Lecture 7 – K-Nearest Neighbors

Instructor: Hamed Hasheminia

Agenda

- Classification problems
- Misclassification error
- KNN algorithm for classifications
- CV for KNN algorithm
- Limitations of KNN algorithm
- KNN algorithm for regression

Classification Problems

- Here the response Variable Y is ***Qualitative*** e.g. e-mail if one of C = (spam,ham), digit is on of C = {0,1,...,9}. Our goals are to:
  - Build a classifier C(X) that assigns a class label from C to a feature unlabeled observations X.
  - Assess the uncertainty in each classification
  - Understand the roles of the different predictors among $X=(X_1,X_2,\ldots,X_p)$

Classification: some details

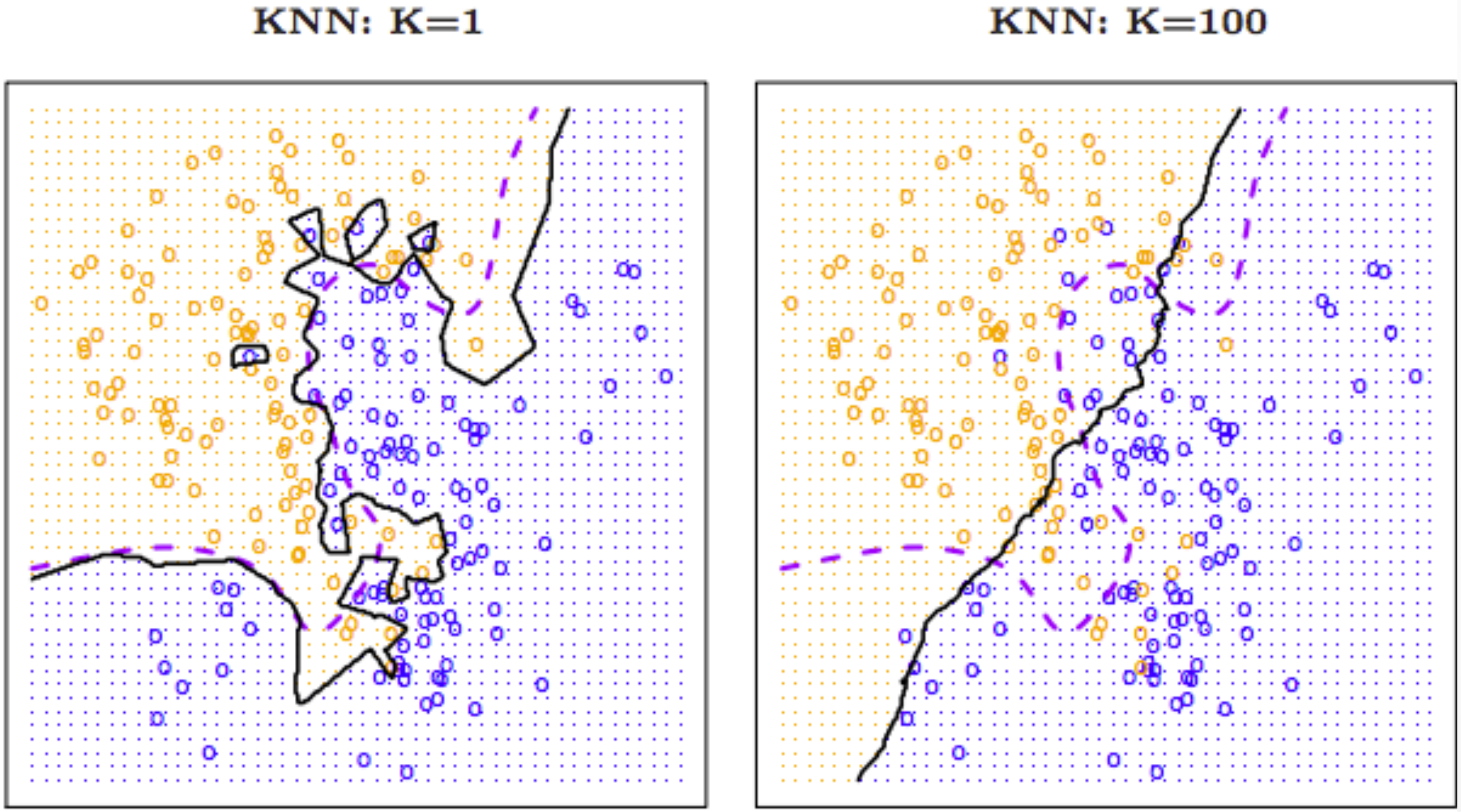
- Typically we measure the performance of $\hat{C}(x)$

