# Méthodes d'apprentissage automatique pour la bioinformatique

## *BIOL-F-524*

Gianluca Bontempi

Département d'Informatique

Boulevard de Triomphe - CP 212

http://www.ulb.ac.be/di

# A K-NN classifier

Suppose a training set is available and that the classification is required for an input vector $x$. Henceafter we will call this input vector a *query* point. The classification procedure of a K-NN classifier can be summarized in these steps:

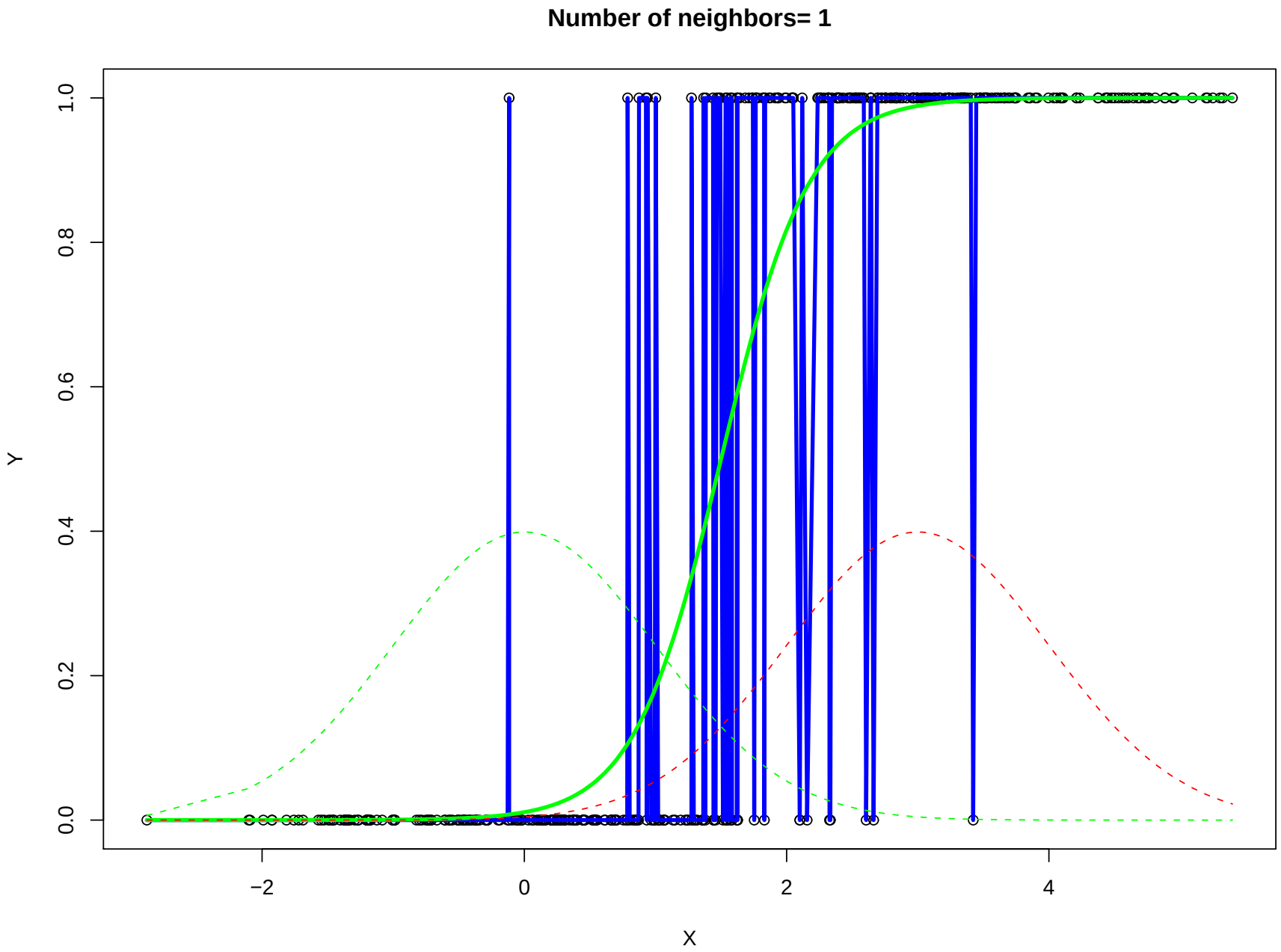
1. Compute the distance between the query and the training samples according to a predefined metric.
2. Rank the observed inputs on the basis of their distance to the query.
3. Select a subset $\{x_{[1]}, \ldots, x_{[K]}\}$ of the $K$ nearest neighbors. Each of these neighbors $x_{[k]}$ has an associated class $y_{[k]}$ ($0$ or $1$).
4. Compute the estimation of the conditional probability by constant fitting

