k-Nearest Neighbor Method

(Course: Introduction to Data Science)

Tirtharaj Dash

BITS Pilani, K.K. Birla Goa Campus

tirtharaj@goa.bits-pilani.ac.in

November 12, 2020

Introduction I

- k-NN is an instance based learning method for supervised learning problems.
- Instance-based learning is often termed lazy learning, as there is typically no "transformation" of training instances.
- Instead, the presented training data is simply stored and, when a new query instance is encountered, a set of similar, related instances is retrieved from memory and used to classify the new query instance.
- Here goal is not to learn any (target) function, rather to just make predictions.

Introduction II

- k-NN assumes that all instances are points in some d-dimensional space and defines neighbors in terms of distance (usually Euclidean in $\mathbb{R}$-space).
- k in k-NN is the number of neighbors considered.

Basic idea I

- The k-NN classification rule is to assign to a test sample the majority category label of its k nearest training samples.
- In practice, k is usually chosen to be odd.
  1. Why?

     Answer. to avoid ties
- The $k = 1$ rule is generally called the nearest-neighbor classification rule.

Basic idea II

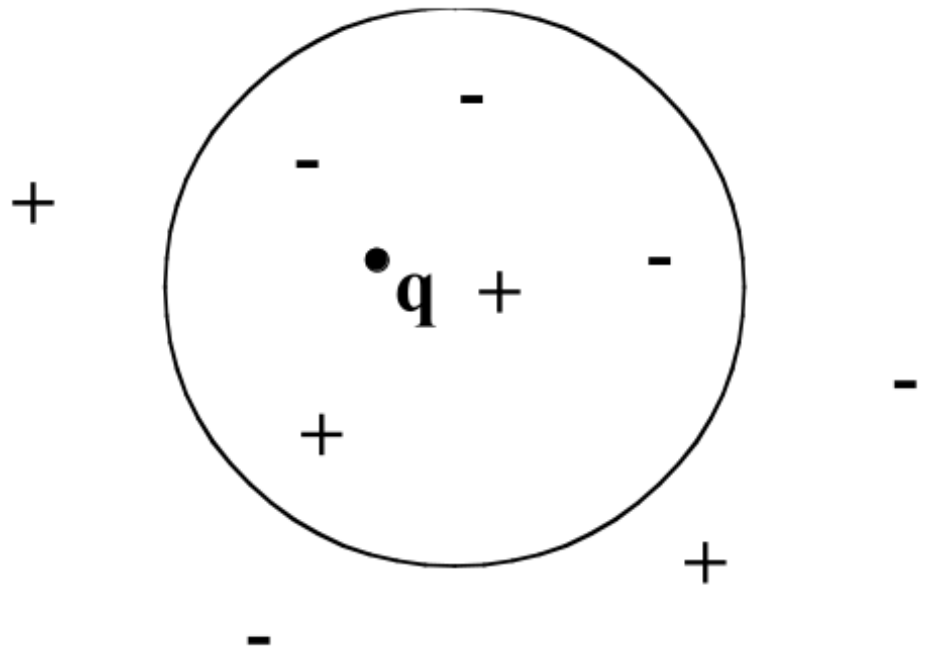
k-NN Procedure:

1. Instance_set $= \emptyset$
2. For each instance with its label $(\boldsymbol{x}, y)$
   a. Instance_set $=$ Instance_set $\bigcup (\boldsymbol{x}, y)$
3. Given a query instance $\boldsymbol{x}'$
   a. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ be the set of instances $\in$ Instance_set, $\boldsymbol{x}'$ is nearest to.
   b. Return majority class from the classes of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$.

Note: Note that this procedure is easy to modify for regression problems. A straightforward modification can be in Line 3.b: return mean of the true outputs of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$.

Basic idea III

What does this mean?



For the above diagram: q is the query instance. The k-NN will return '+' if $k = 1$, and '–' if $k = 5$.

Effects of scale I

- An instance consists of many features. Different features may have different measurement scales.
  - E.g. patient weight in kg (range $[50, 200]$) vs. blood protein values in ng/dL (range $[-3, 3]$)
- Effects:
  - Patient weight will have a much greater influence on the distance between samples.
  - May bias the performance of the classifier.

Effects of scale II

2. How will you solve the problem?

   Answer. Standardisation of data.

How are neighbors determined?

- Using a distance measure: Read about various distance measures that we have already studied in previous lectures.
- There are various issues with distance measures that we will see next. However, students are advised to read about the cases when the instance is described by heterogeneous attributes, i.e. mixture of discrete and real-valued attributes.

Some remarks

- k-NN works well on many practical problems and is fairly noise tolerant (depending on the value of k).
- k-NN is subject to the curse of dimensionality (i.e., presence of numerous irrelevant attributes).
- k-NN needs adequate distance measure.
- k-NN relies on efficient indexing of the (training) instances.

Homework

3. Implement k-NN for Iris data classification problem. You may choose to use Scikit-learn. However, I would suggest that you write the complete code from scratch.
4. Solve the following numerical problem. Assume the distance metric to be L_1-norm distance.

   Given the dataset: $((35, 14), 1)$, $((36, 28), 2)$, $((7, 10), 1)$, $((15, 14), 2)$, $((16, 28), 2)$, $((37, 1), 1)$. Use $k = 3$ to find out the class for the query data $(17, 19)$.