Using the misclassification error rate:

$$\mathrm{Err}_{\mathsf{Te}} = \mathrm{Ave}_{i \in \mathsf{Te}} I[y_i \neq \hat{C}(x_i)]$$

What is K-Nearest Neighbors?

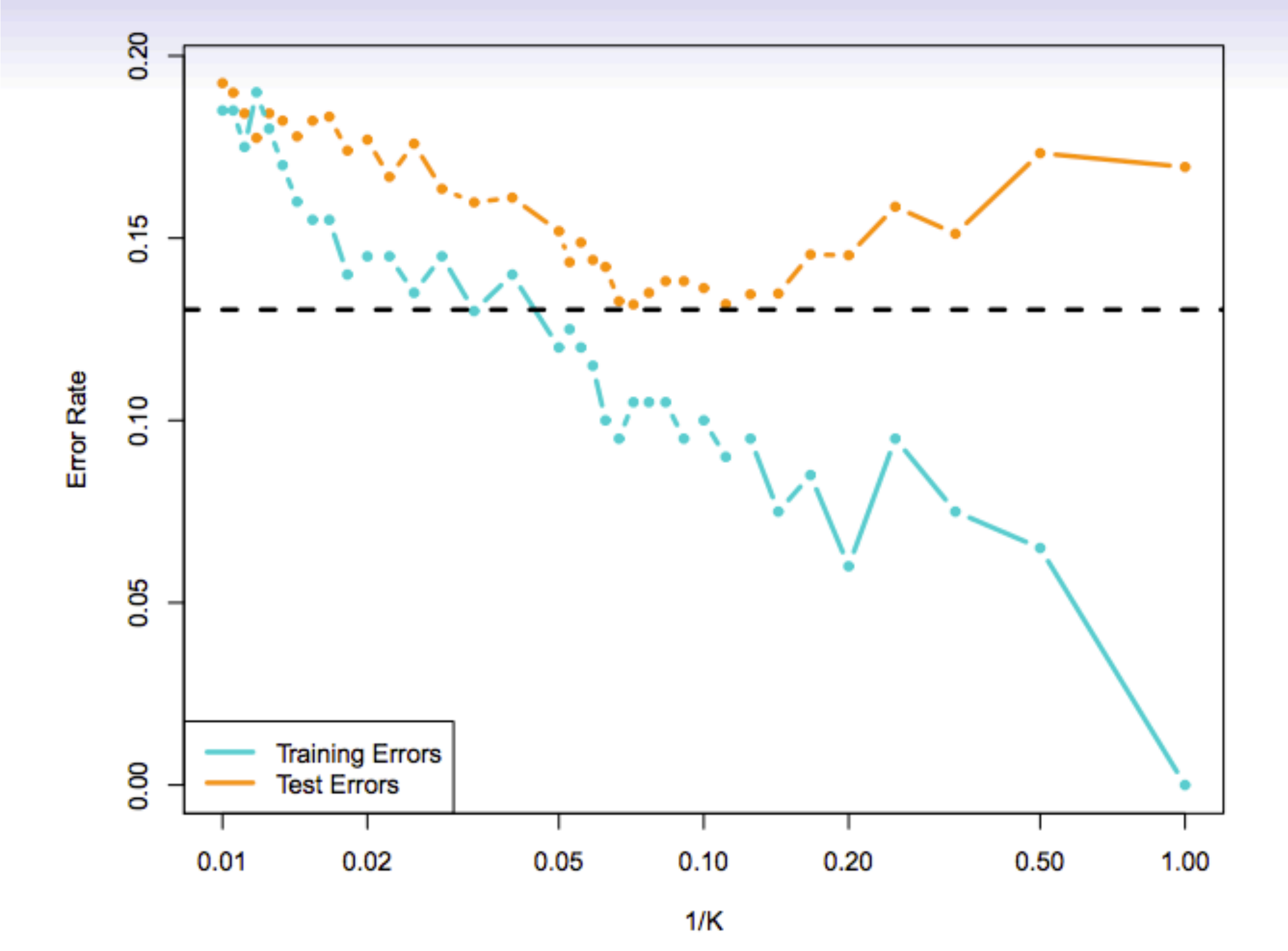
- K Nearest Neighbors (KNN) is a fairly straightforward algorithm used for classification:
  - For a given point, calculate the distance to all other points.
  - Given those distances, pick the k closest points.
  - Calculate the probability of each class label five those points
  - The original point is classified as the class label with the largest probability ("votes")

