k-Nearest Neighbor (kNN)

Sept. 2015
Youn-Hee Han
http://link.koreatech.ac.kr

Eager vs. Lazy Learning

- Eager Learners
  - when given a set of training data, it will construct a generalization model before receiving new (e.g., test) data to classify
    - Classification by decision tree induction
    - Common Linear Regression and Logistic Regression
    - Support Vector Machines (SVM)

- Lazy Learners
  - Simply stores training data (or only minor processing) and waits until it is given a test tuple
  - less time in training but more time in predicting
    - KNN
    - Locally weighted regression

Eager vs. Lazy Learning

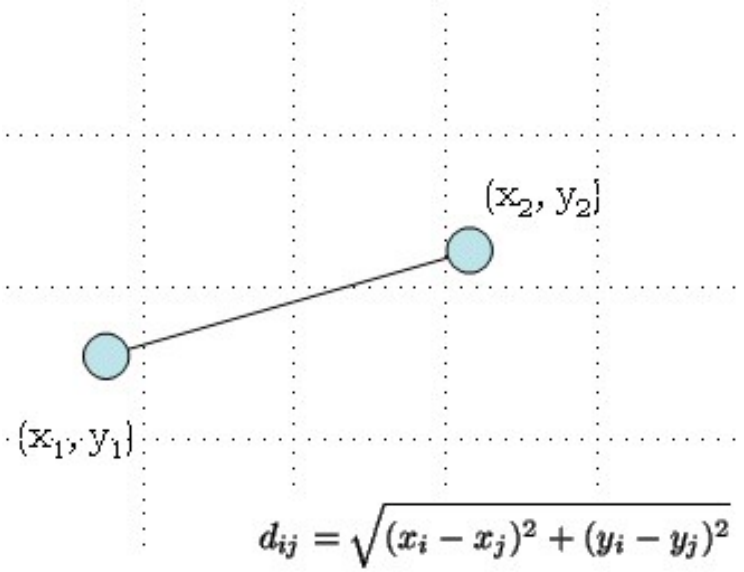
- Accuracy
  - Lazy learners effectively uses a richer hypothesis space since it uses many local linear functions to form its implicit global approximation
  - Eager learners must commit to a single hypothesis that covers the entire instance space

- Instance-Based Methods (Lazy Learners)
  - Store training examples and delay the processing ("lazy evaluation") until a new instance must be classified
  - KNN

k-Nearest Neighbor Algorithm

◈Algorithm (1/2)

- All training data instances correspond to points in the n-D (Euclidean) space
- The nearest neighbor are defined in terms of Euclidean distance, d_{ij}



