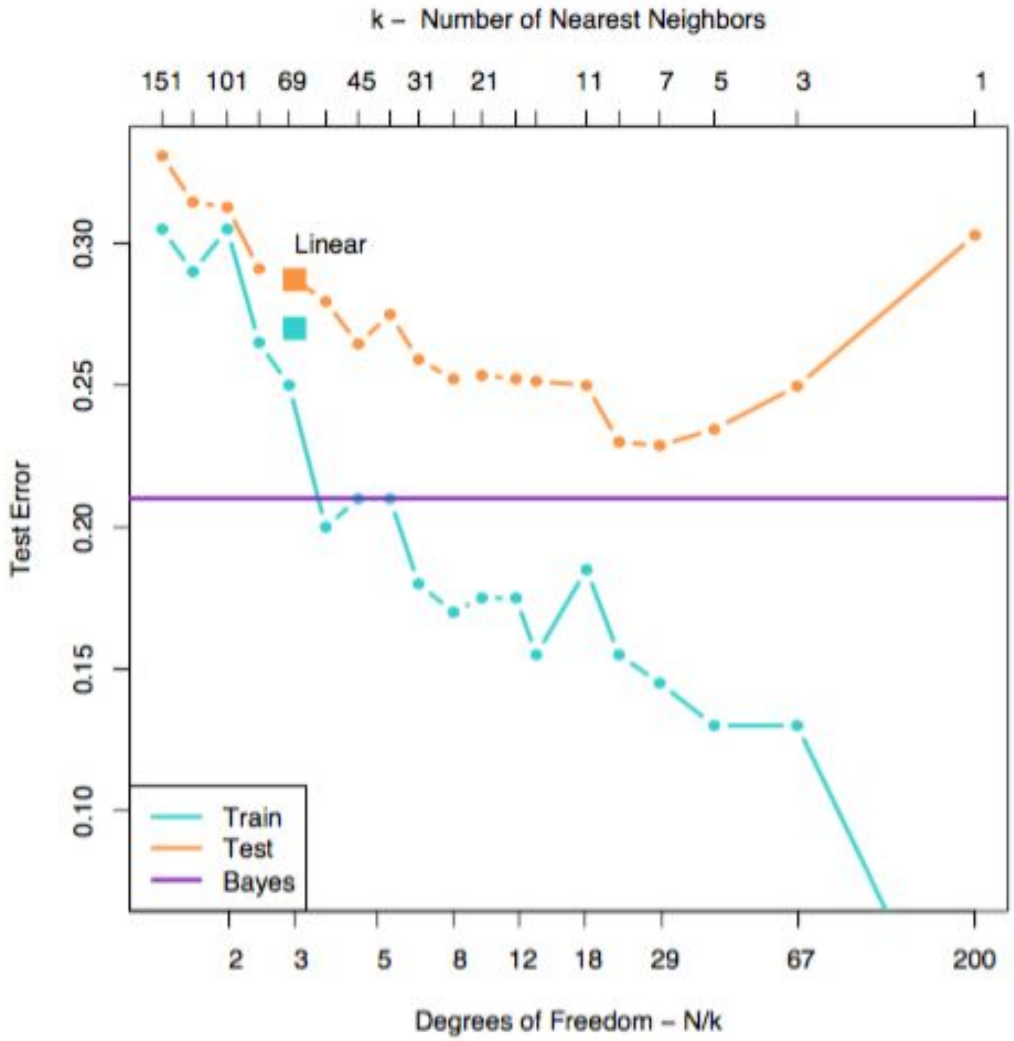
source:



