K-NN

15.097 MIT, Spring 2012, Cynthia Rudin
Credit: Seyda Ertekin

K-Nearest Neighbors

• Amongst the simplest of all machine learning algorithms. No explicit training or model.
• Can be used both for classification and regression.
• Use x’s K-Nearest Neighbors to vote on what x’s label should be.
K-Nearest Neighbors

- Classify using the majority vote of the k closest training points

(a) 1-nearest neighbor  (b) 2-nearest neighbor  (c) 3-nearest neighbor

K-Nearest Neighbors

- K-NN algorithm does not explicitly compute decision boundaries. The boundaries between distinct classes form a subset of the Voronoi diagram of the training data.

Each line segment is equidistant to neighboring points.
K-Nearest Neighbors

• **For regression:** the value for the test example becomes the (weighted) average of the values of the K neighbors.

Making K-NN More Powerful

• A good value for K can be determined by considering a range of K values.
  – K too small: we’ll model the noise
  – K too large: neighbors include too many points from other classes

• There are problems when there is a spread of distances among the K-NN. Use a distance-based voting scheme, where closer neighbors have more influence.

• The distance measure has to be meaningful – attributes should be scaled
  – Eg. Income varies 10,000-1,000,000 while height varies 1.5-1.8 meters
Pros/Cons to K-NN

Pros:
• Simple and powerful. No need for tuning complex parameters to build a model.
• No training involved (“lazy”). New training examples can be added easily.

Pros/Cons to K-NN

Cons:
• Expensive and slow: $O(md)$, $m=$ # examples, $d=$ # dimensions
  – To determine the nearest neighbor of a new point $x$, must compute the distance to all $m$ training examples. Runtime performance is slow, but can be improved.
  • Pre-sort training examples into fast data structures
  • Compute only an approximate distance
  • Remove redundant data (condensing)
## K-NN Applications

- **Handwritten character classification using nearest neighbor in large databases.**

- **Fast content-based image retrieval based on equal-average K-nearest-neighbor search schemes.**
  CBIR (Content based image retrieval), return the closest neighbors as the relevant items to a query.

- **Use of K-Nearest Neighbor classifier for intrusion detection.**
  Classify program behavior as normal or intrusive.

- **Fault Detection Using the k-Nearest Neighbor Rule for Semiconductor Manufacturing Processes.**
  Early fault detection in industrial systems.