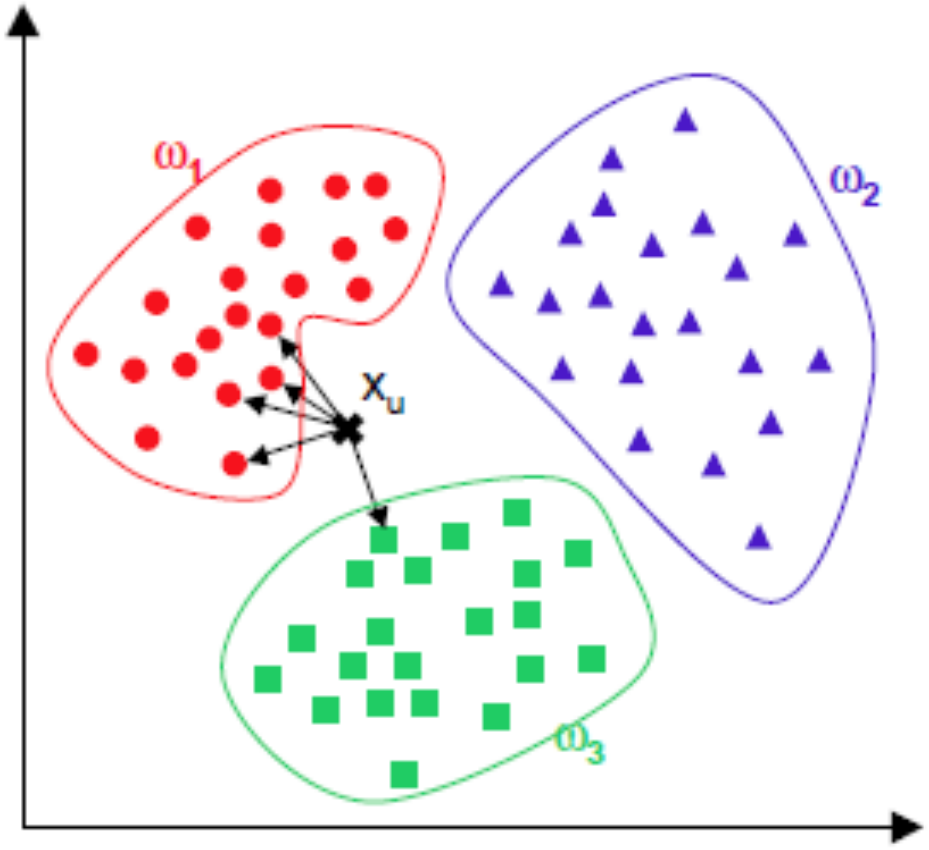
k-Nearest Neighbor Algorithm

- Algorithm (2/2)
  - For an unknown case y…
  - If k=1, select the nearest neighbor
  - If k>1,
    - For classification, select the most frequent neighbor.
      - ranking yields k feature vectors and store a set of k class labels
      - pick the class label which is most common in this set (“vote”)
      - classify y as belonging to this class
    - For regression, calculate the average of k neighbors.