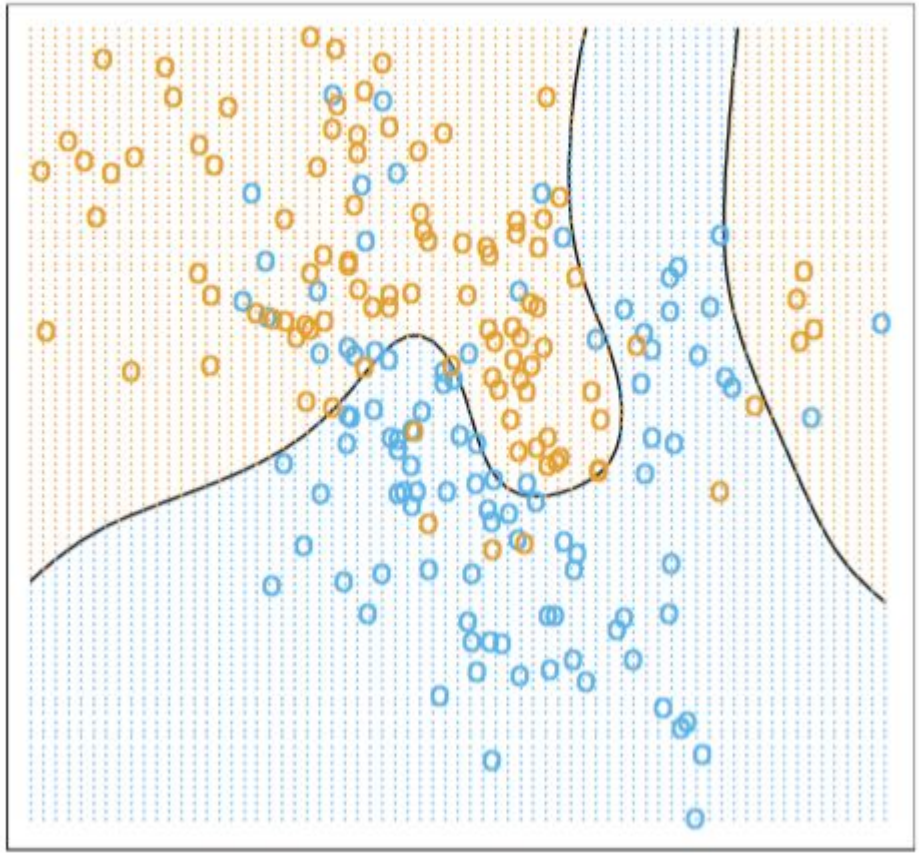
15-Nearest Neighbor Classifier

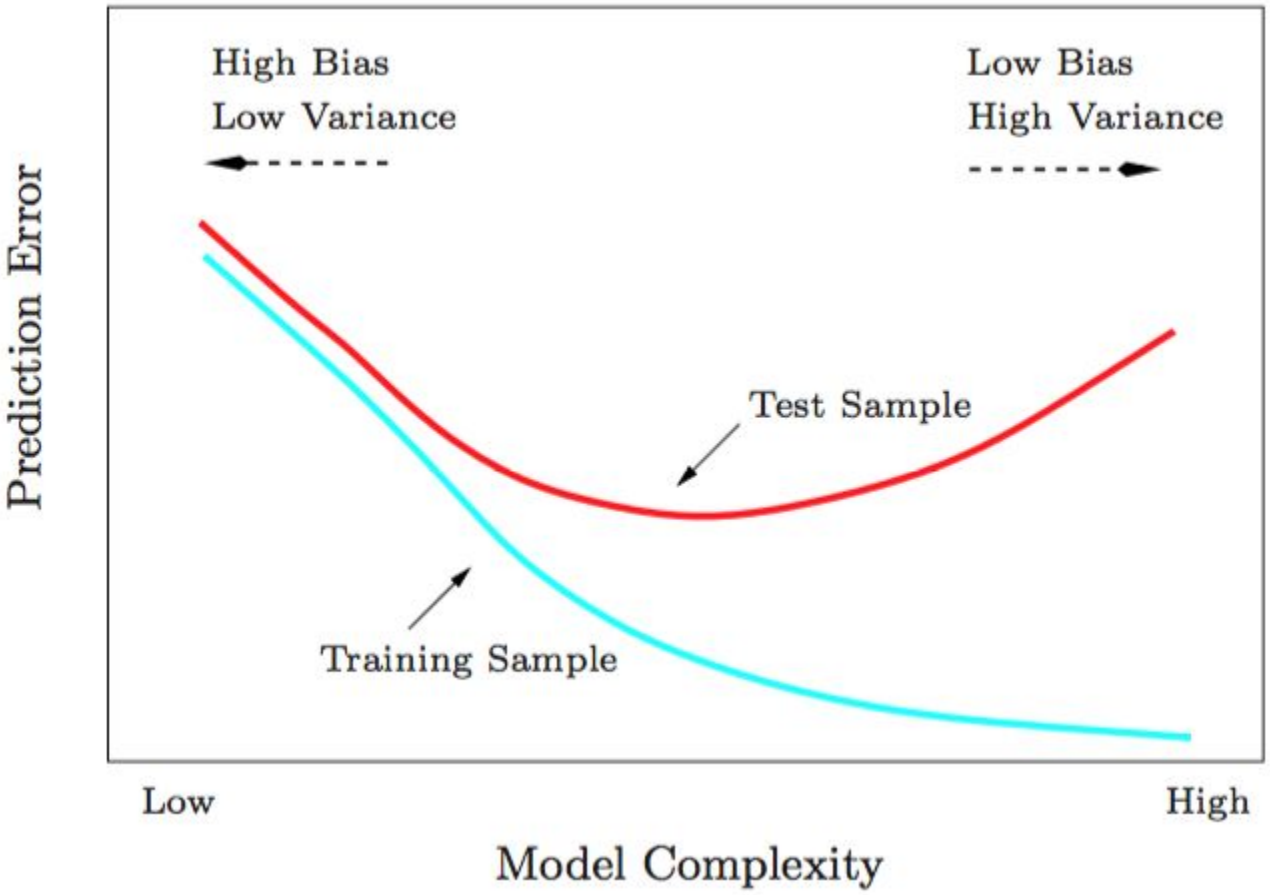


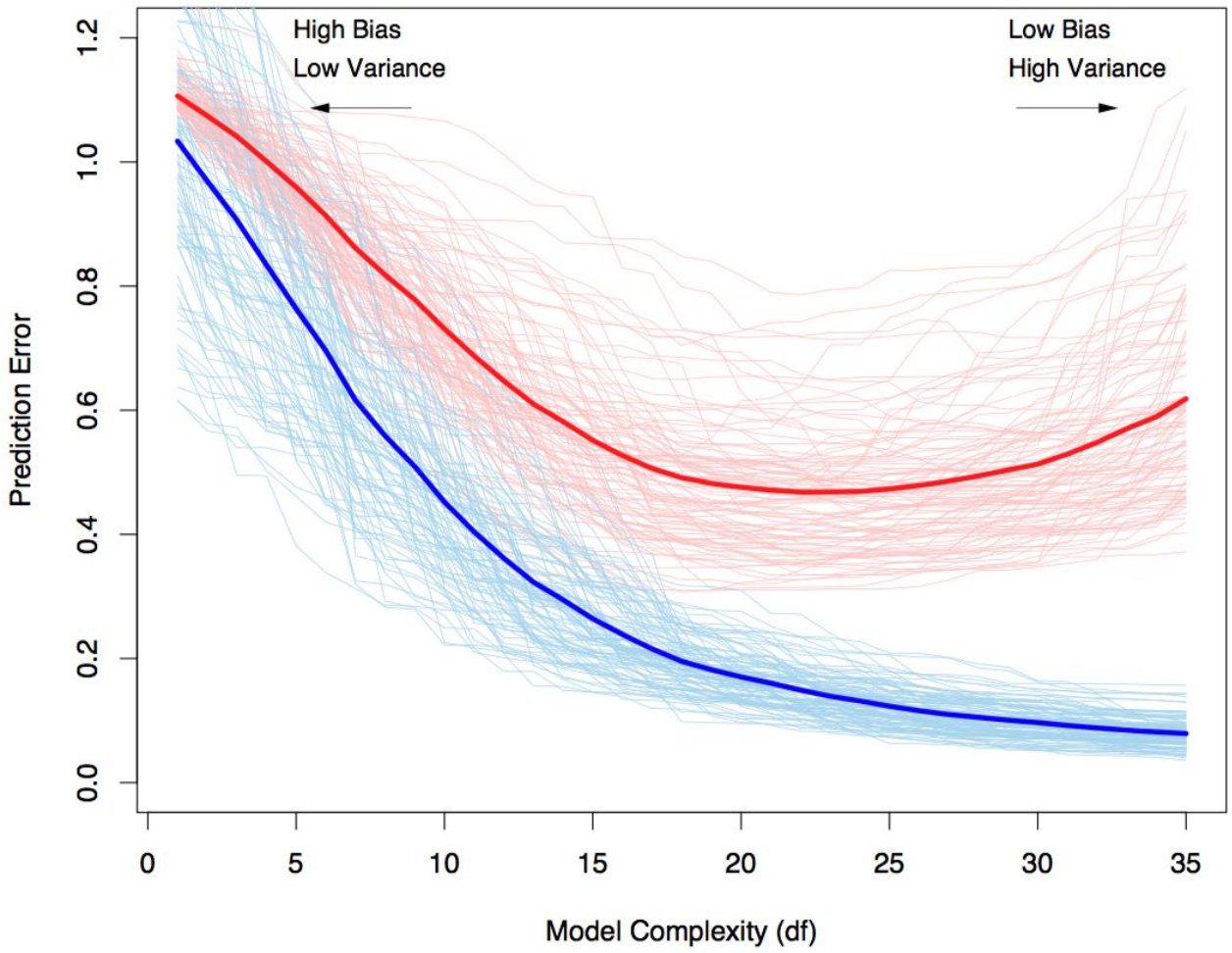

1−Nearest Neighbor Classifier

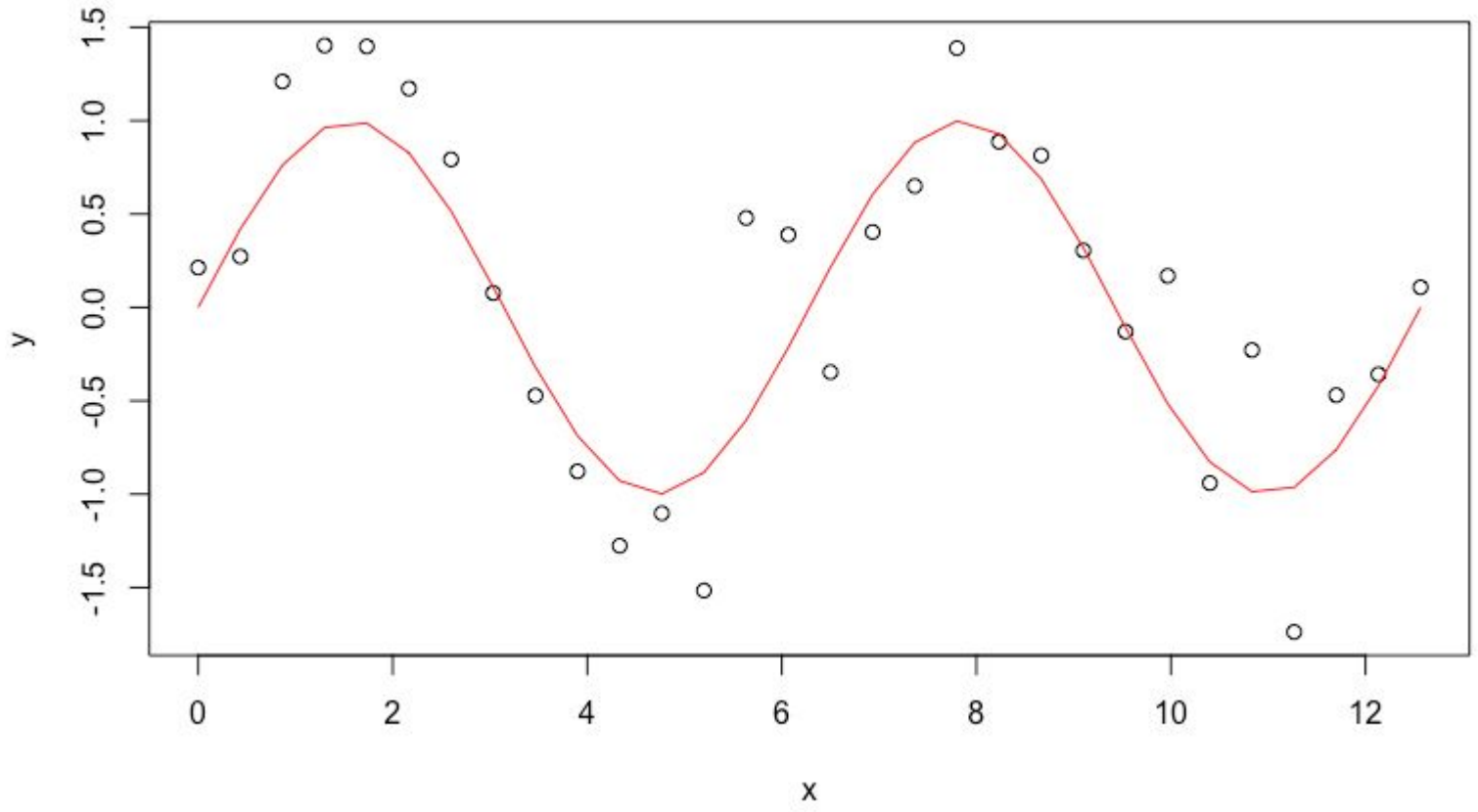

k – Number of Nearest Neighbors
151
101
69
45
31
21
11
7
5
3
1
Linear
Test Error
0.10
0.15
0.20
0.25
0.30
Train
Test
Bayes
2
3
5
8
12
18
29
67
200
Degrees of Freedom – N/k