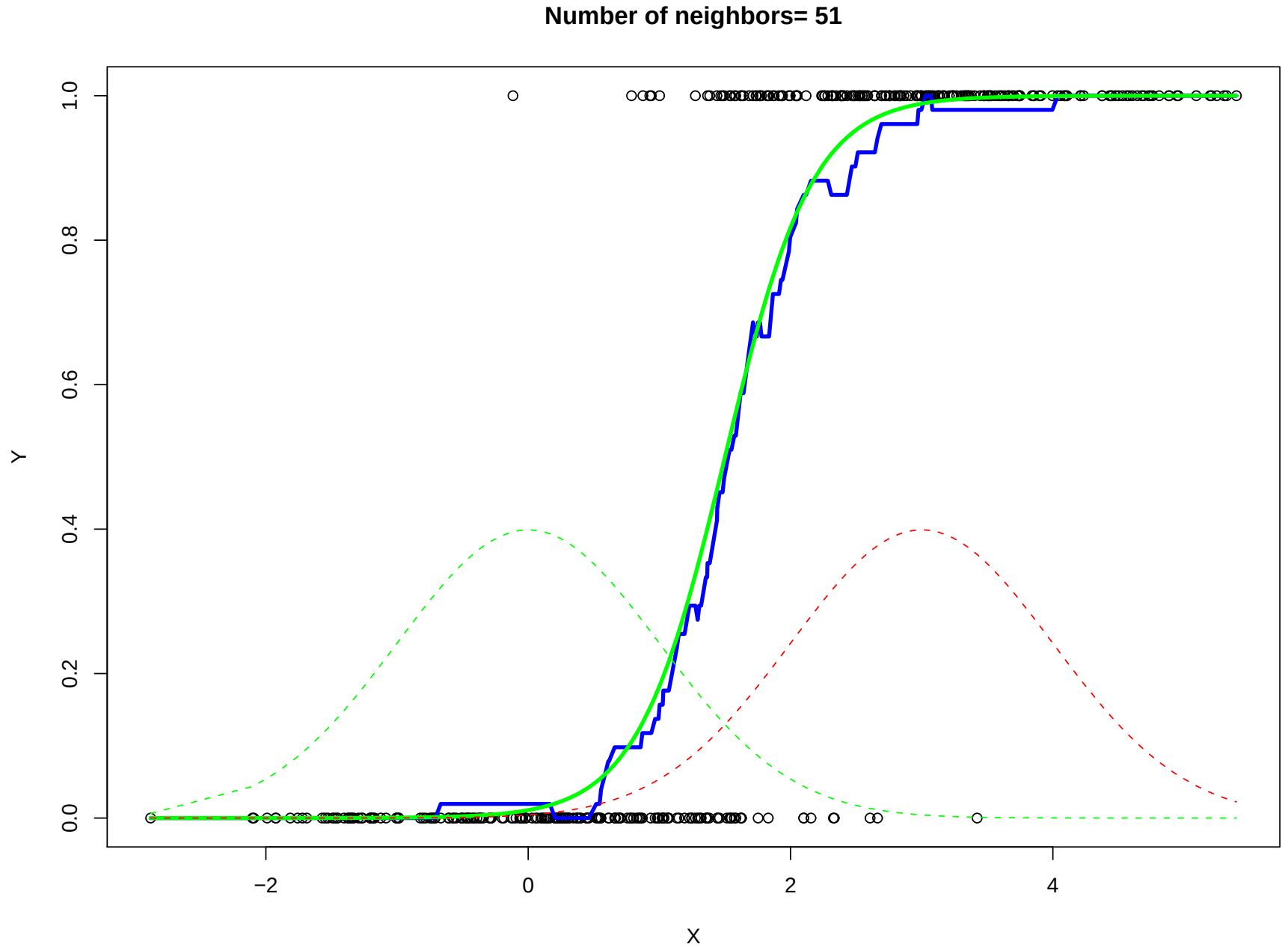
$$\hat{p}_1(x) = \frac{\sum_{k=1}^{K} y_{[k]}}{K}$$