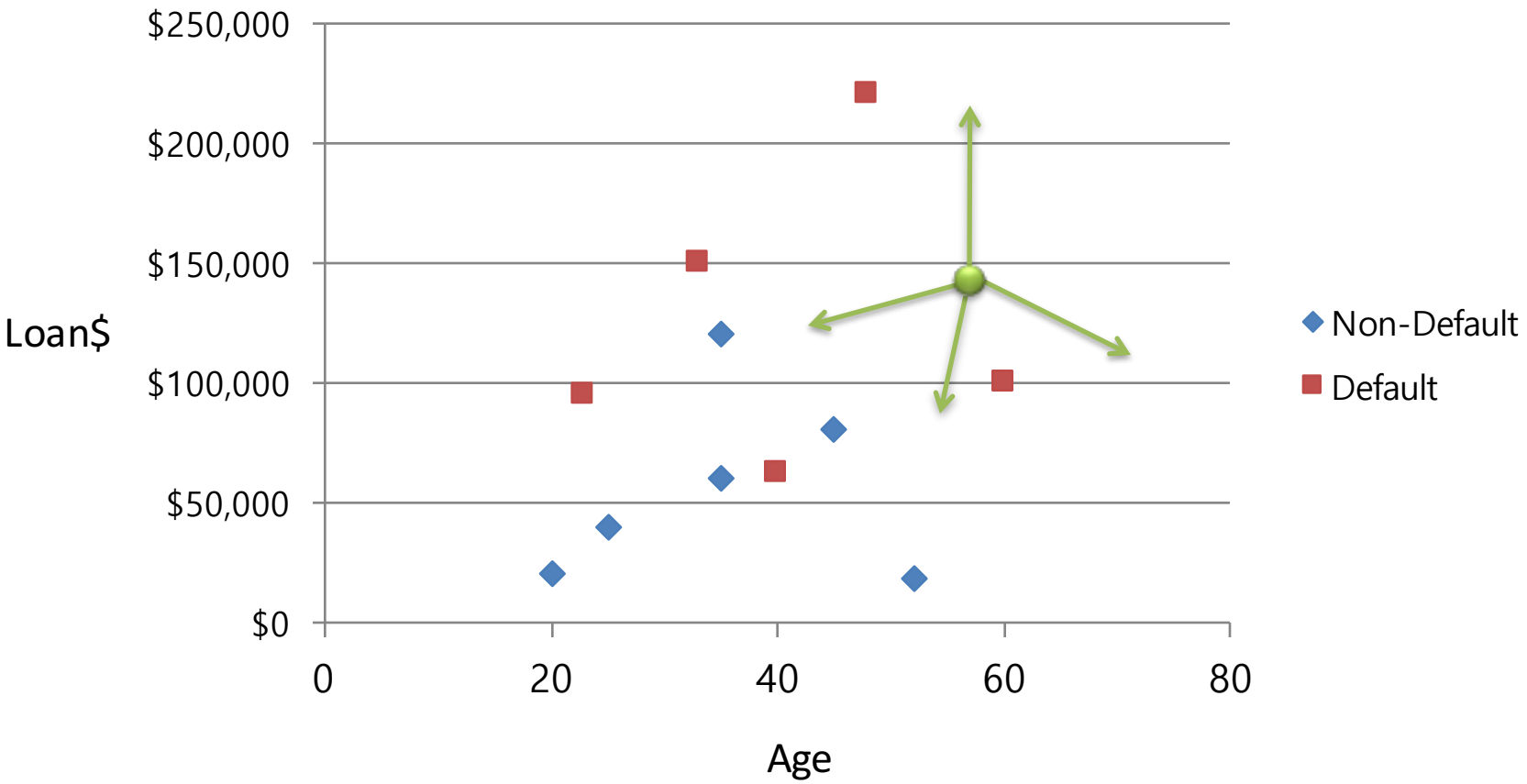
The *k*-Nearest Neighbor Algorithm

◈Example:

K=5



kNN Classification



KNN Classification – Distance

Age	Loan	Default	Distance
25	$40,000	N	102000
35	$60,000	N	82000
45	$80,000	N	62000
20	$20,000	N	122000
35	$120,000	N	22000
52	$18,000	N	124000
23	$95,000	Y	47000
40	$62,000	Y	80000
60	$100,000	Y	42000
48	$220,000	Y	78000
33	$150,000	Y	8000
48	**$142,000**	**?**	

Euclidean Distance

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

KNN Classification – Standardized Distance

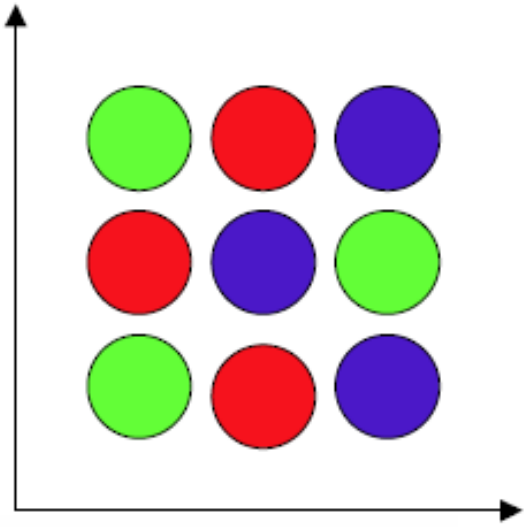
Age	Loan	Default	Distance
0.125	0.11	N	0.7652
0.375	0.21	N	0.5200
0.625	0.31	N	0.3160
0	0.01	N	0.9245
0.375	0.50	N	0.3428
0.8	0.00	N	0.6220
0.075	0.38	Y	0.6669
0.5	0.22	Y	0.4437
1	0.41	Y	0.3650
0.7	1.00	Y	0.3861
0.325	0.65	Y	0.3771
0.7	**0.61**	**?**	