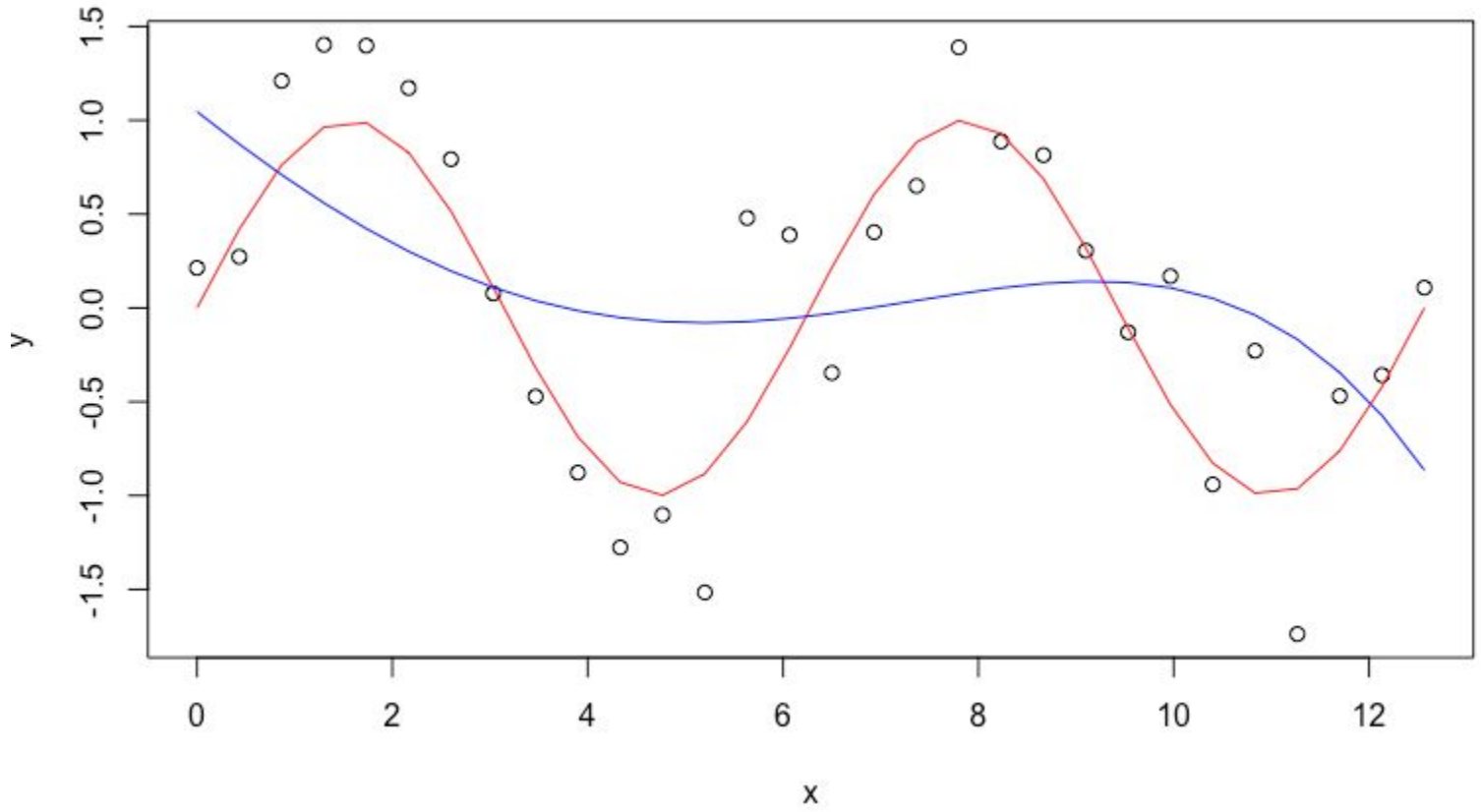
Bayes Optimal Classifier



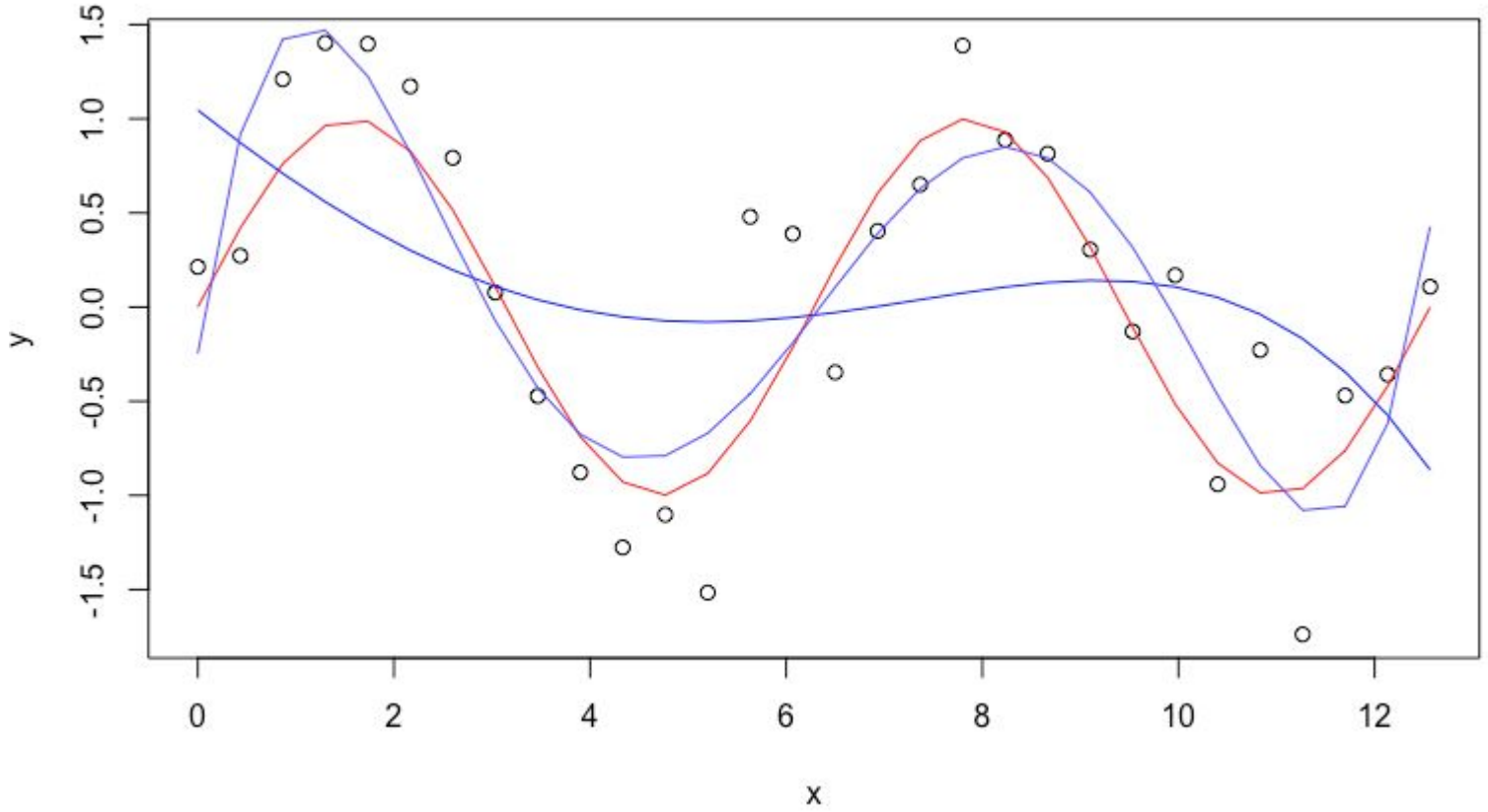

High Bias
Low Variance
Low Bias
High Variance
Prediction Error
Test Sample
Training Sample
Low
High
Model Complexity


High Bias
Low Variance
Low Bias
High Variance
Prediction Error
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25
30
35
Model Complexity (df)