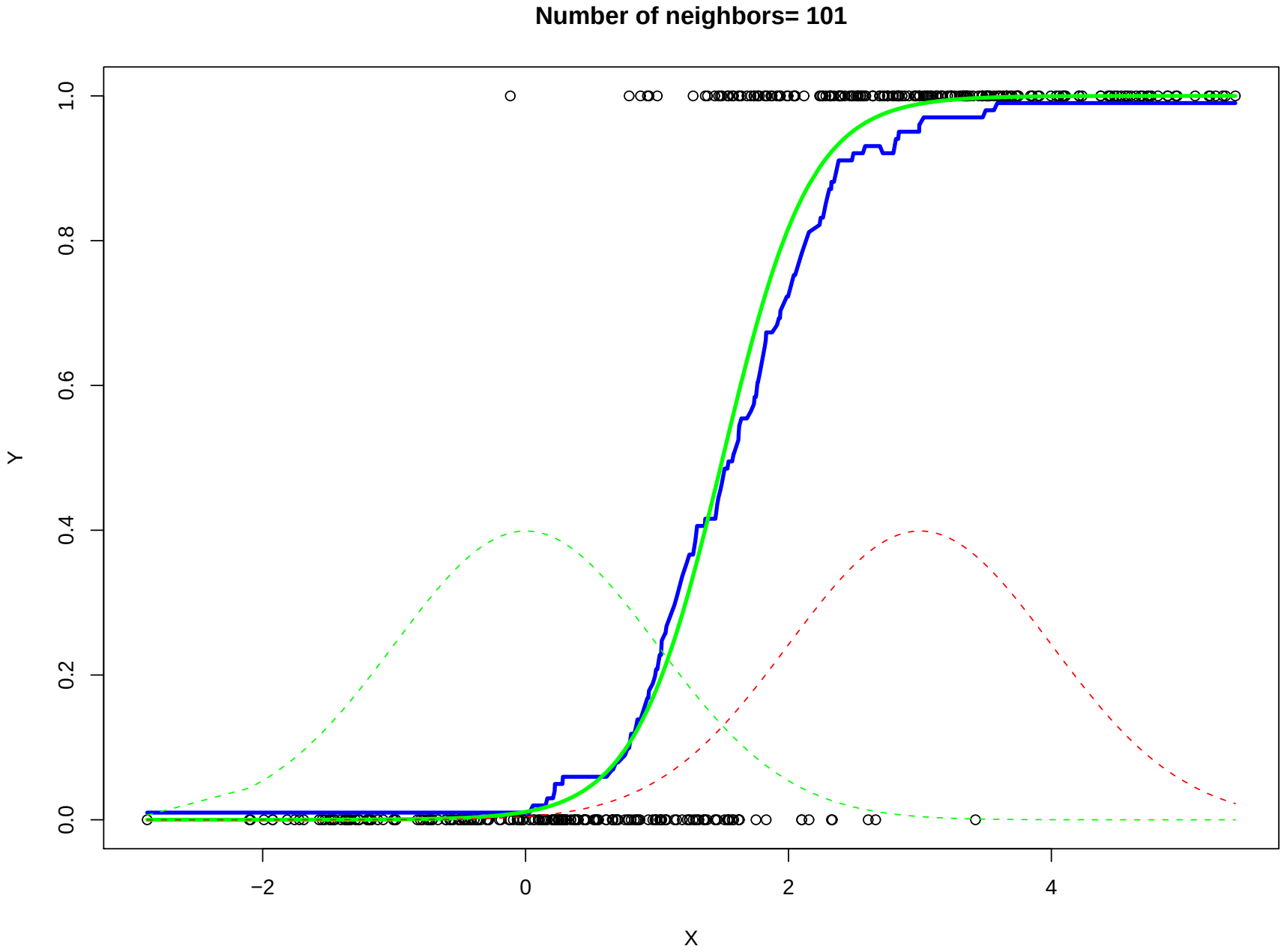
or linear fitting

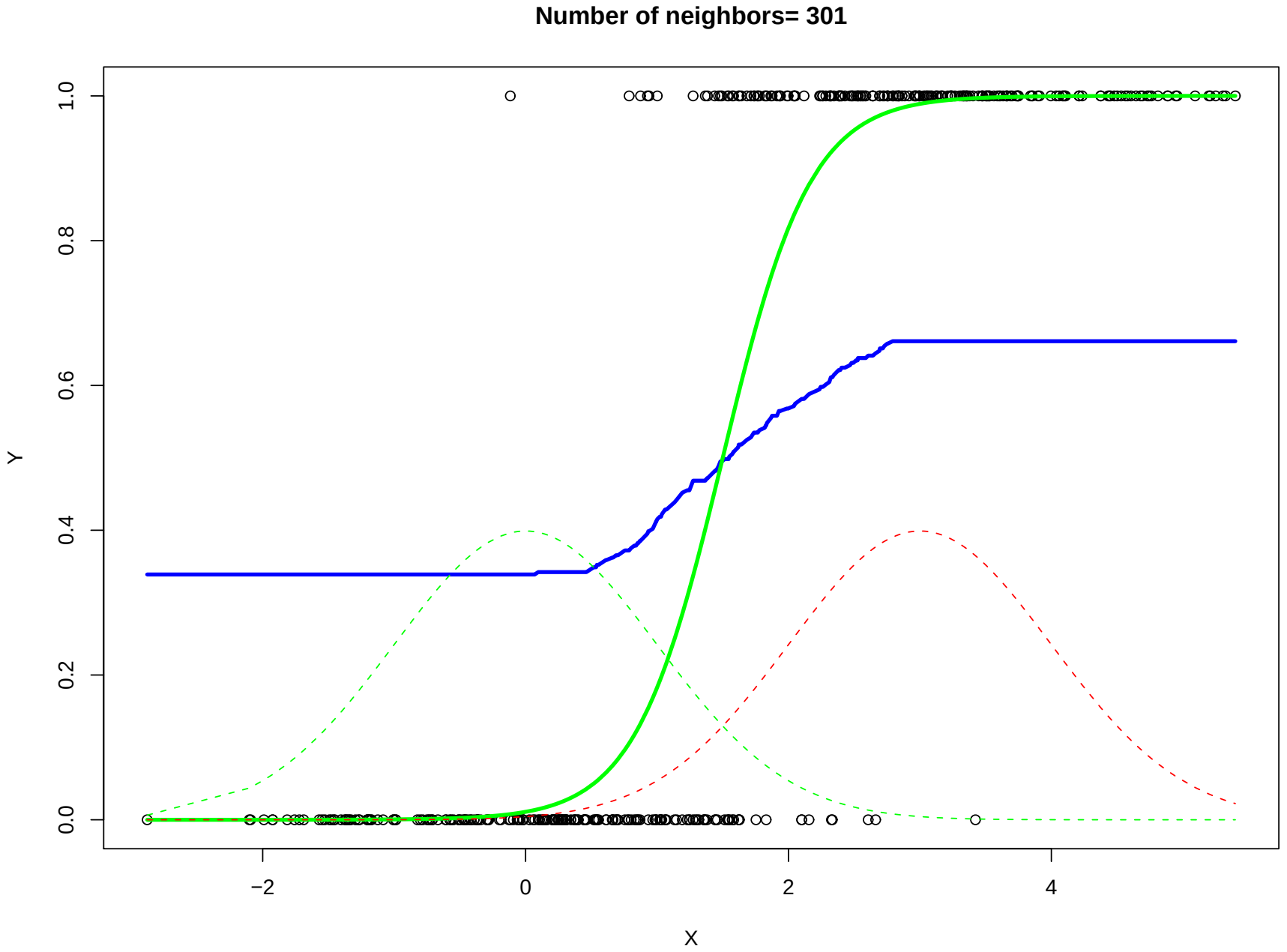