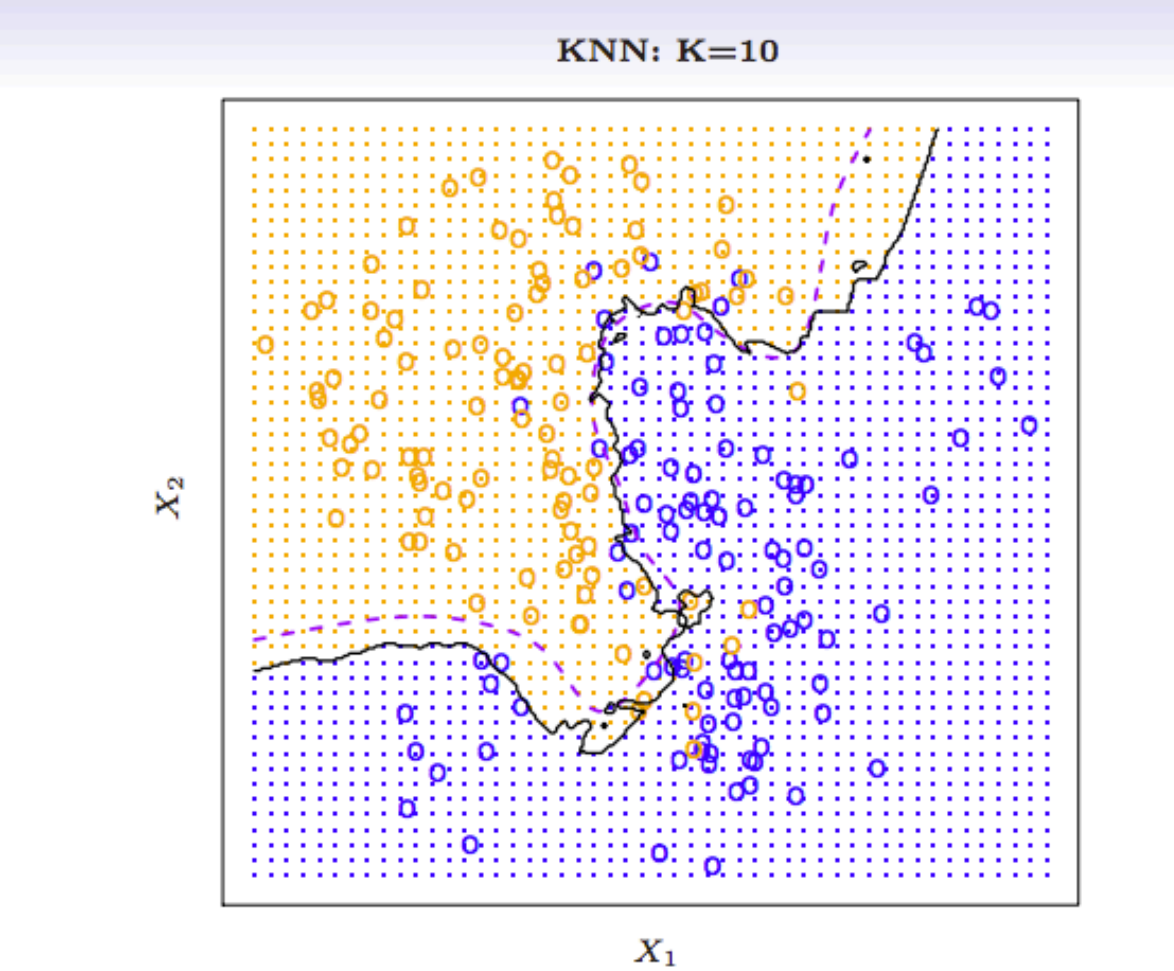


KNN: K=1

KNN: K=100

How to choose optimal K




KNN: K=10
X_2
X_1

What happen in ties?