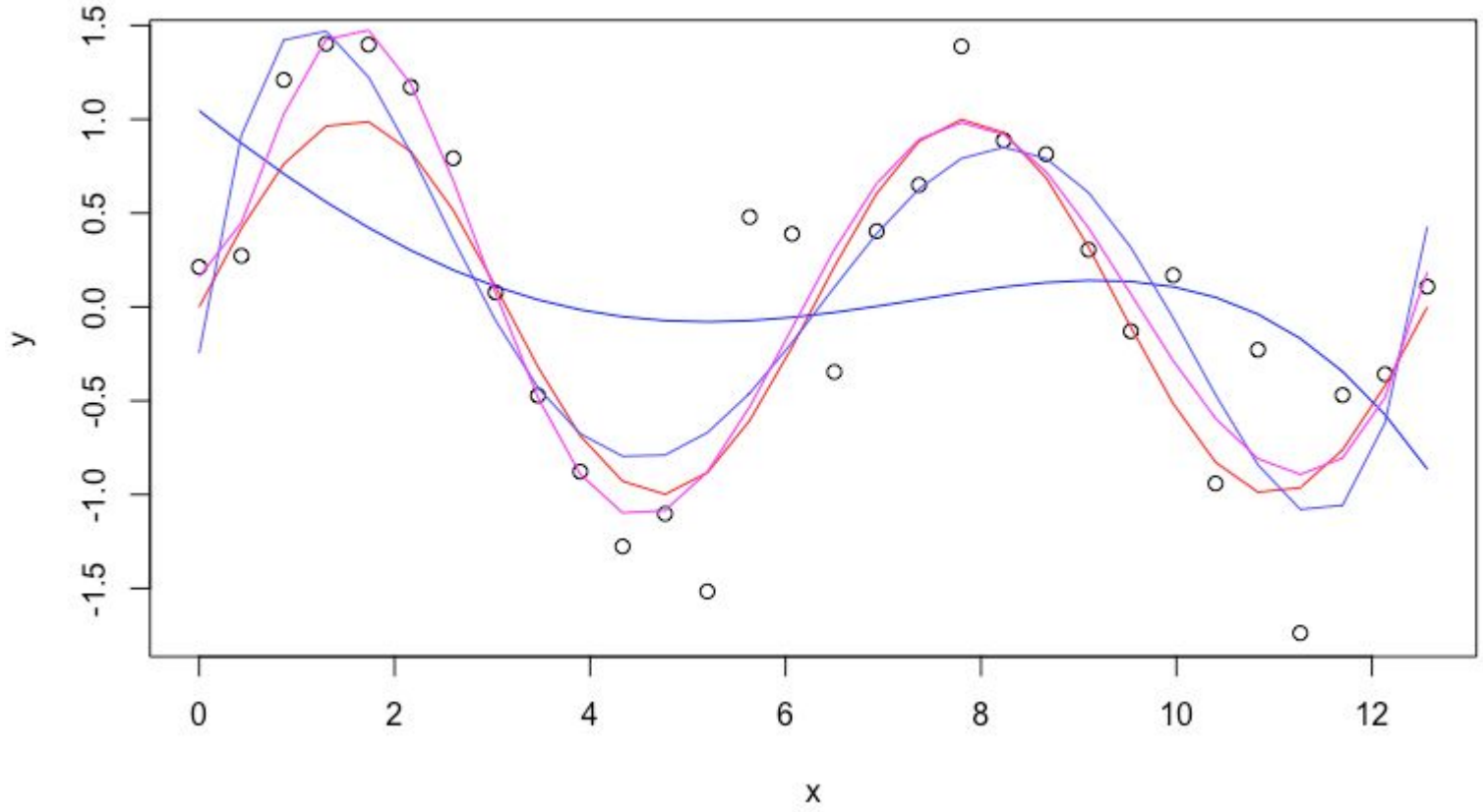

y
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
2
4
6
8
10
12
x



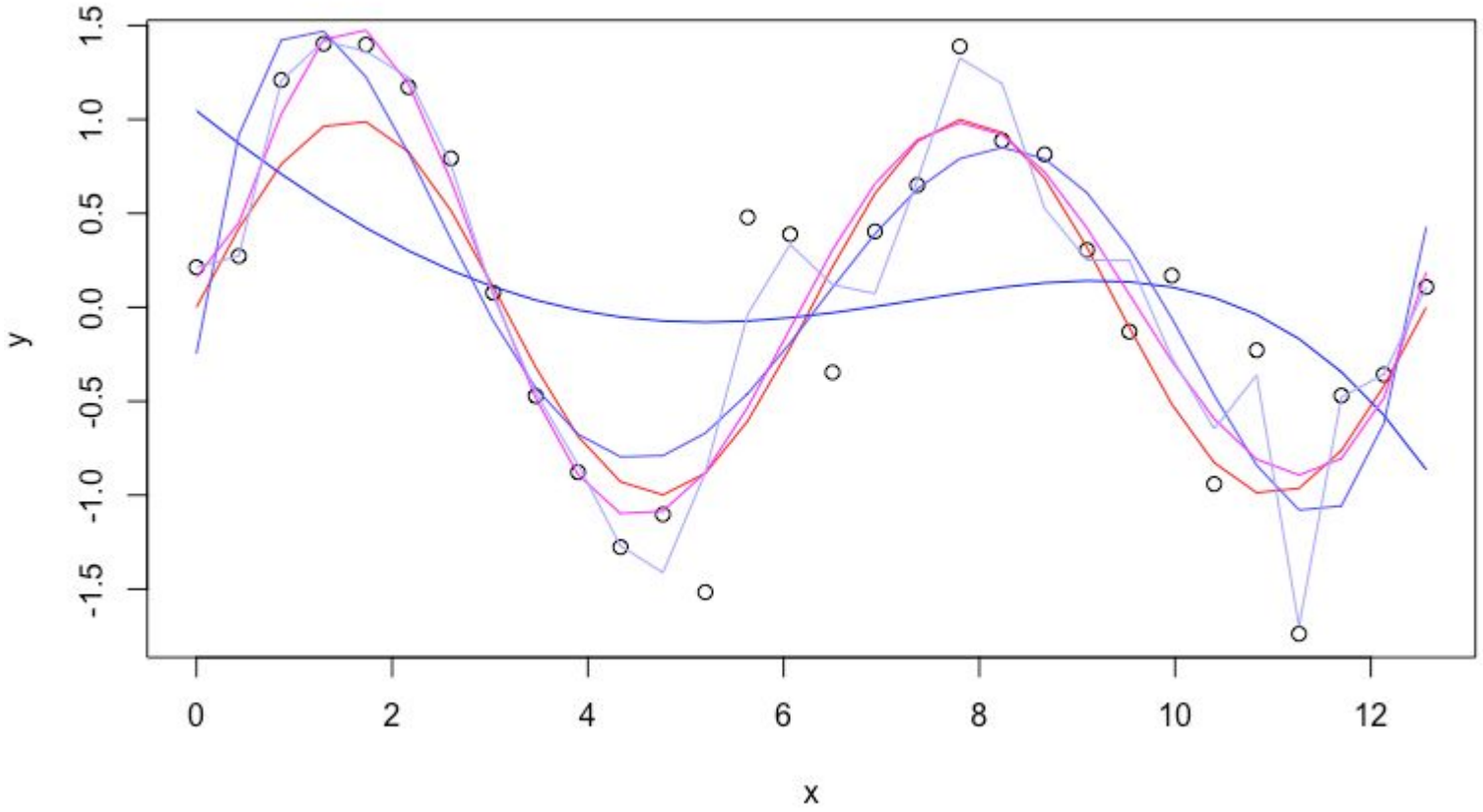
d = 4



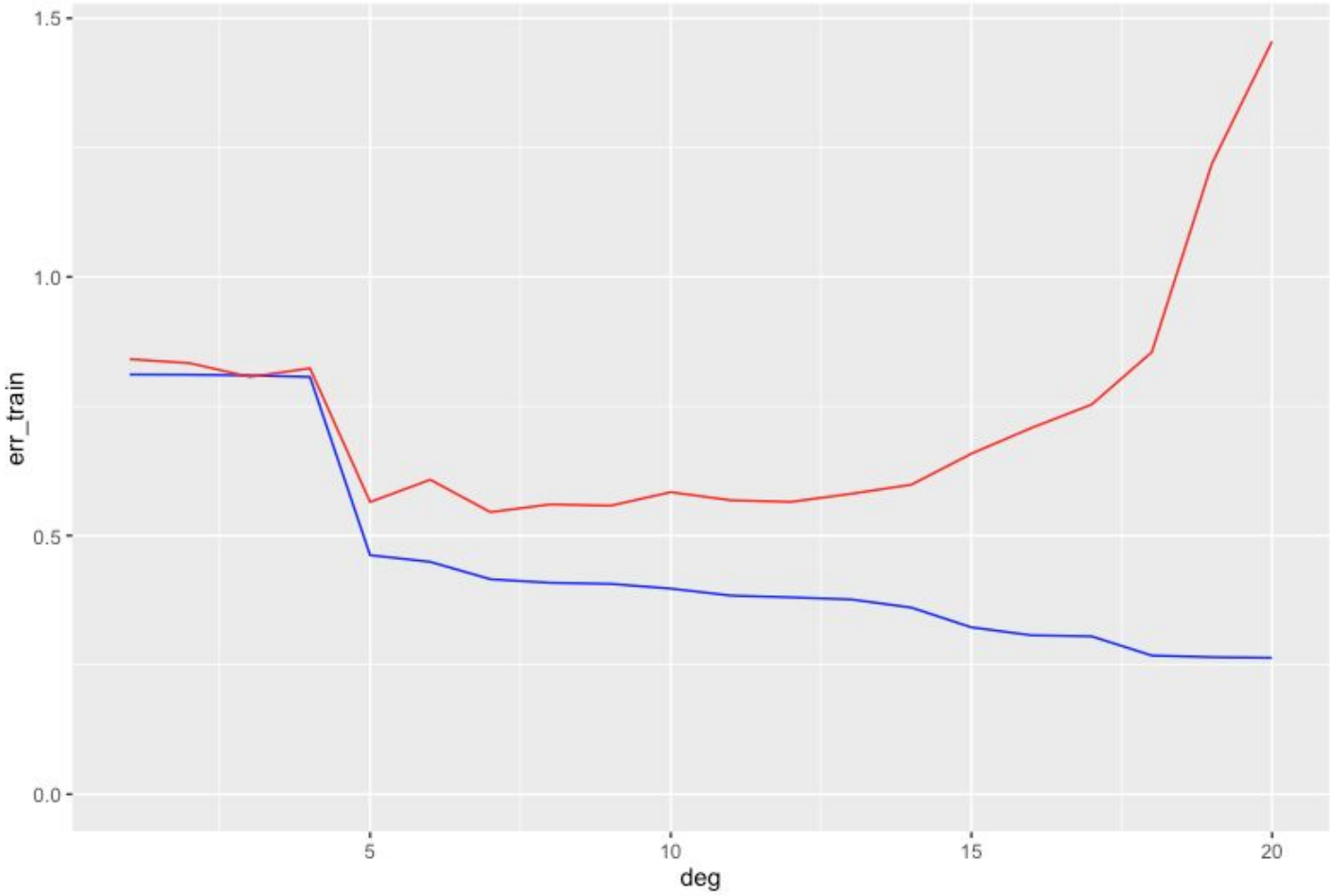
