K Nearest Neighbors (KNN) classifier

Jelena Jovanovic

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

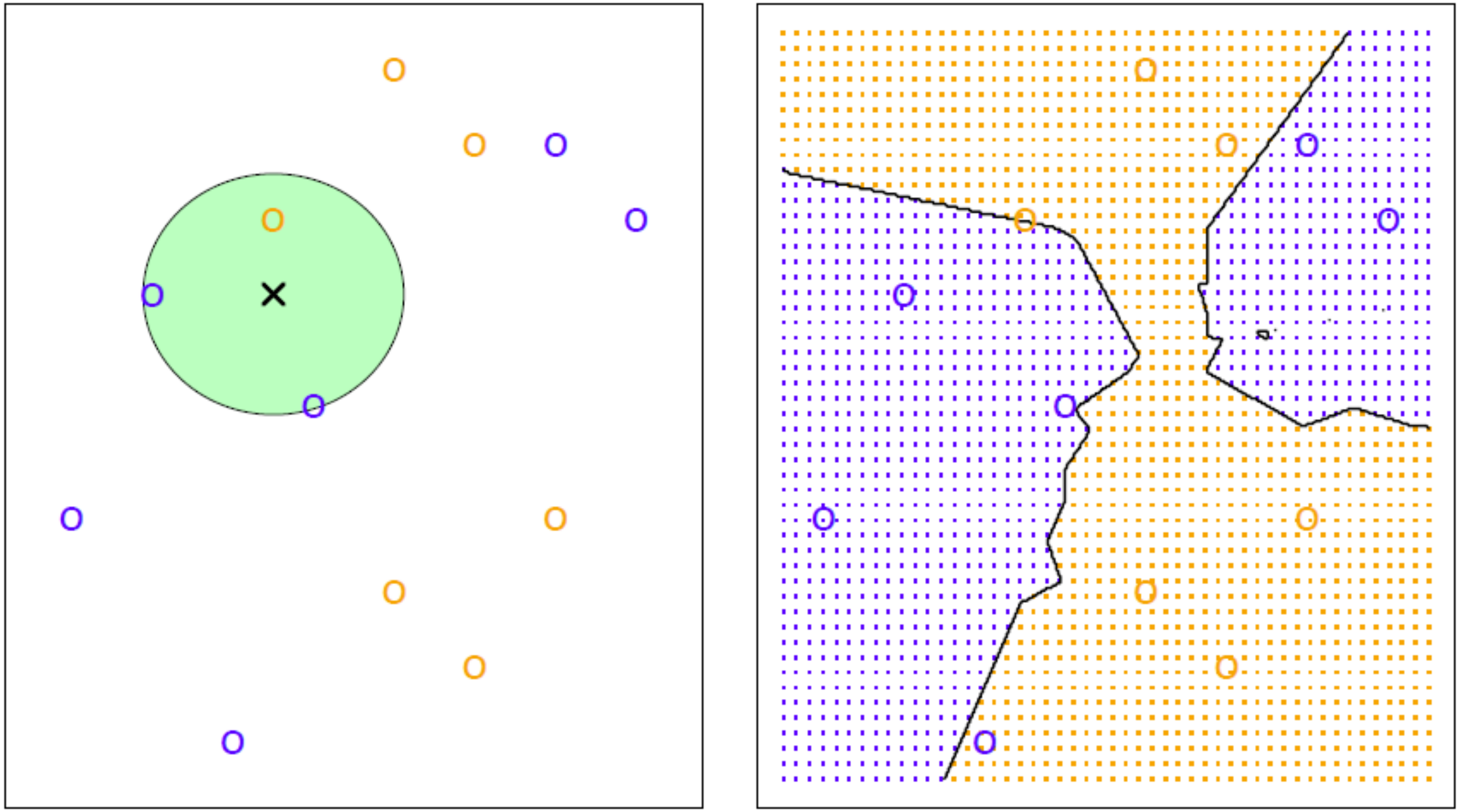
INSTANCE-BASED LEARNING

- kNN belongs to the category of *Instance-based* ML methods
- Common characteristics of these methods:
  - All instances (observations) from the training set are kept in the computer memory
  - When there is a new instances to be classified
    - the method searches for *k* instances from the training set that are closest to the new instance
    - class of the new instance is estimated based on the *k* nearest training instances

K NEAREST NEIGHBORS (KNN)