$$\hat{p}_1(x) = \hat{a}x + \hat{b}$$

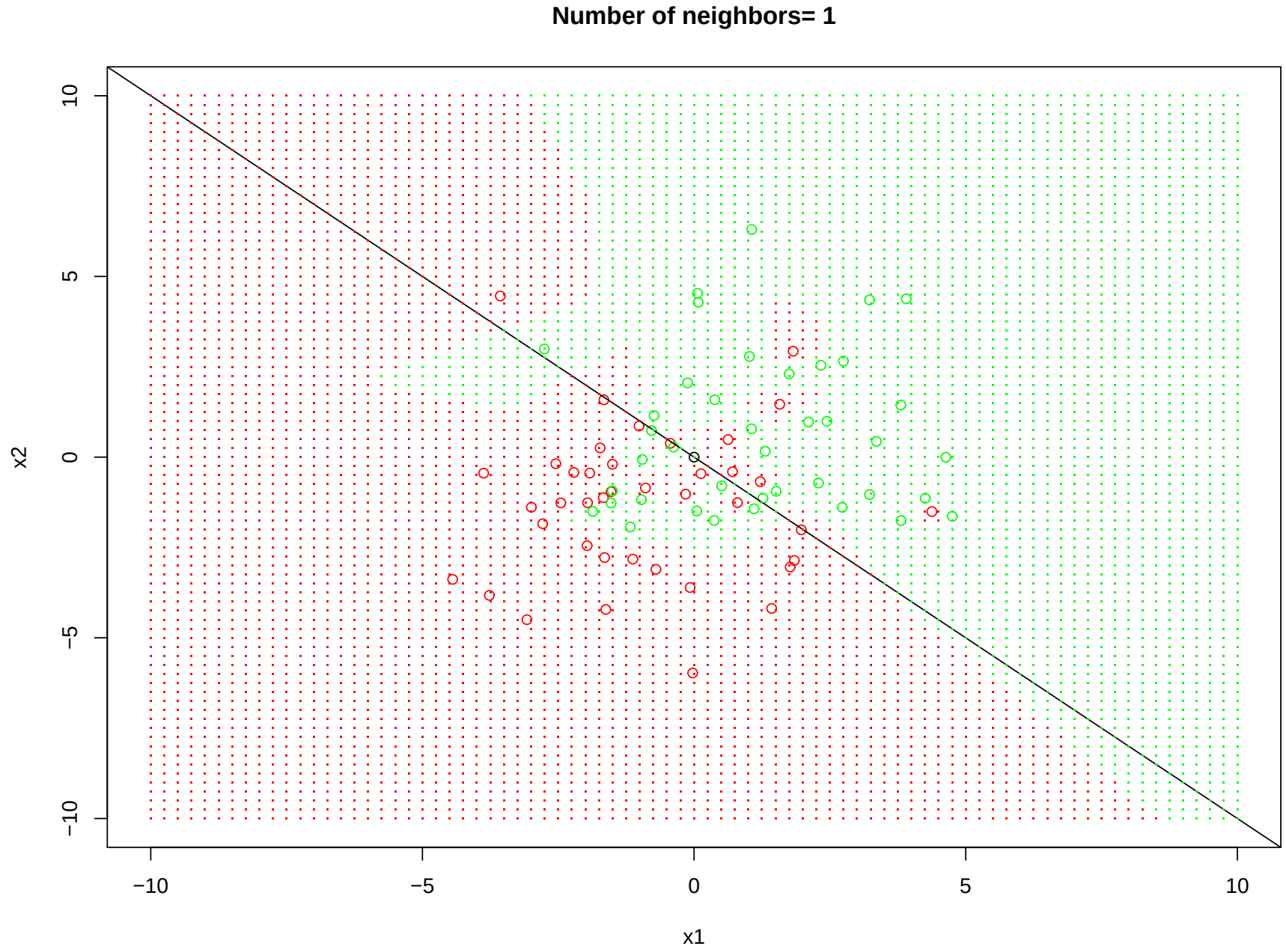
where the parameters $\hat{a}$ and $\hat{b}$ are locally fitted by least-squares.