d = 5

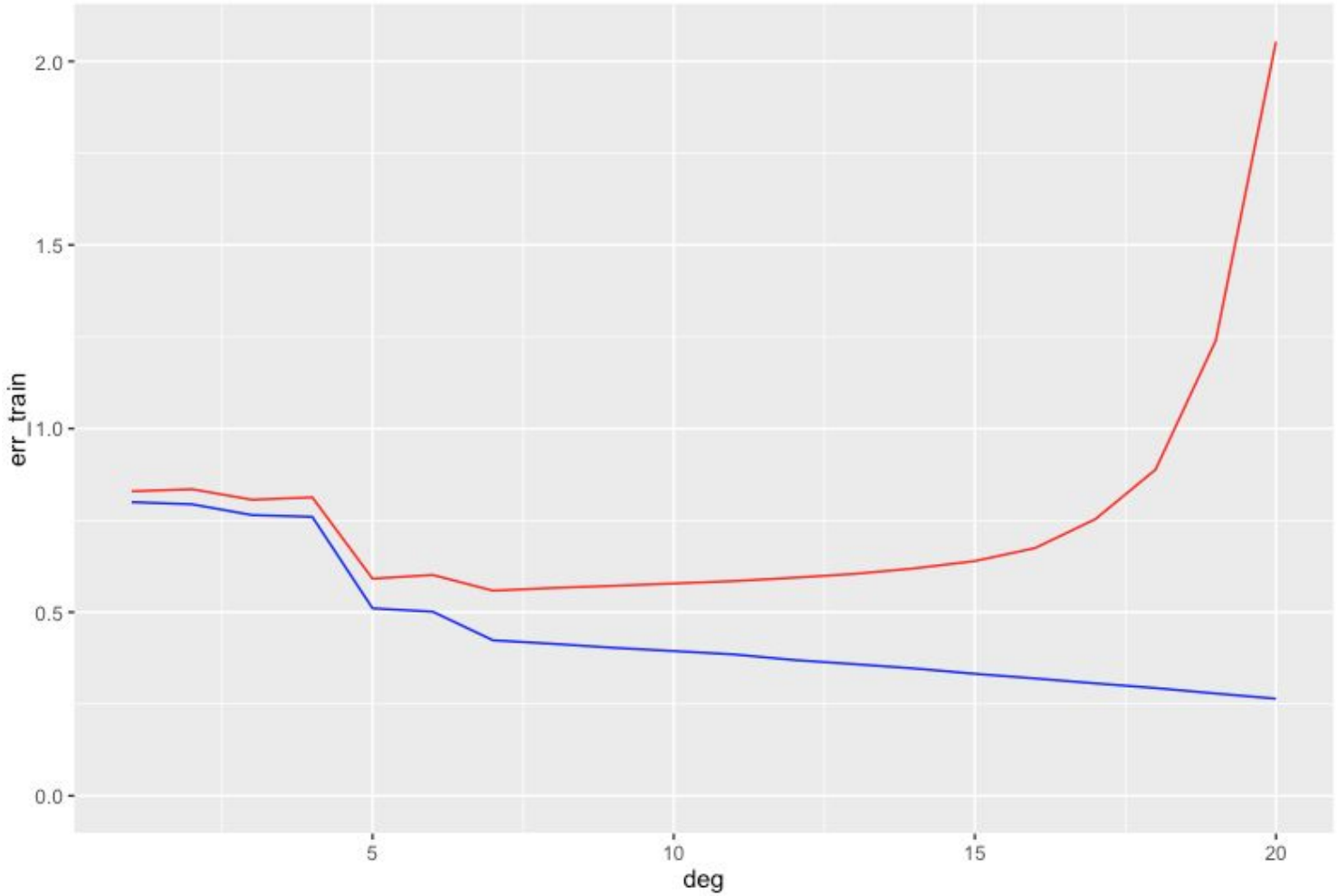


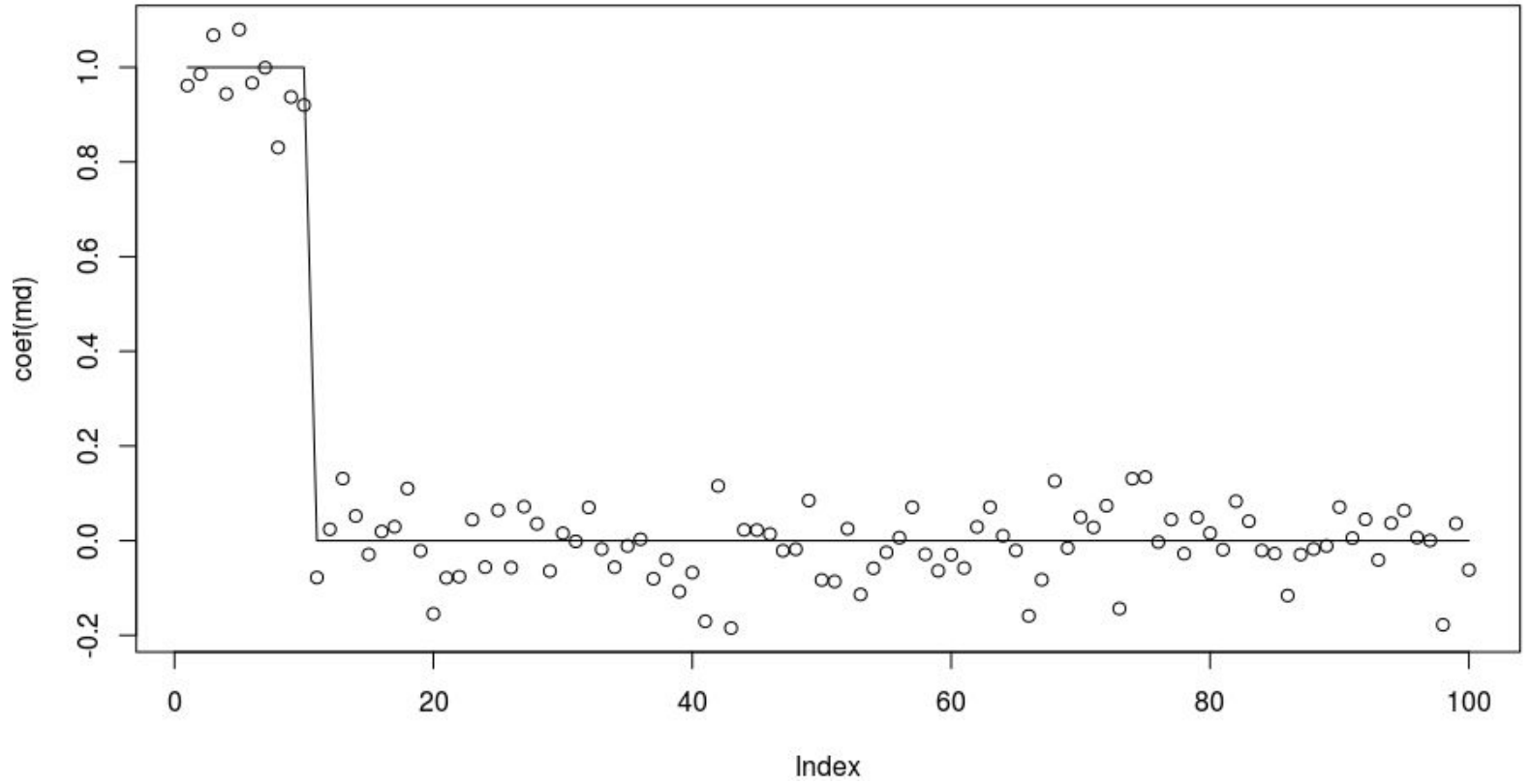
d = 8

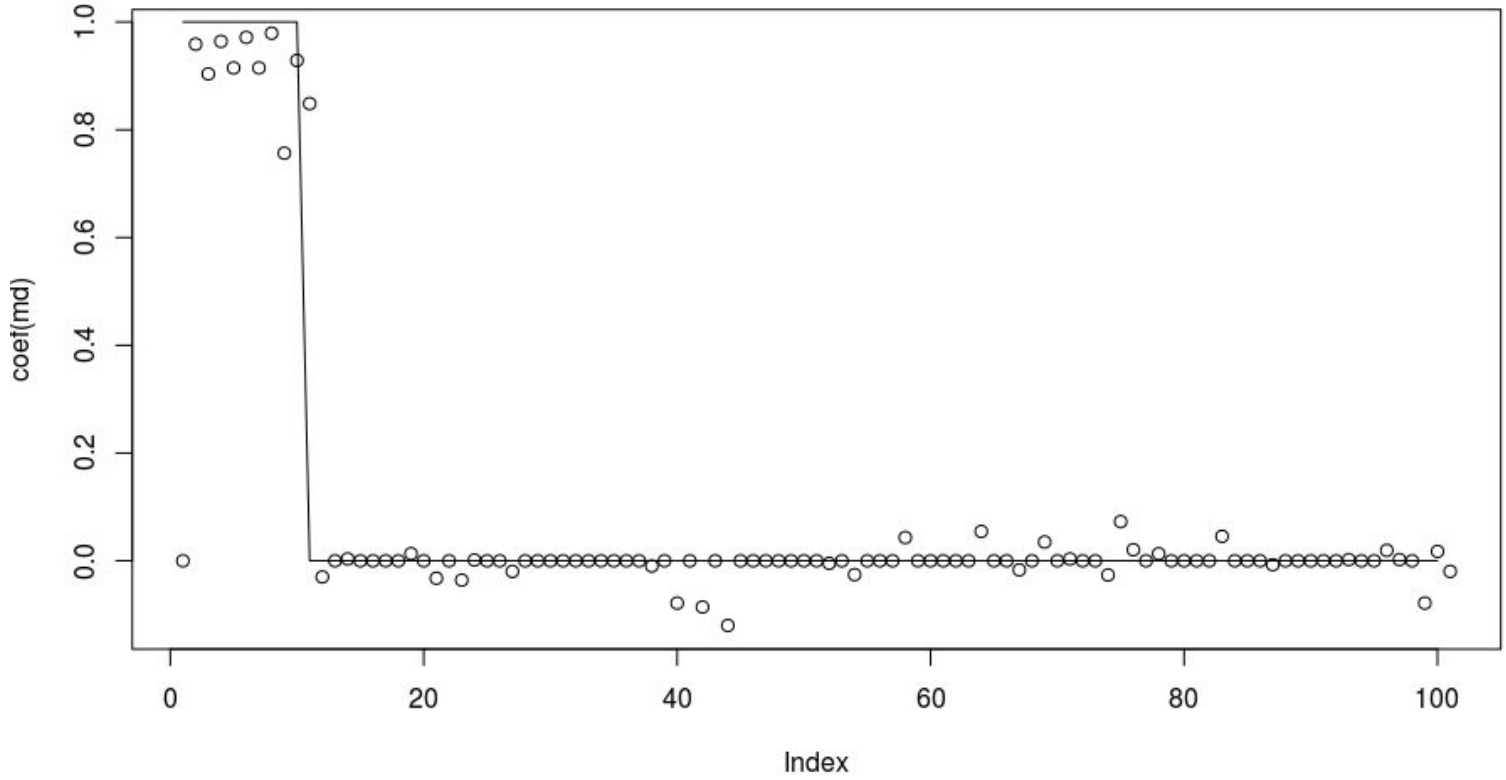


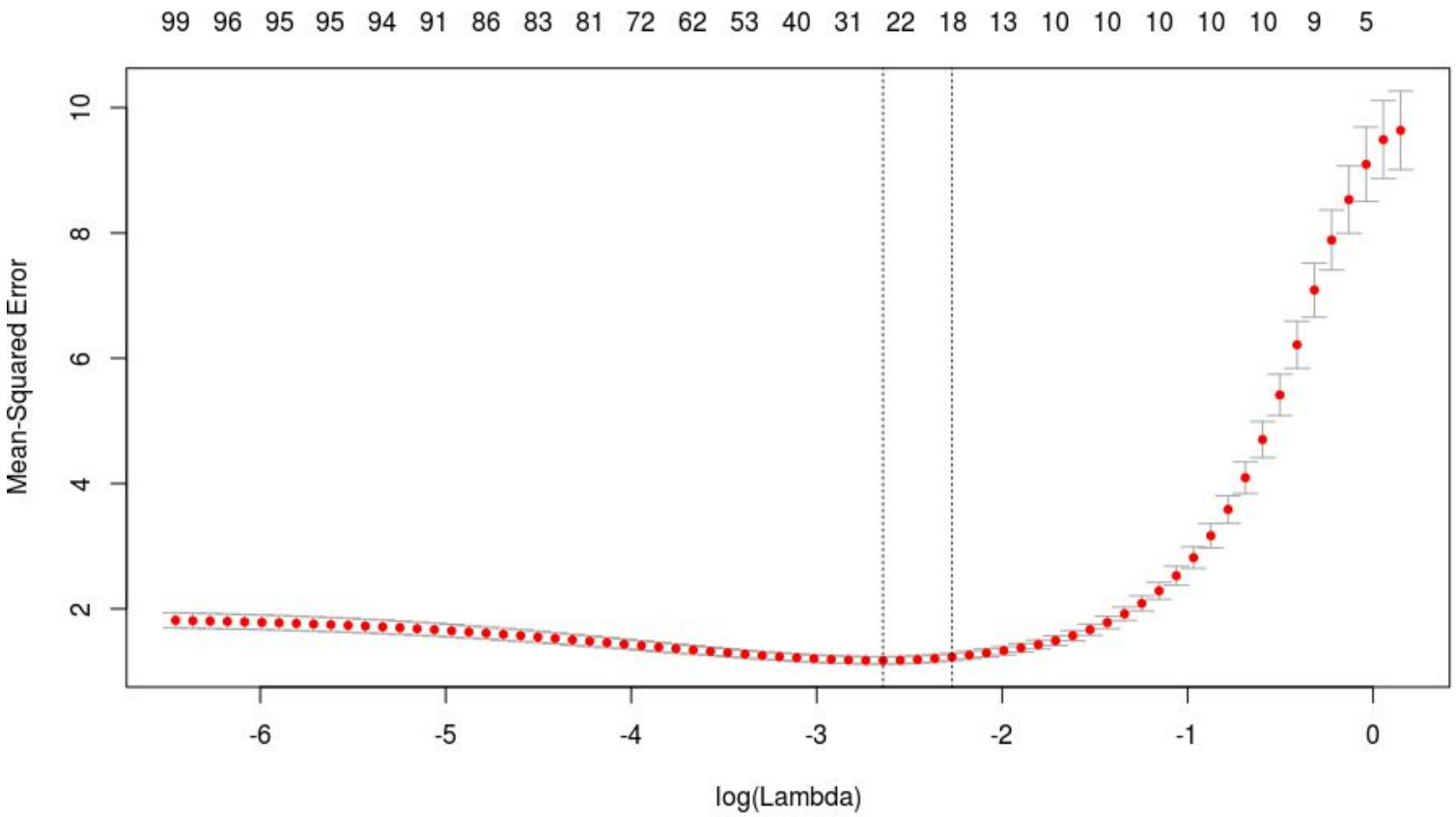
d = 20