- One of the basic *Instance-based* ML methods
- It considers all instances as points in an *n*-dimensional space
  - *n* is the number of attributes that describe the instances
- It requires an appropriate metric / technique for computing the closeness / distance between two instances (points in n-dimensional space)
- A new instance is assigned to the most popular / dominant class (*majority class*) among its nearest neighbors

KNN EXAMPLE FOR K = 3



Images are taken from the book “An Introduction to Statistical Learning”, p.40

ESTIMATING THE 'CLOSENESS' OF INSTANCES

Euclidian distance

$$d(x_i, x_j) \equiv \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}$$

- x_i and x_j are instances for which the distance is computed
- $a_1(x), a_2(x), \ldots, a_n(x)$ is the attribute (feature) vector that represents an instance

ESTIMATING THE 'CLOSENESS' OF INSTANCES

Manhattan (taxi-cab) distance

$$d(x_i, x_j) = \sum_{k=1}^{n} \left| a_{ik} - a_{jk} \right|$$

x_i are x_j instances;

$a(x_i)$ and $a(x_j)$ are feature vectors

that represent x_i and x_j

Euclidian vs Manhattan distance

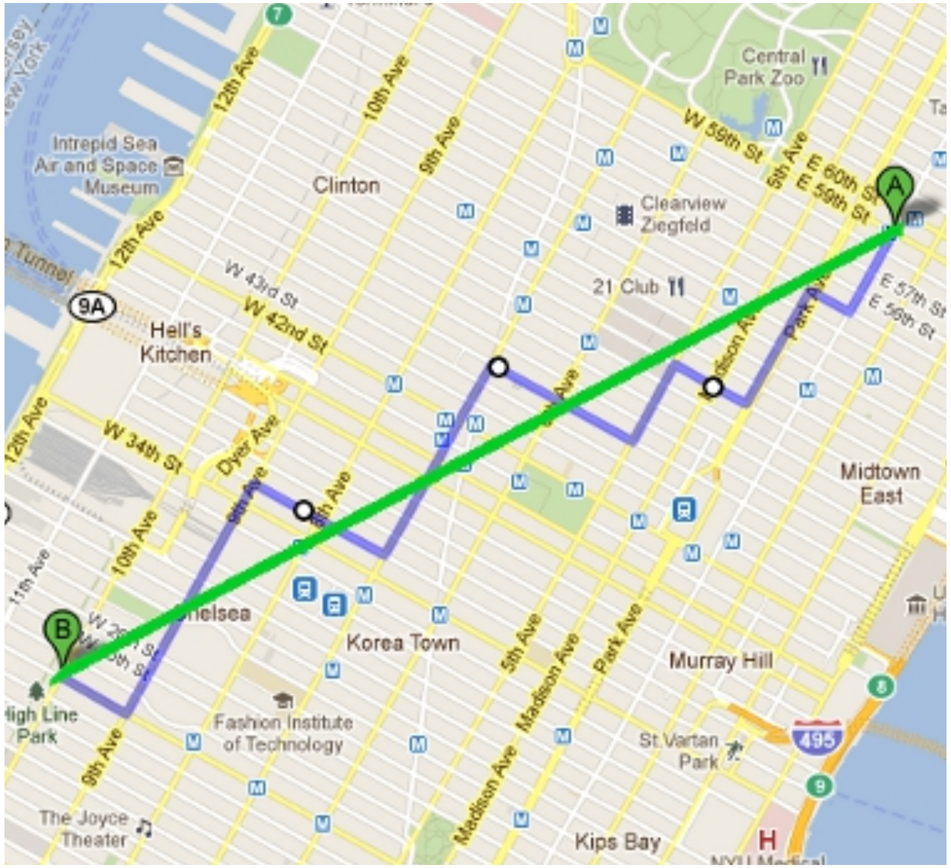


Image source:
http://blog.csdn.net/kikitamoon/article/details/42119415

ESTIMATING THE 'CLOSENESS' OF INSTANCES

- Euclidian and Manhattan metrics are among the most popular; however, depending on the particular classification problem, other metrics, often derived from these basic ones, can be used
- For instance, if we want to put more emphasis on big differences between instances, while paying less attention to smaller differences, we should use metrics with higher degree
  - E.g., metrics that compute 3^{rd} or 4^{th} degree of the sum of attributes differences

