K-NEAREST NEIGHBORS

Regression and Classification

INTRODUCTION

- Supervised learning problem
- Output can be either qualitative or quantitative: k-NN can be applied both for data classification and regression
- Shortly saying:
  - while classifying or predicting a new input, we look into the k nearest neighbors of that input and mimic their behavior

K-NN FOR CLASSIFICATION

- In theory we would always like to predict qualitative responses using the Bayes classifier
- However, we do not know the conditional distribution of Y given X for real data

$$P(Y|X)$$

- Many approaches attempt to estimate the conditional distribution of Y given X, and then classify a given observation to the class with the largest probability
- One such method is the k-NN classifier

K-NN CLASSIFIER

- Select parameter k
- Select a distance measure for defining the vicinity of an observation x_0
- Identify the k neighbors in the data that are closest to x_0, represented by N_0
- Estimate the conditional probability for class j as the fraction of points in N_0 whose response values equal j:

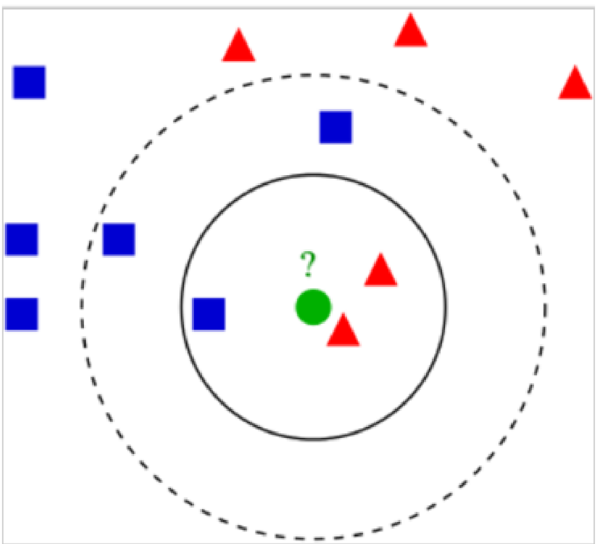
$$P(Y = j | X = x_0) = \frac{1}{k} \sum_{y_k \in N_0} I(y_k = j)$$

- Classify the observation x_0 to the class with the largest conditional probability

K-NN CLASSIFIER

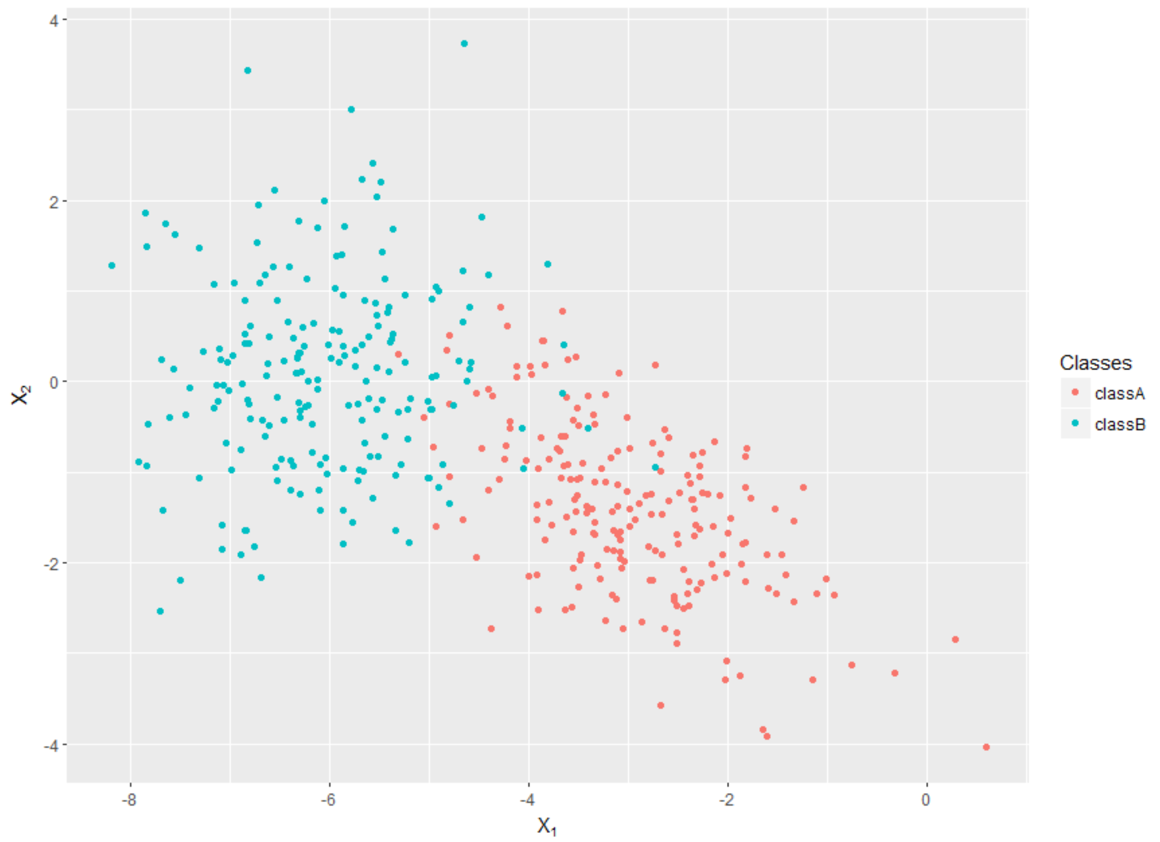
- For many applications, a commonly used distance metric for continuous variables is Euclidian distance
- In some cases it could be useful to assign weights to the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones
- A common weighting scheme will be giving each neighbor a weight of $1/d$, where d is the distance to the neighbor