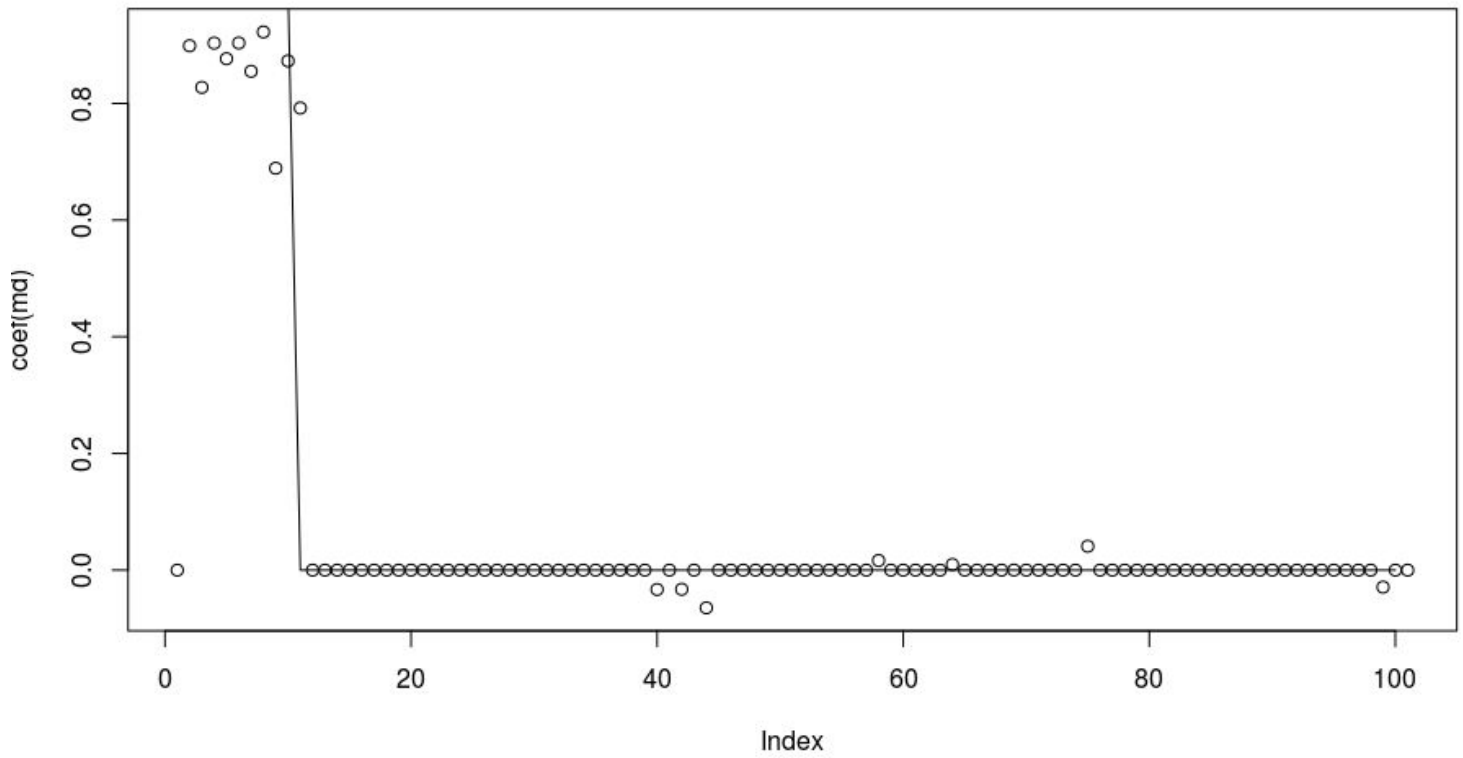

1.5
1.0
0.5
0.0
err_train
5
10
15
20
deg


2.0
1.5
1.0
0.5
0.0
err_train
5
10
15
20
deg


coef(md)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
20
40
60
80
100
Index