NORMALIZING ATTRIBUTE VALUES

- Different attributes are often expressed in different scales
  - E.g., attributes describing a real estate can be number of squared meters, number of rooms, estimated value (price), …
- If we directly apply Euclidian or any other metric, attributes with wider range of values (price) would completely diminish the influence of attributes whose values are in significantly smaller range (num. of rooms)
- Through normalization, values of all the attributes are reduced to the [0,1] range
- A common approach to normalizing attributes:

$$a_i = \frac{v_i - min\ v_i}{max\ v_i - \min v_i}$$

where v_i is the value of the attribute a_i

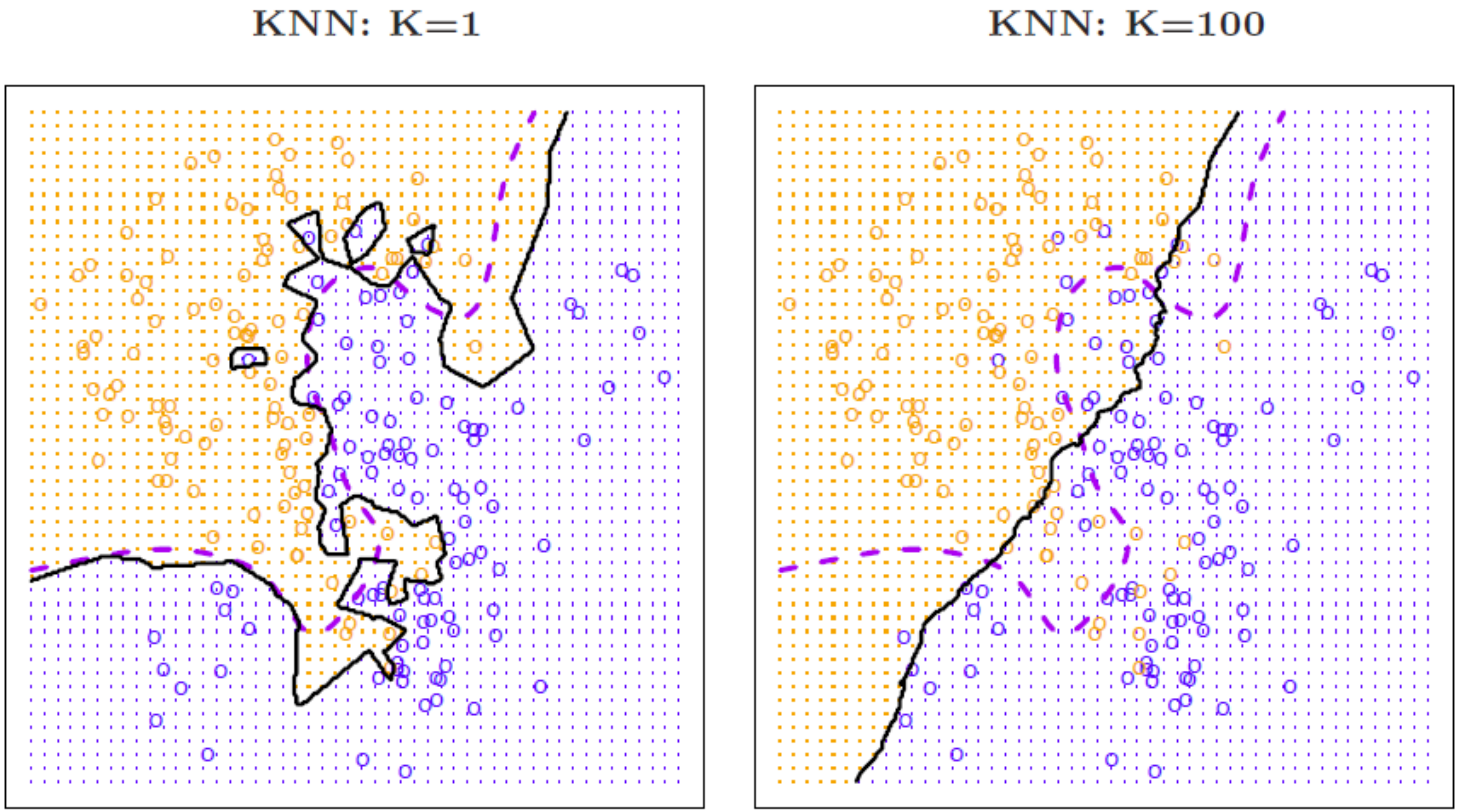
NOMINAL ATTRIBUTES

- In the case of nominal (categorical) attributes, the difference between attribute values (a_{pi} and a_{pj}) of two instances (x_i and x_j) is
  - 1, if $a_{pj} \neq a_{pi}$
  - 0, if $a_{pj} = a_{pi}$

How to choose the parameter K?