ILLUSTRATION

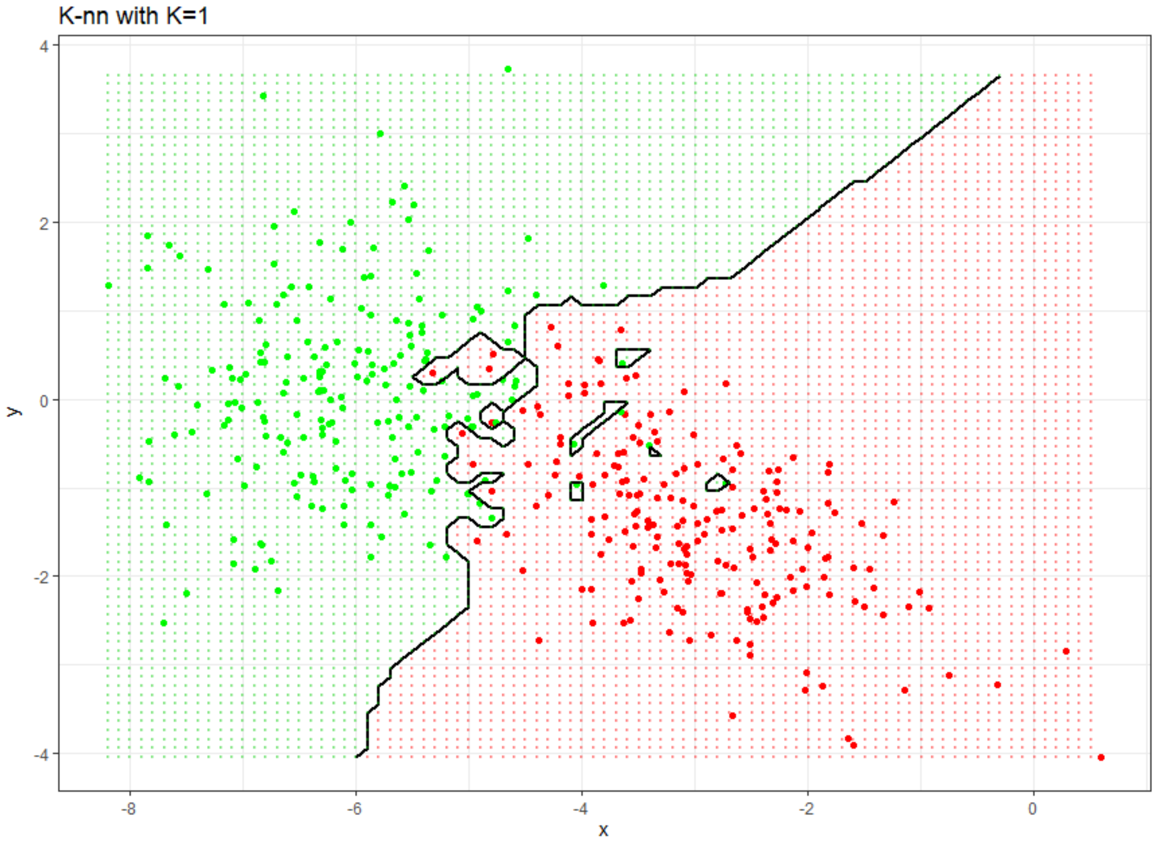


- The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles.
- If $k = 3$ (solid line circle) it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle.
- If $k = 5$ (dashed line circle) it is assigned to the first class (3 squares vs. 2 triangles inside the outer circle).