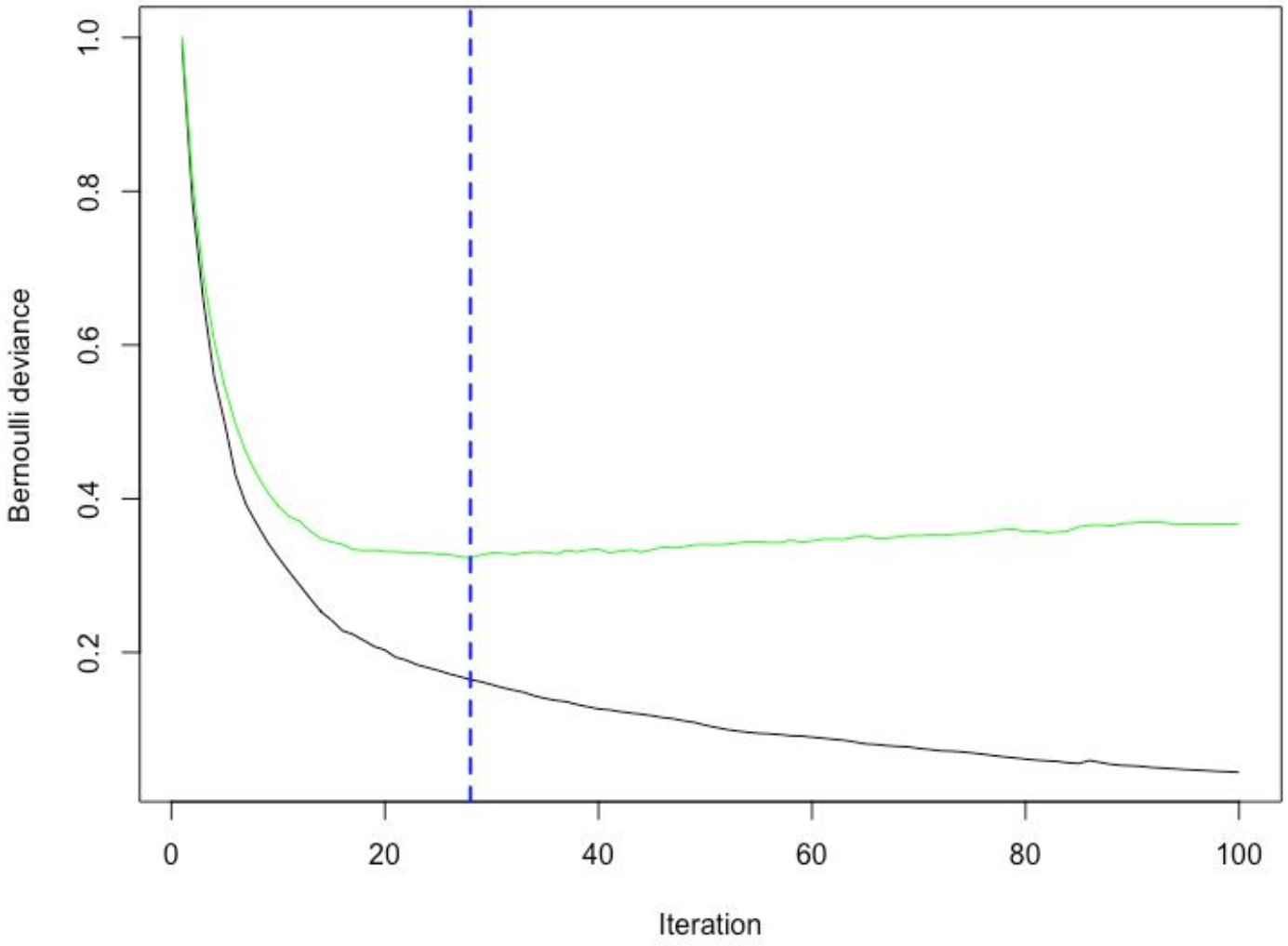

coef(md)
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
Index

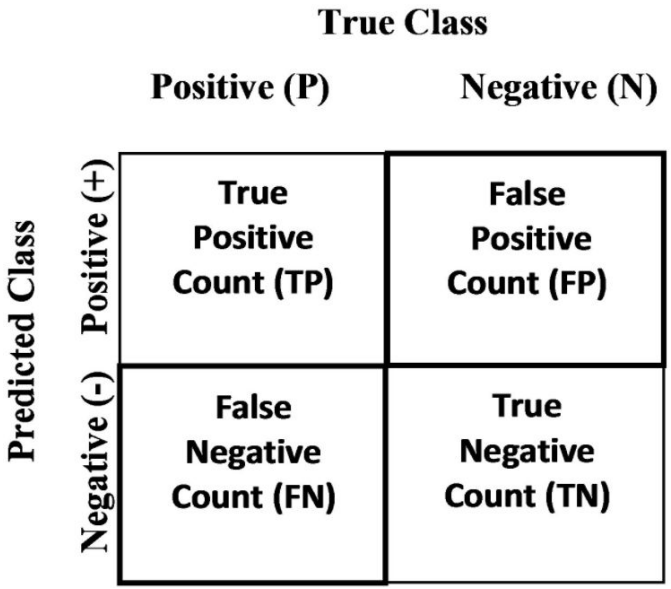

99 96 95 95 94 91 86 83 81 72 62 53 40 31 22 18 13 10 10 10 10 10 9 5
Mean-Squared Error
2
4
6
8
10
-6
-5
-4
-3
-2
-1
0
log(Lambda)