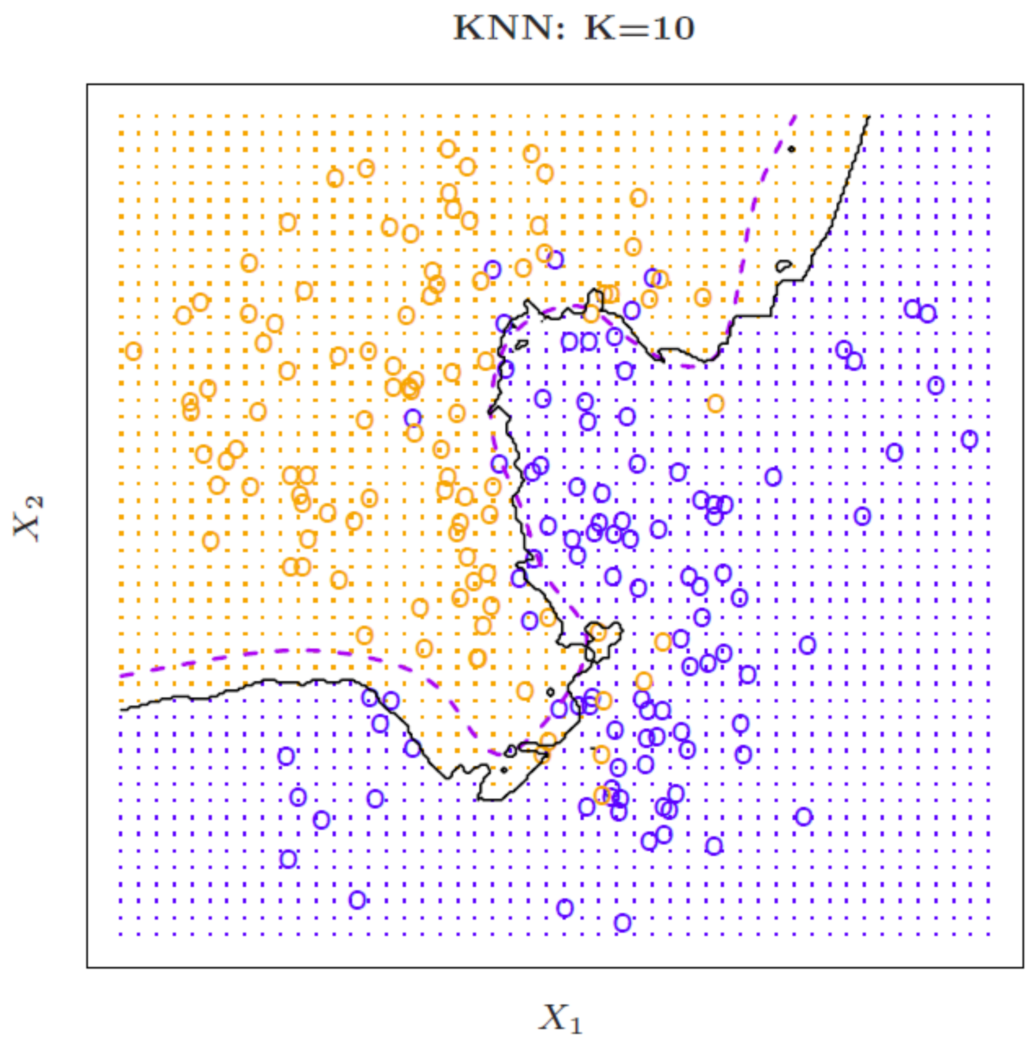
- Cross-validation is typically used to find *K* that guaranties good performance of the classifier, while avoiding the problem of *over-fitting*
- It is recommended to chose an odd number for K, in order to facilitate the selection of the majority class among the *K* nearest neighbors
- In general, the smaller the value of *K*, the more flexible the kNN method will be, i.e., less prone to *over-fitting*