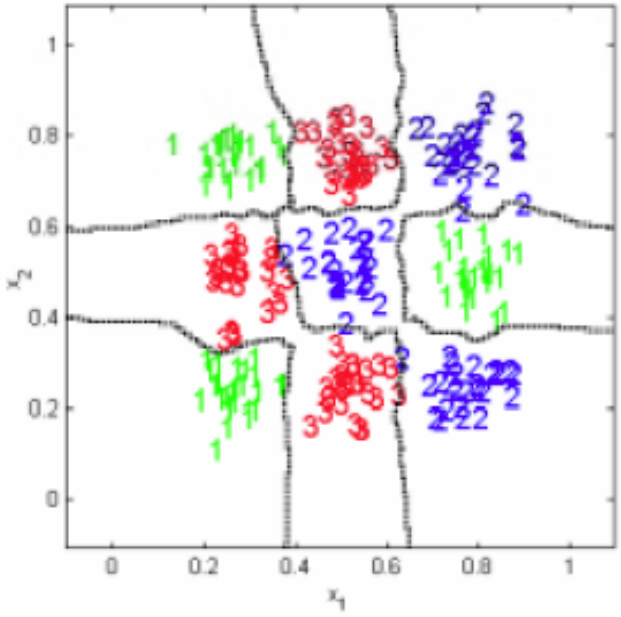
Standardized Variable

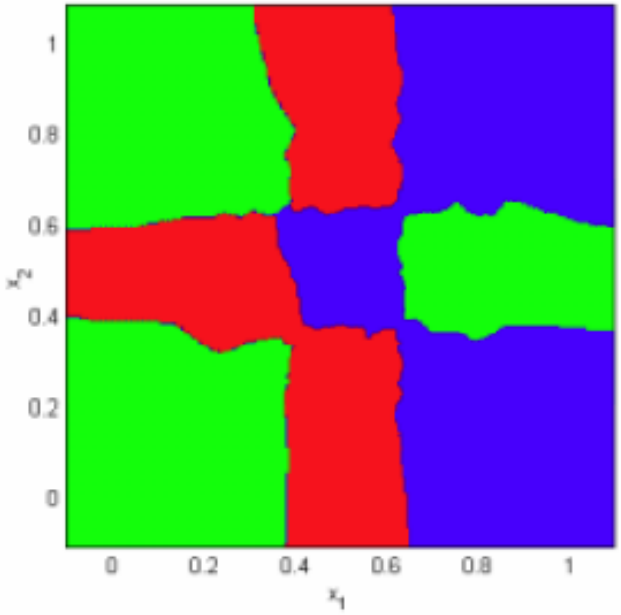
$$X_s = \frac{X - Min}{Max - Min}$$

k-NNR in action: example 1

- We have generated data for a 2-dimensional 3-class problem, where the class-conditional densities are multi-modal, and non-linearly separable, as illustrated in the figure
- We used the k-NNR with
  - k = five
  - Metric = Euclidean distance
- The resulting decision boundaries and decision regions are shown below