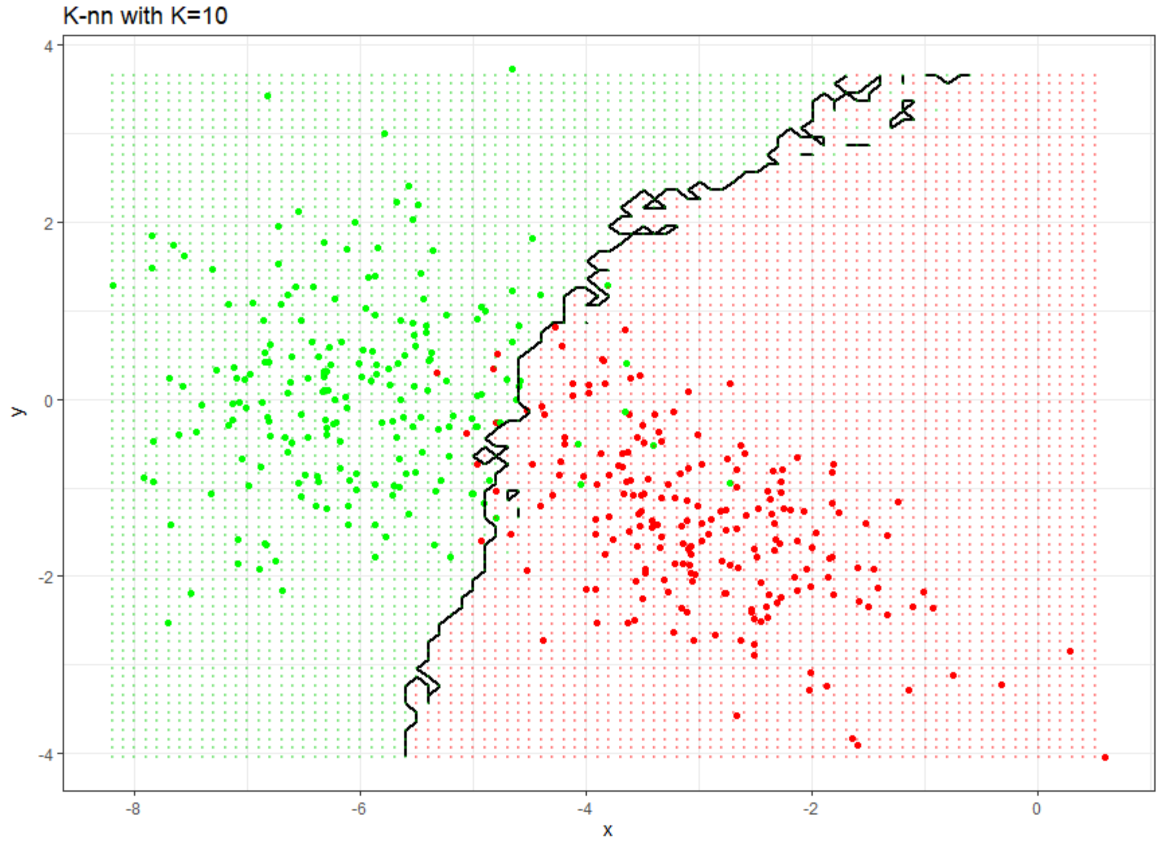
CLASSIFICATION EXAMPLE



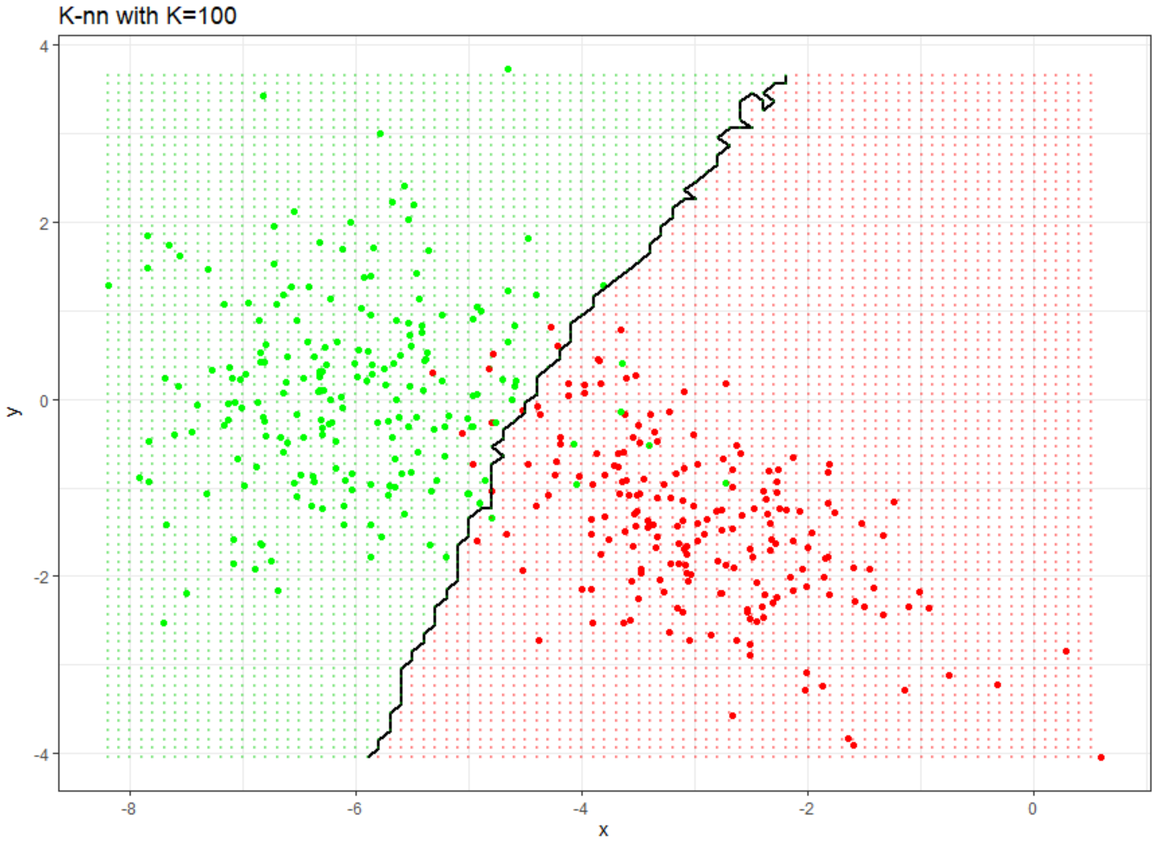
EXAMPLE: K=1



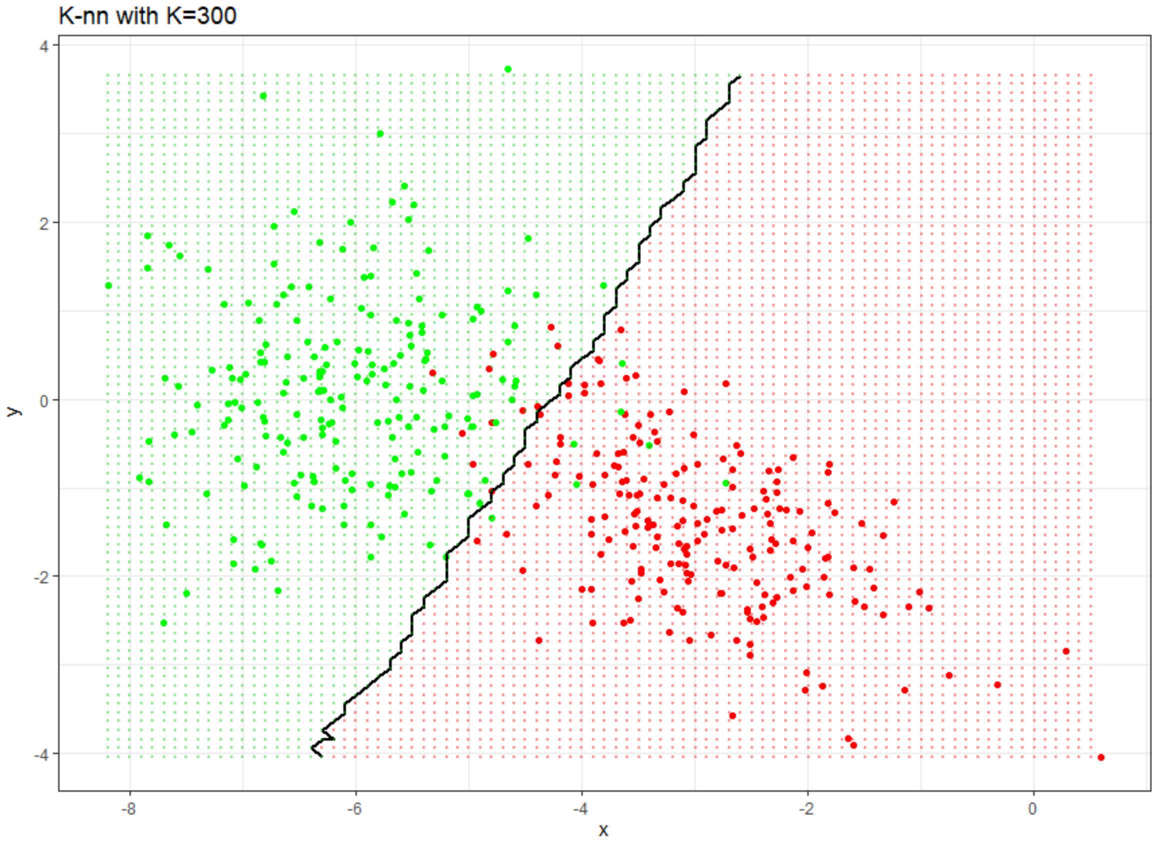
EXAMPLE: K=10



EXAMPLE: K=100



EXAMPLE: K=300



EXAMPLE: EXPLANATION

- On the previous 4 slides, the same example of binary classification problem is shown. k-NN is applied with $k = 1, 10, 100$ and 300.
- The best choice of k depends on data
- Larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct
- In binary classification problems, it is helpful to choose k to be an odd number as this avoids tied votes
- The optimal value of parameter k can be determined via cross-validation

CLASSIFICATION MEASURES

- K=5
- Confusion matrix for a training data

	Model Predicts Class A	Model Predicts Class B
Actual Class A	191	8
Actual Class B	9	192

- Class A is the positive
- $Accuracy = 0.958$
- $Sensitivity = 0.96$
- $Specificity = 0.96$

CLASSIFICATION MEASURES

- $K = 5$
- Confusion matrix for a test data $(0.8 - 0.2)$

	Model Predicts Class A	Model Predicts Class B
Actual Class A	35	2
Actual Class B	5	38

- Class A is the positive
- $Accuracy\ =\ 0.91$
- $Sensitivity\ =\ 0.88$
- $Specificity\ =\ 0.95$