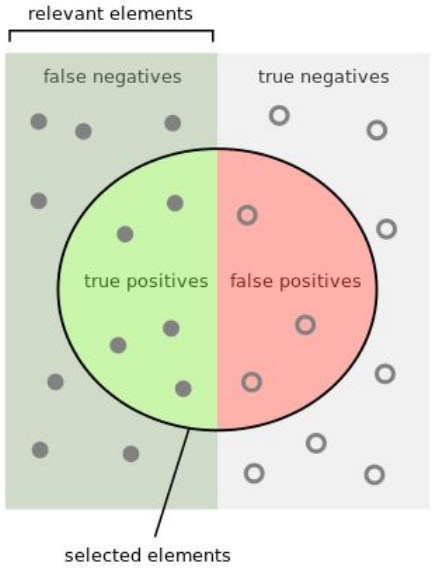

coef(ma)
0.0
0.2
0.4
0.6
0.8
0
20
40
60
80
100
Index

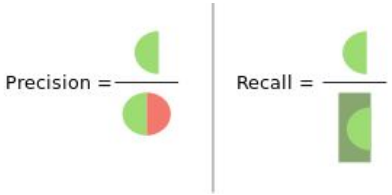

1
2
3
4
5
Train
Train
Validation
Train
Train

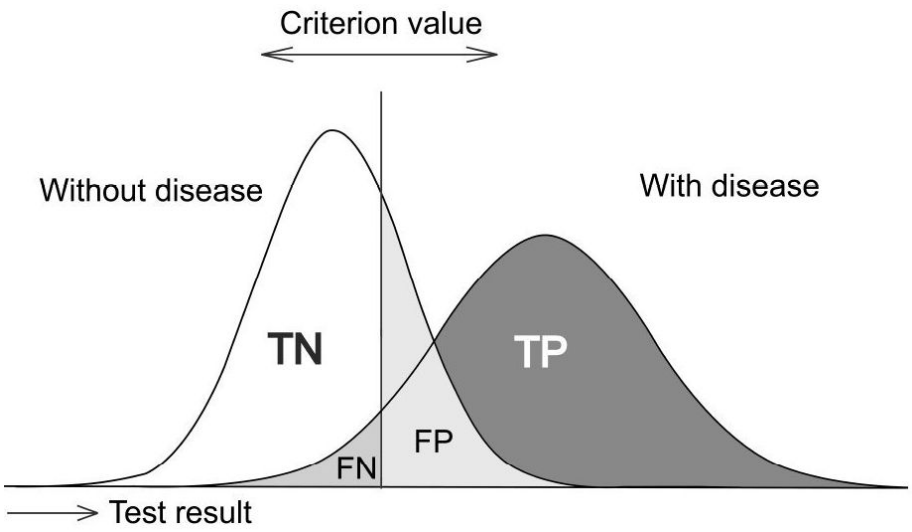

Train
Validation
Test

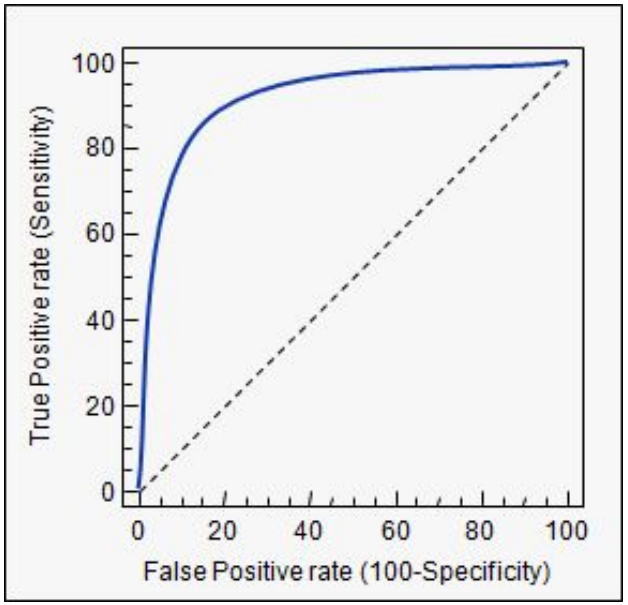

Bernoulli deviance
1.0
0.8
0.6
0.4
0.2
0
20
40
60
80
100
Iteration

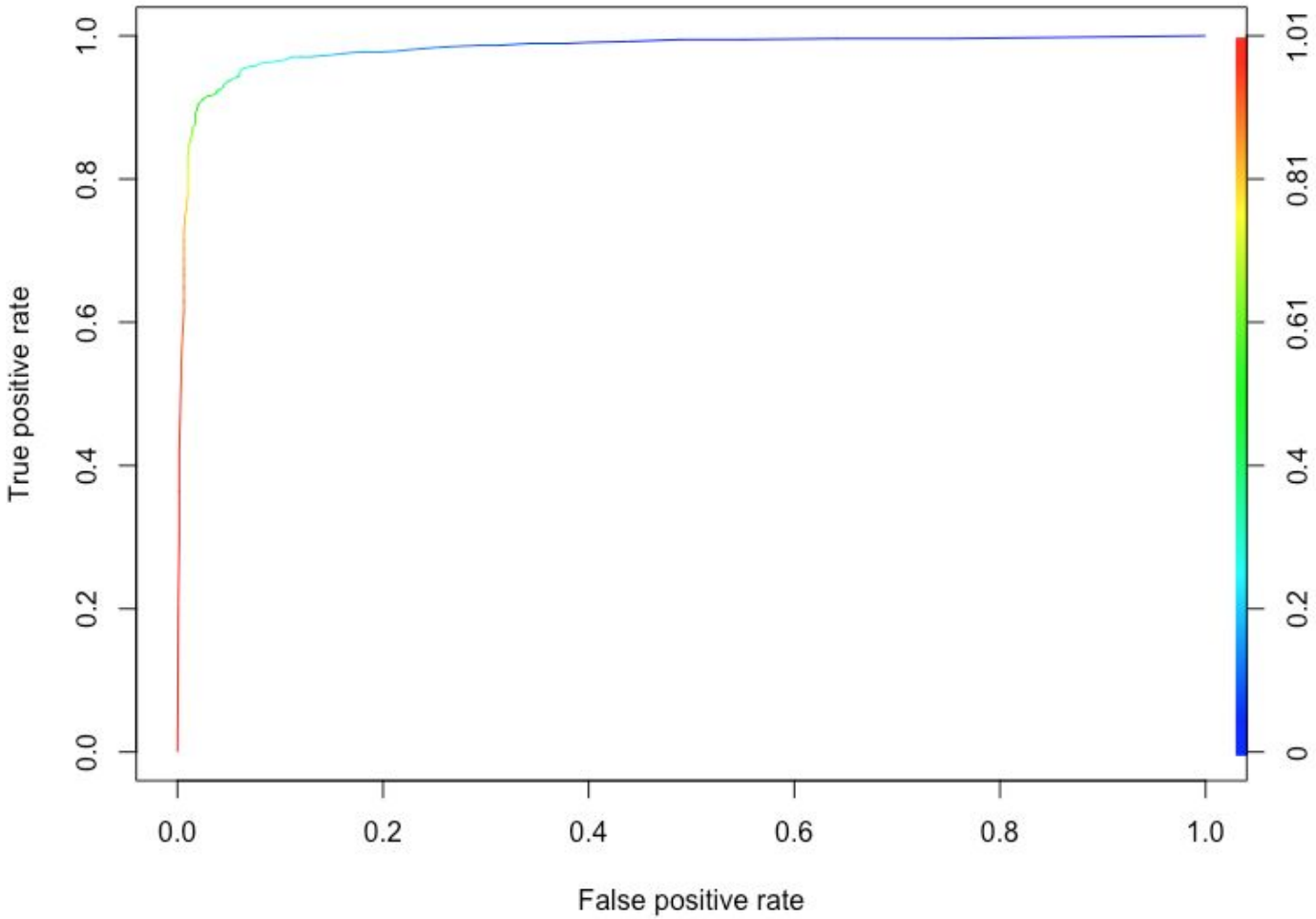

True Class
Positive (P)
Negative (N)
Predicted Class
Positive (+)
Negative (-)
True Positive Count (TP)
False Positive Count (FP)
False Negative Count (FN)
True Negative Count (TN)

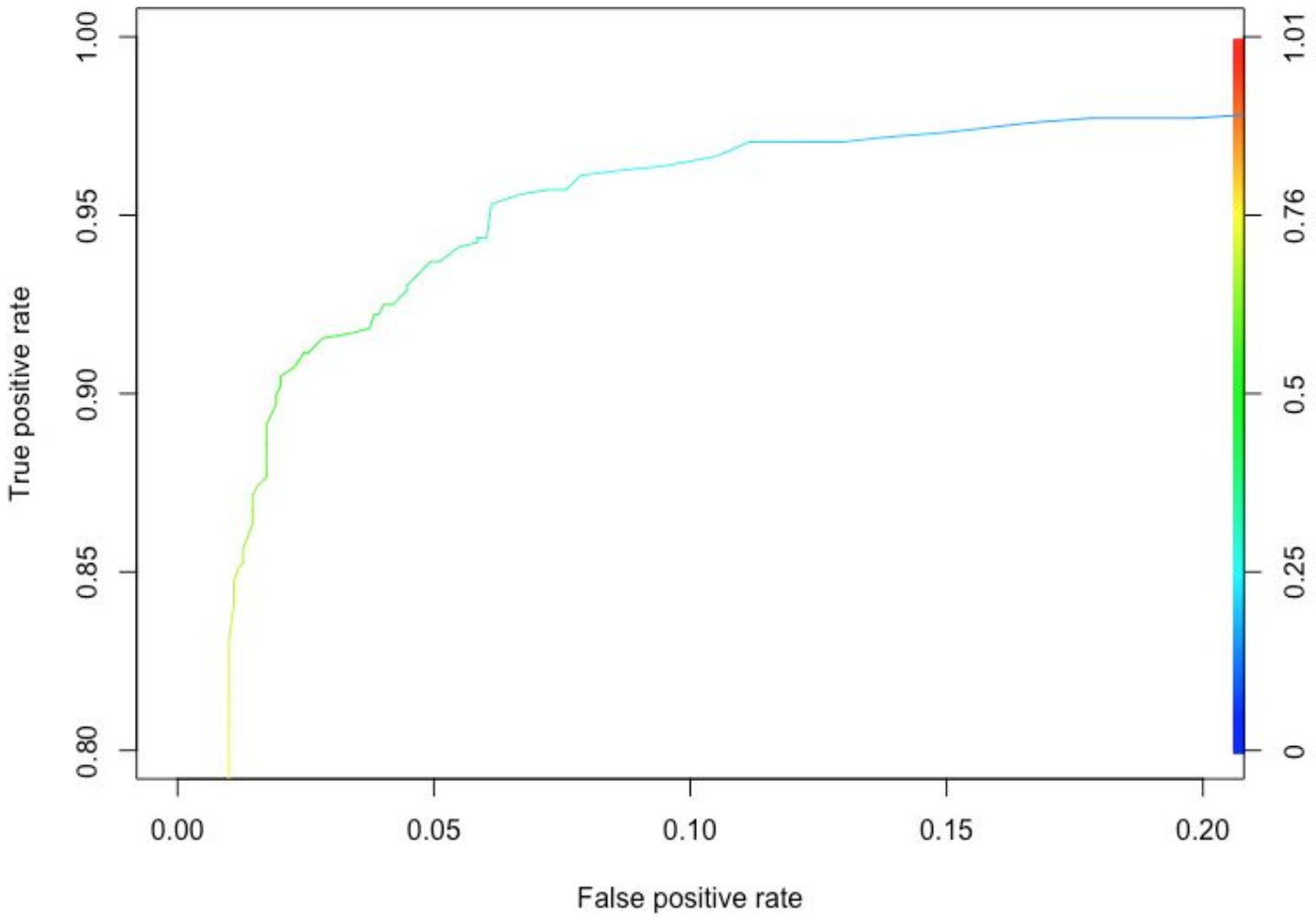

relevant elements
false negatives
true negatives
true positives
false positives
selected elements

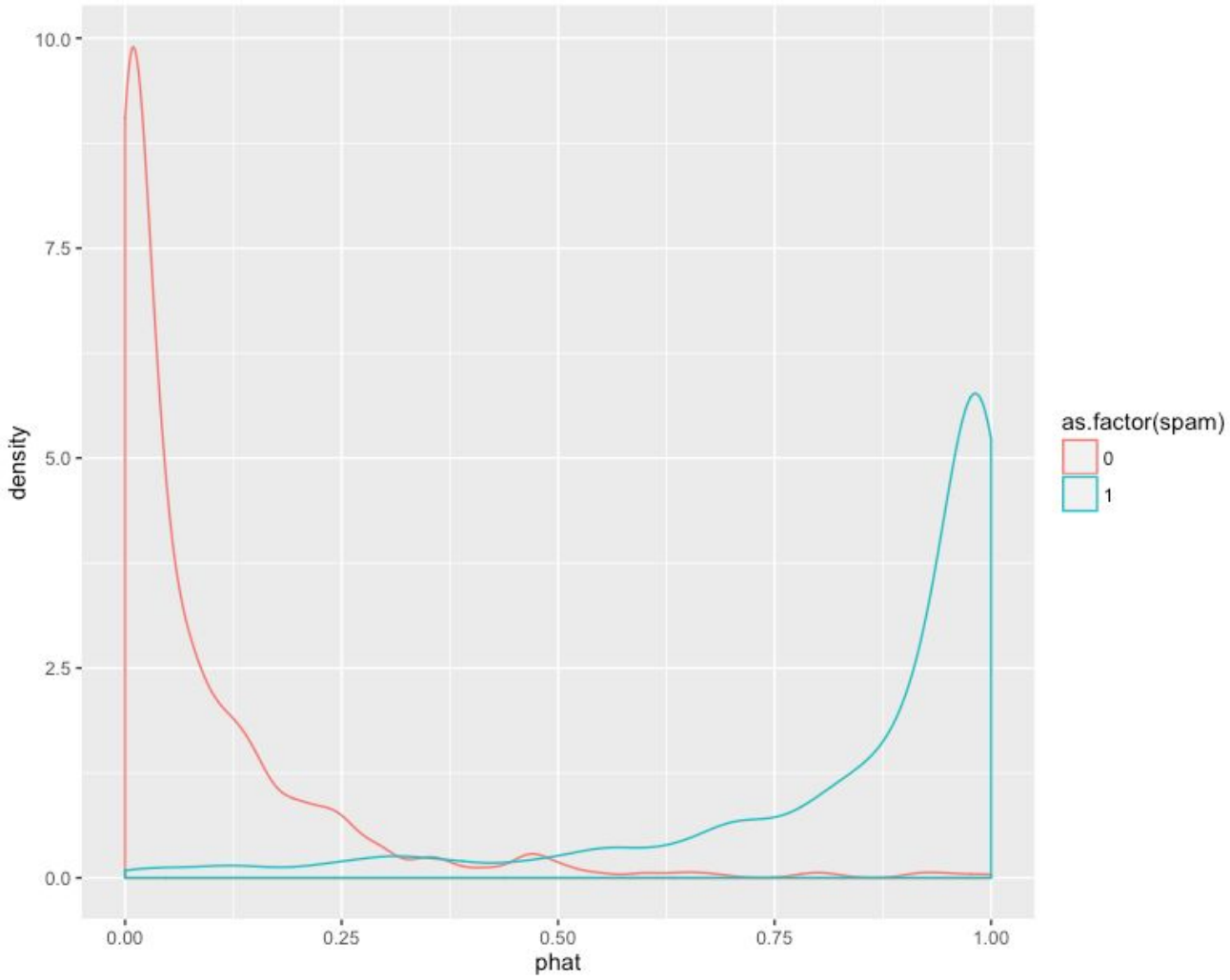

Precision =
Recall =


Criterion value
Without disease
With disease
TN
TP
FP
FN
Test result


True Positive rate (Sensitivity)
False Positive rate (100-Specificity)
0
20
40
60
80
100


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0
0.2
0.4
0.61
0.81
1.01


True positive rate
1.00
0.95
0.90
0.85
0.80
0.00
0.05
0.10
0.15
0.20
False positive rate
1.01
0.76
0.5
0.25
0


10.0
7.5
5.0
2.5
0.0
density
0.00
0.25
0.50
0.75
1.00
phat
as.factor(spam)
0
1