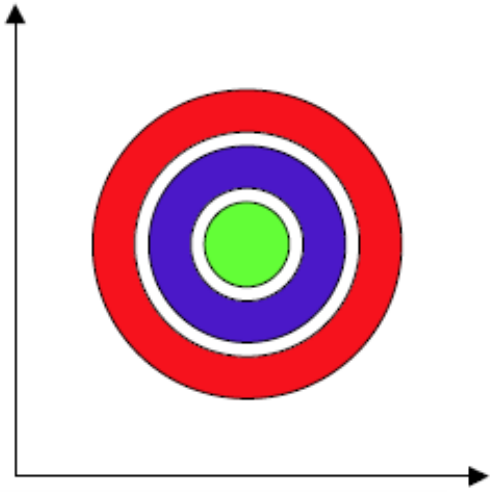


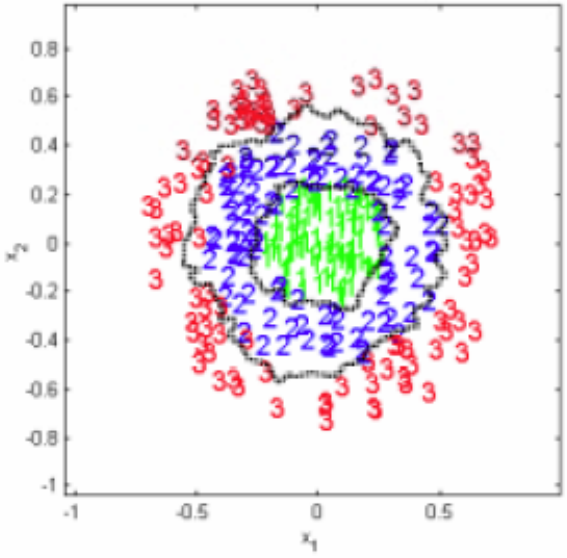


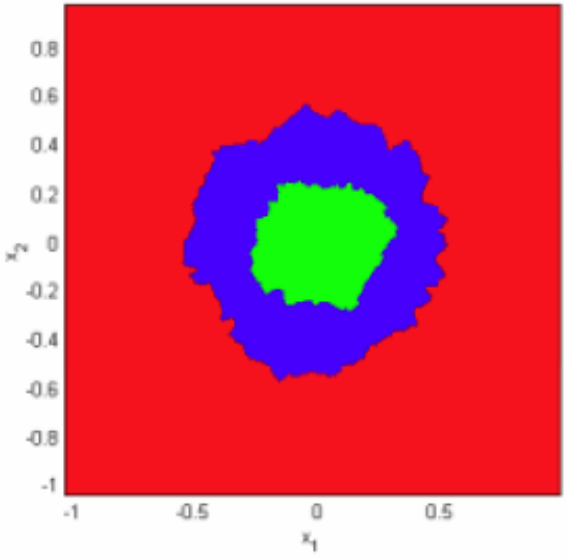


k-NNR in action: example 2

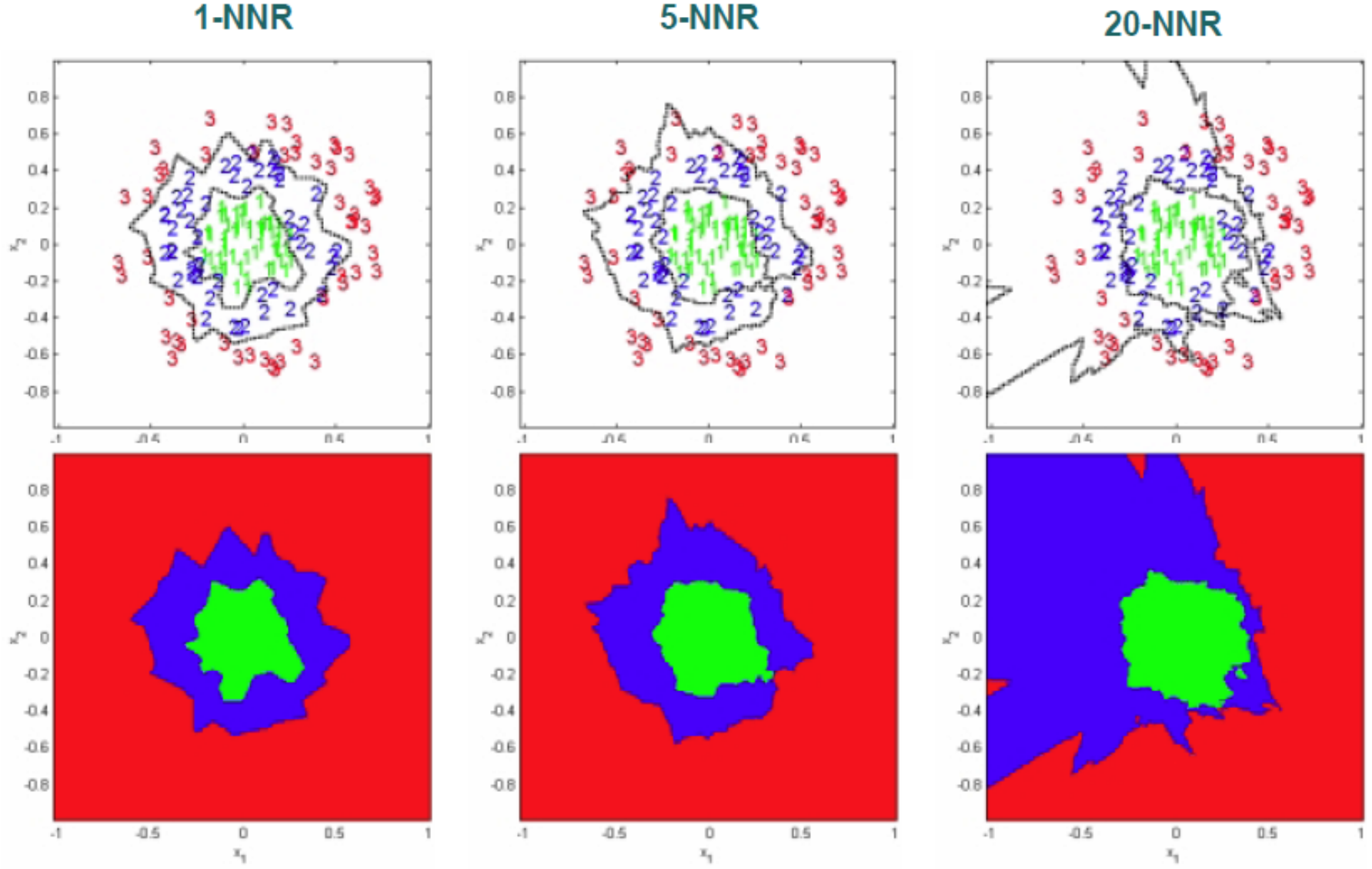
- We have generated data for a 2-dimensional 3-class problem, where the class-conditional densities are unimodal, and are distributed in rings around a common mean. These classes are also non-linearly separable, as illustrated in the figure
- We used the k-NNR with
  - k = five
  - Metric = Euclidean distance
- The resulting decision boundaries and decision regions are shown below