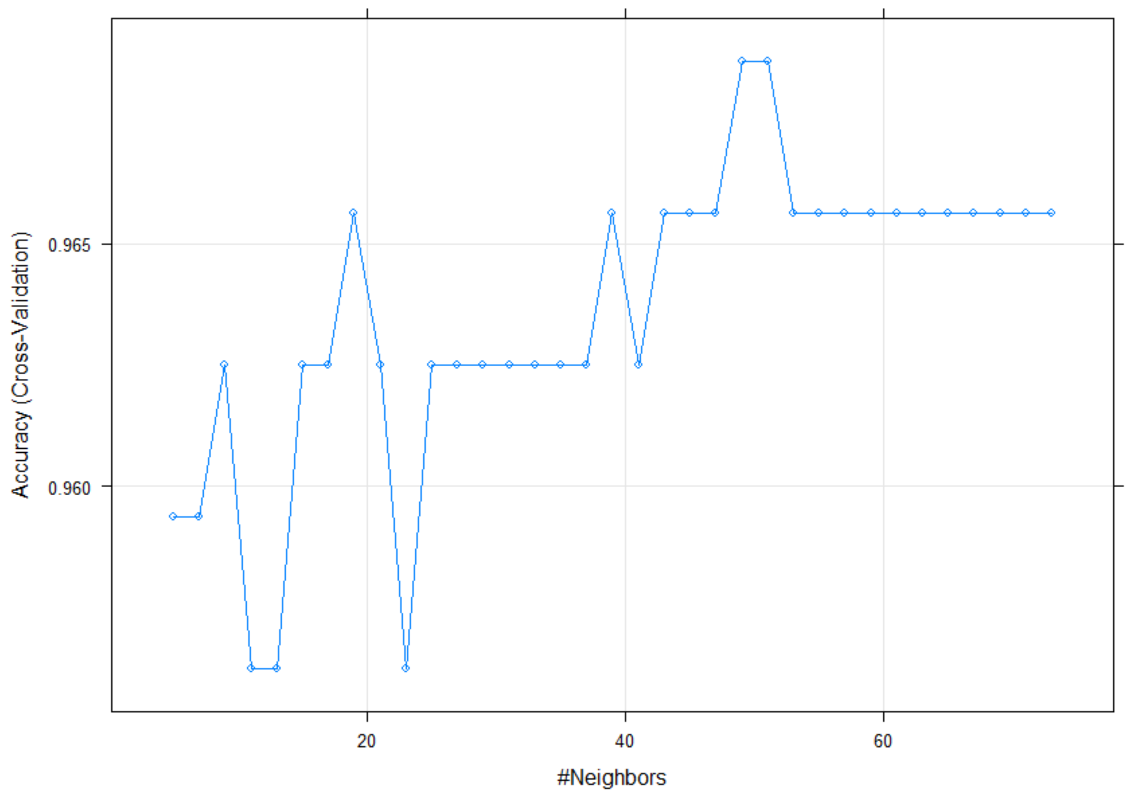
CV ON CLASSIFICATION PROBLEMS

- We can use cross validation in a classification situation in a similar manner
- For example, in case of LOOCV

$$CV_{(n)} = \frac{1}{n}\sum_{k=1}^{n} Err_k$$

where $Err_k = I(y_k \neq \hat{y}_k)$

OPTIMAL K BY CROSS-VALIDATION



OPTIMAL K

- $K = 51$
- Confusion matrix for a test data $(0.8 - 0.2)$

	Model Predicts Class A	Model Predicts Class B
Actual Class A	36	1
Actual Class B	4	39

- Class A is the positive
- $Accuracy = 0.94$
- $Sensitivity = 0.9$
- $Specificity = 0.98$

K-NN ALGORITHM FOR REGRESSION

- The k-NN algorithm can be used both for classification and regression
- In both cases, the input consists of the k closest training examples in the feature space
- In k-NN regression, the output is the average of the values of its k nearest neighbors

$$\hat{y}_0 = \frac{1}{k} \sum_{k \in N_0} y_k$$

ADVANTAGES

- Simple and data driven
- It has power of both linear and non-linear approaches
- Can be applied for both classification and regression
- KNN is a completely non-parametric approach: no assumptions are made about the shape of the decision boundary

DISADVANTAGES

- Complexity significantly increases as the number of variables increase. Large numbers of data sets can take a long to find the record you a near
- It is sensitive to the local structure of the data
- Difficult to find the reasonable value of k
- KNN does not tell us which predictors are important as we did it for logistic regression

COMPARISON WITH LINEAR REGRESSION

- Linear Regression
  - Parametric approach
  - Less flexible than k-NN
  - Has big bias and small variance
- K-NN
  - Non-parametric approach
  - More flexible than linear regression for small k