- In Sklearn, in the case of ties, it will designate the class based on what the algorithm saw first in the training set.
- We can also implement "weights", so that the total distance plays a more significant role.

A few issues with the KNN algorithm

1. Nearest neighbor algorithm can be pretty good for small number of features i.e. $p < 5$ and large-ish *N*
2. They become extremely lousy when p is large.

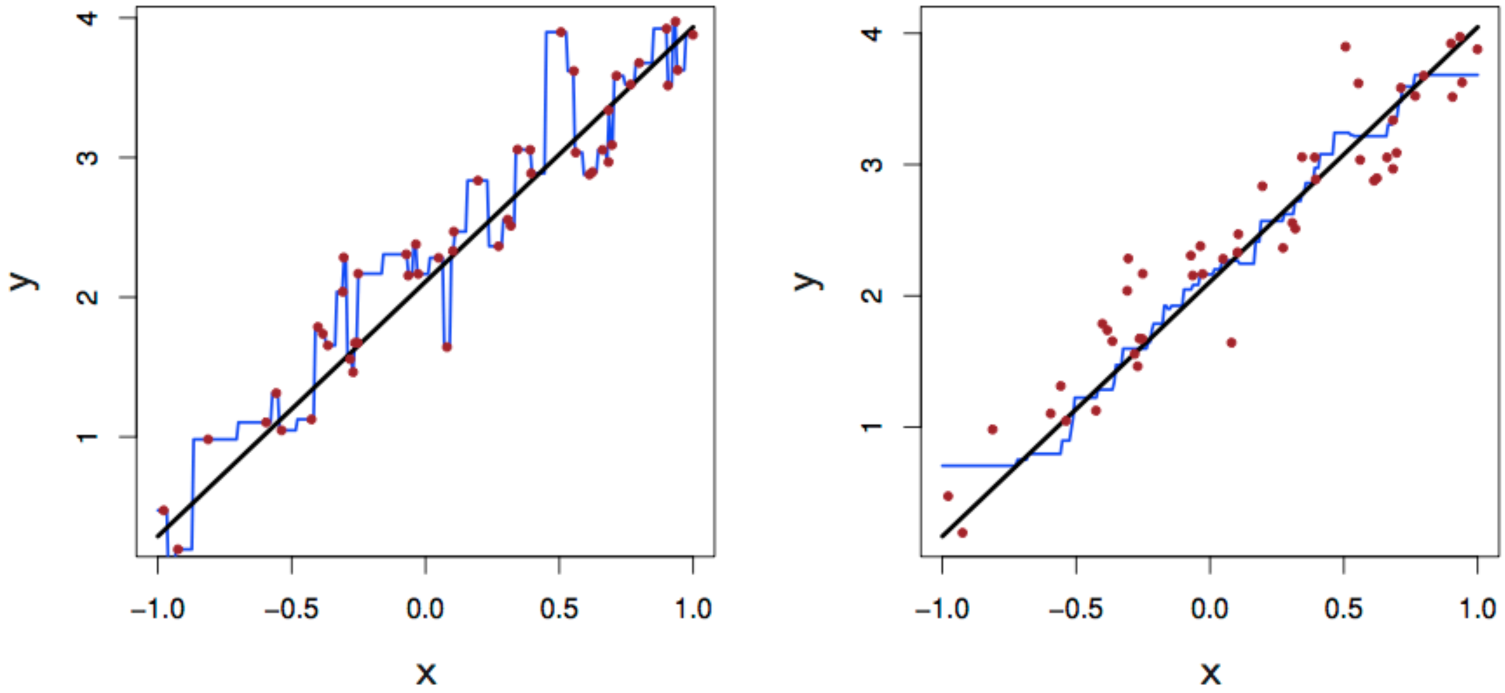
   Reason: ***Curse of dimensionality***. Nearest neighbors tend to be far away in high dimensions.
3. KNN algorithms can be affected by units of your dataset. We