k-NNR versus 1-NNR



kNN Features

- Features
  - for two-class problems, if we choose k to be odd (i.e., k=1, 3, 5,…) then there will never be any "ties"
  - Learning is simple (no learning at all!)
    - "training" is trivial for the kNN classifier
    - we just use the training set as a "lookup table" when we want to classify a new feature vector
  - KNN is conceptually simple, yet able to solve complex problems
  - Can work with relatively little information
  - Memory and CPU cost
  - Sensitive to feature representation

Value of k

- How can I determine the value of k, the number of neighbors?
  - In general, the larger the number of training tuples is, the larger the value of k is
- Theoretical Considerations
  - as k increases
    - we are averaging over more neighbors
    - the effective decision boundary is more “smooth”
  - as N (the number of training data set) increases, the optimal k value tends to increase in proportion to logN

kNN Regression

- Real-valued prediction for a given unknown data
  - Returns the mean values of the k nearest neighbors

Age	Loan	House Price Index	Distance
25	$40,000	135	102000
35	$60,000	256	82000
45	$80,000	231	62000
20	$20,000	267	122000
35	$120,000	139	22000
52	$18,000	150	124000
23	$95,000	127	47000
40	$62,000	216	80000
60	$100,000	139	42000
48	$220,000	250	78000
33	$150,000	264	8000
48	**$142,000**	**?**	

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

kNN Regression

Age	Loan	House Price Index	Distance
0.125	0.11	135	0.7652
0.375	0.21	256	0.5200
0.625	0.31	231	0.3160
0	0.01	267	0.9245
0.375	0.50	139	0.3428
0.8	0.00	150	0.6220
0.075	0.38	127	0.6669
0.5	0.22	216	0.4437
1	0.41	139	0.3650
0.7	1.00	250	0.3861
0.325	0.65	264	0.3771
0.7	**0.61**	**?**	

$$X_s = \frac{X - Min}{Max - Min}$$

kNN Variations

- Distance-weighted nearest neighbor algorithm
  - Weight the contribution of each of the k neighbors according to their distance to the unknown y
  - Give greater weight to closer neighbors

- Weighted features
  - if some of features are more important, more weight is considered for the features
  - if some of features are irrelevant, less weight is considered for the features

kNN Names

- kNN Different Names
  - K-Nearest Neighbors
  - Memory-Based Reasoning
  - Example-Based Reasoning
  - Instance-Based Learning
  - Case-Based Reasoning
  - Lazy Learning

kNN Examples

- Speech Recognition
  - Training data: words recorded and labeled in a database
  - Test data: words recorded from new speakers, new locations
- Zipcode Recognition
  - Training data: zipcodes manually selected, scanned, labeled
  - Test data: actual letters being scanned in a post office
- Credit Scoring
  - Training data: historical database of loan applications with payment history or decision at that time
  - Test data: you