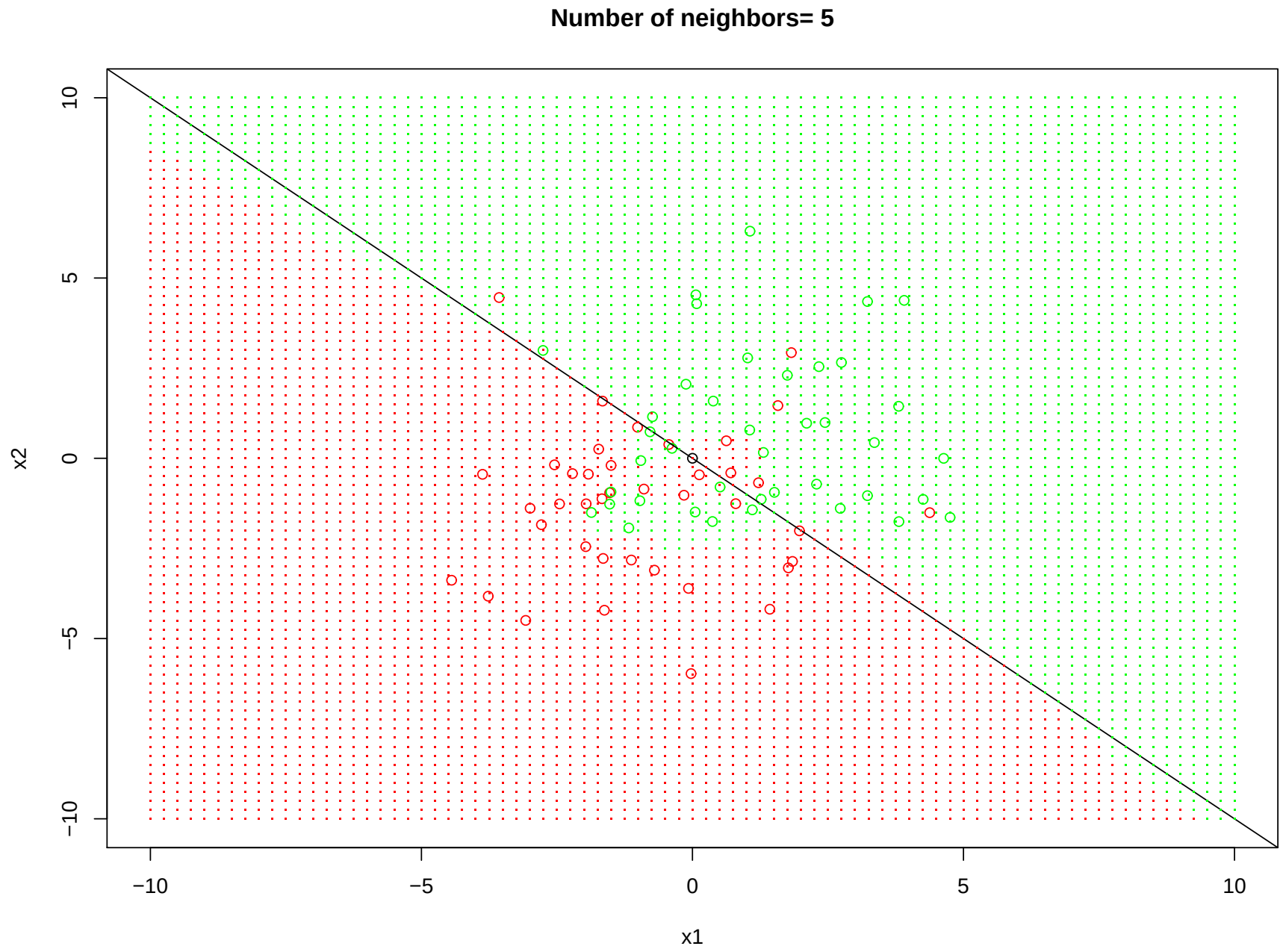

Number of neighbors= 1
Y
X
0.0
0.2
0.4
0.6
0.8
1.0
−2
0
2
4

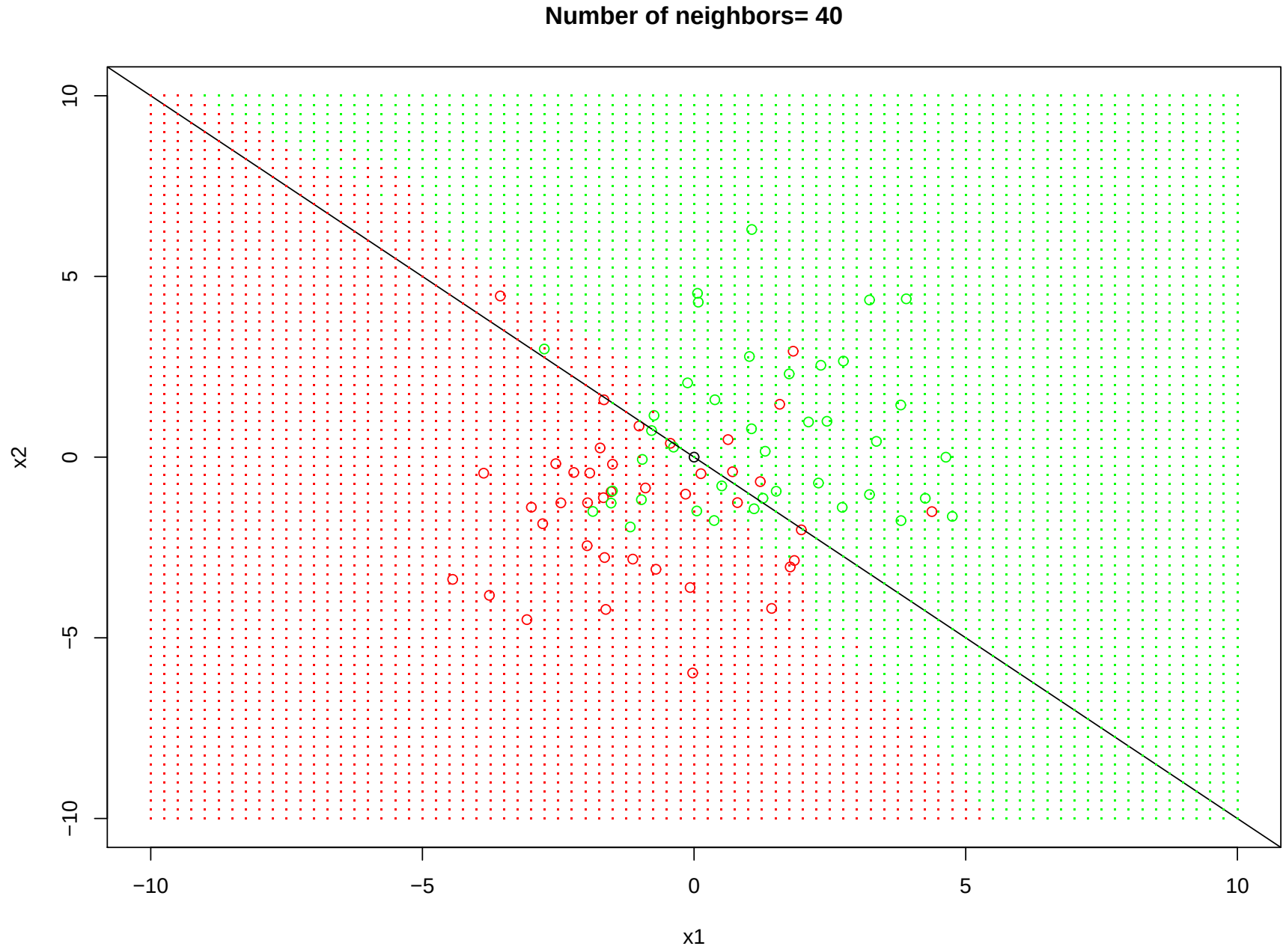

Number of neighbors= 51
Y
X
0.0
0.2
0.4
0.6
0.8
1.0
−2
0
2
4


Number of neighbors= 101
Y
X
0.0
0.2
0.4
0.6
0.8
1.0
−2
0
2
4


Number of neighbors= 301
Y
X
0.0
0.2
0.4
0.6
0.8
1.0
−2
0
2
4


Number of neighbors= 1
x2
10
5
0
−5
−10
−10
−5
0
5
10
x1


Number of neighbors= 5
x2
10
5
0
−5
−10
−10
−5
0
5
10
x1


Number of neighbors= 40
x2
10
5
0
−5
−10
−10
−5
0
5
10
x1