EXAMPLE: KNN OVER-FITTING (K = 1) AND UNDER-FITTING (K = 100)



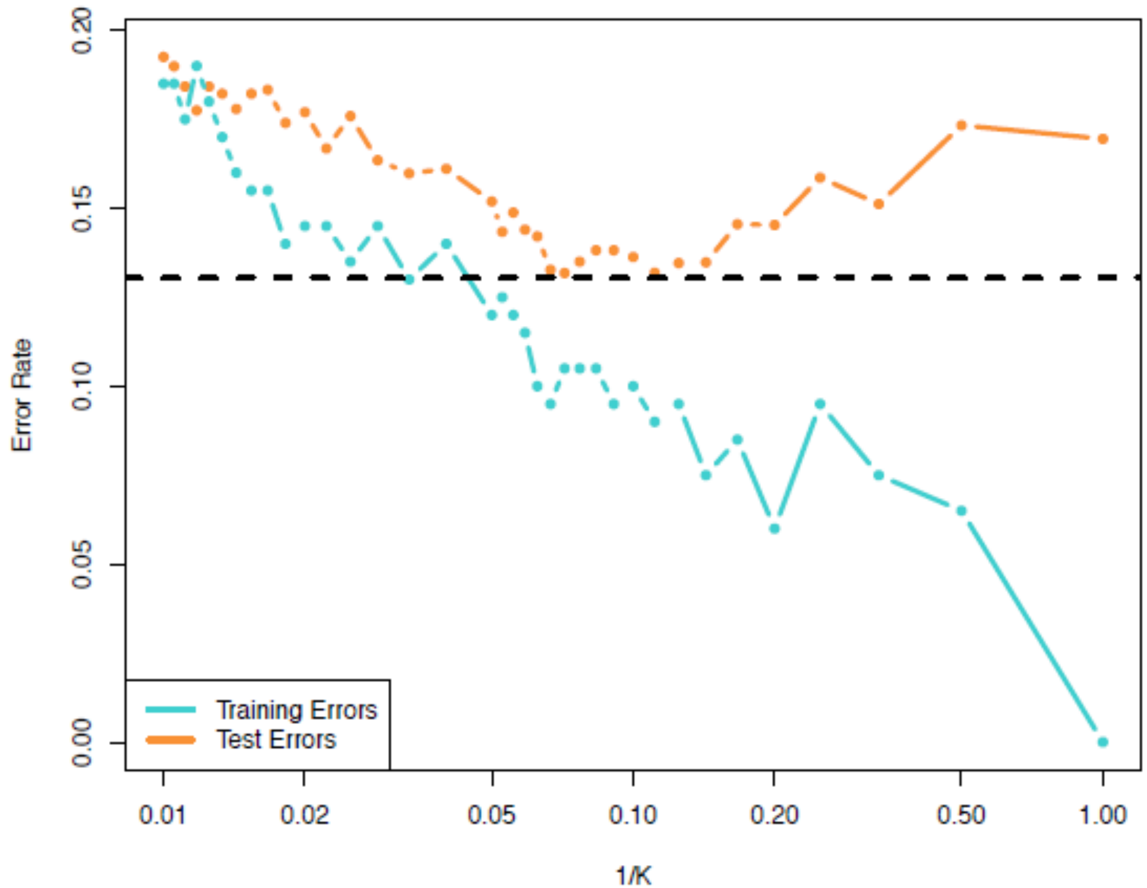
The image is taken form the book “An Introduction to Statistical Learning”, p. 41

EXAMPLE: THE OPTIMAL VALUE FOR K (K = 10) FOR THE GIVEN DATASET



The image is taken form the book “An Introduction to Statistical Learning”, p. 41

COMPARISON OF THE MODEL'S ERRORS ON THE TRAINING AND TEST DATA SETS



As *K* decreases, the error on the training set also goes down; however, on the test set, the decrease in the error stops in one moment (K=10), and from that point onwards the test error keeps increasing

The image Is taken form the book "An Introduction to Statistical Learning", p. 42

ACKNOWLEDGEMENT

These slides are partially based on:

- Chapter 2 of the book “An Introduction to Statistical Learning” (link)
- Chapter 4.7 of the book “Data Mining: Practical Machine Learning Tools and Techniques. 3rd Edition” (link)
- Slide deck “k Nearest Neighbor” downloaded from SlideShare.net (link)