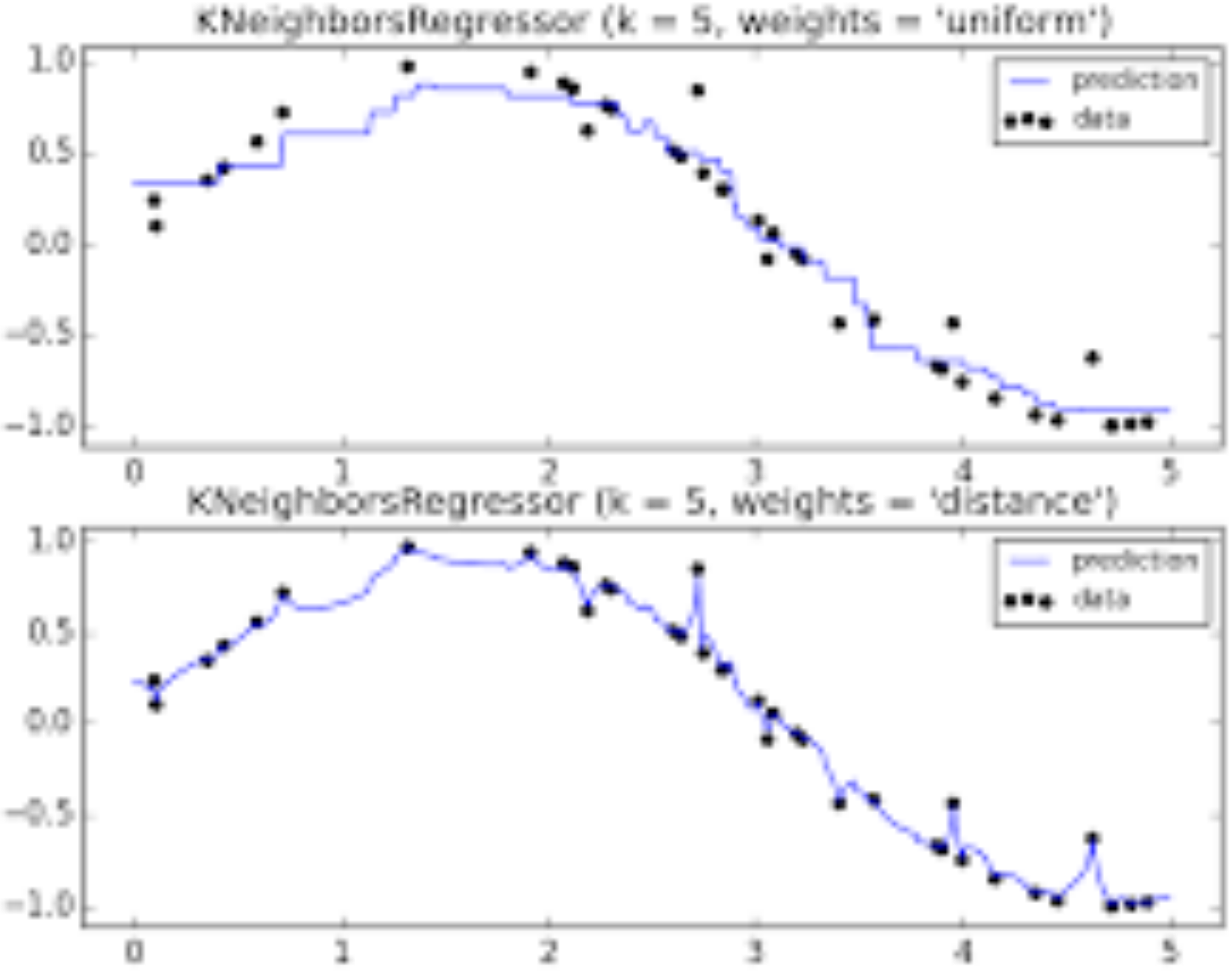
Can resolve this by standardizing our data before training our algorithm. One way to standardize values is (x-min)/(max-min)

K-Nearest Neighbor Algorithm for Regression

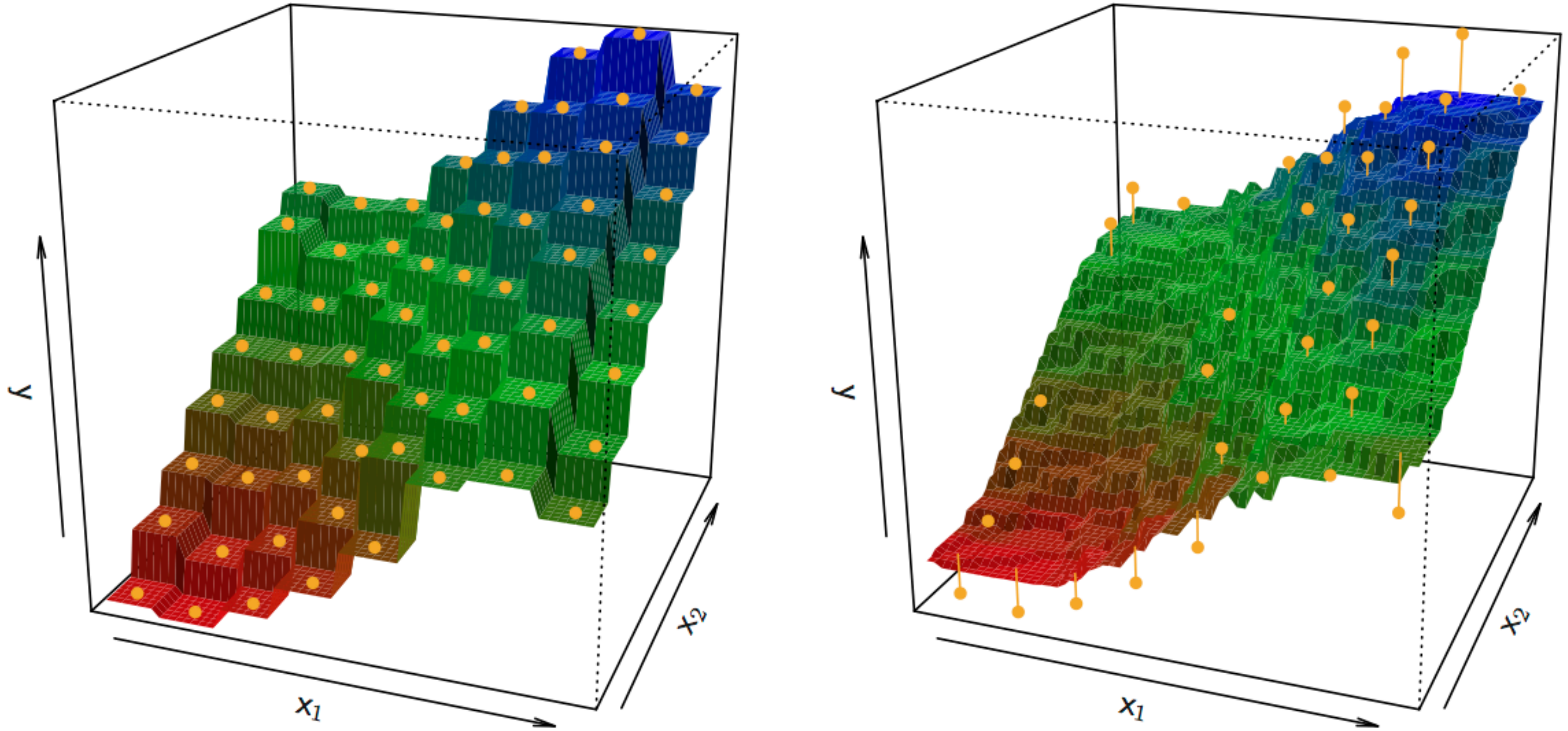
- You can use KNN algorithm for regression as well.
- The algorithm works exactly the same way it works for classification – the only difference is it uses the average of the output of the k-closest observation to your point as your prediction.
- How do we decide on the correct size of k?
  - - Cross-Validation or Validation

KNN Fits in One Dimension (k =1 and k = 9)




KNeighborsRegressor (k = 5, weights = 'uniform')
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5
prediction
data
KNeighborsRegressor (k = 5, weights = 'distance')
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5
prediction
data

KNN Fits for k =1 and k = 9



Summary

- Classification problems
- Misclassification error
- KNN algorithm for Classification
- CV to choose the best k
- Limitations of KNN
- KNN algorithm for regression