Data Mining
(EECS 6412)

K-Nearest Neighbor Classifier
(Chapter 9.5)

K-Nearest Neighbor Classifiers

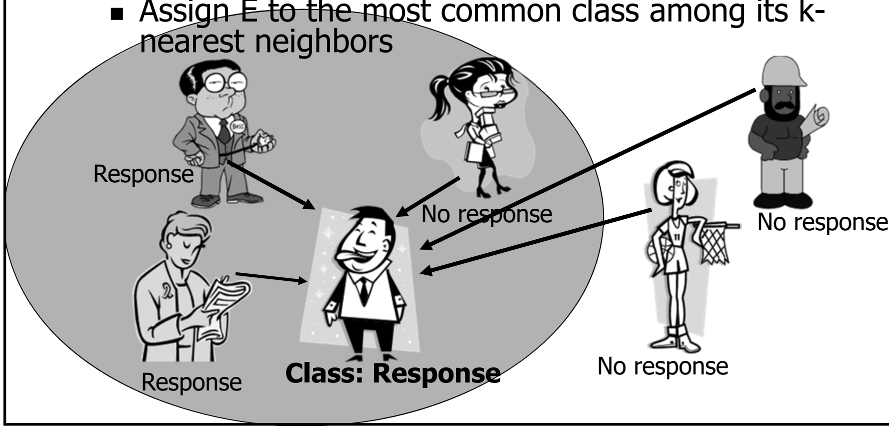
Learning by analogy:

Tell me who your friends are and I'll tell you who you are

A new example is assigned to the most common class among the (K) examples that are most similar to it.

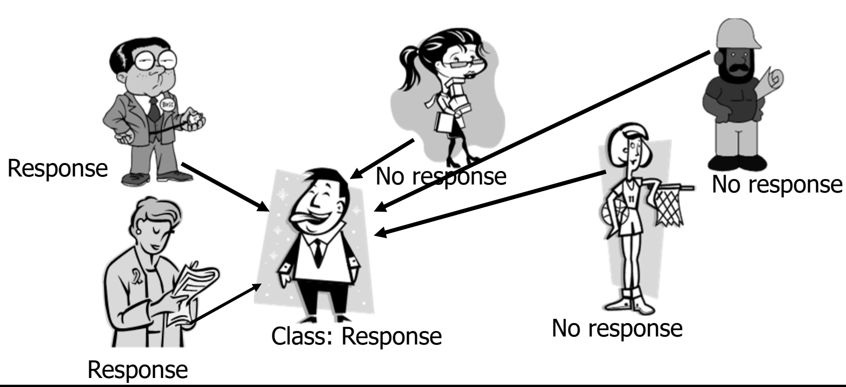
K-Nearest Neighbor Algorithm (k-NN)

- To determine the class of a new example E:
  - Calculate the distance between E and all examples in the training set
  - Select k nearest examples to E in the training set
  - Assign E to the most common class among its k-nearest neighbors



K-Nearest Neighbor: Instance Based Learning

- No model is built: Store all training examples
- Any processing is delayed until a new instance must be classified.



Distance Between Neighbors

- Each example is represented with a set of numerical attributes

John:
Age=35
Income=95K
No. of credit cards=3

Rachel:
Age=41
Income=215K
No. of credit cards=2

- “Closeness” can be defined in terms of the *Euclidean* distance between two examples.
  - The Euclidean distance between $X=(x_1, x_2, x_3, \ldots x_n)$ and $Y=(y_1, y_2, y_3, \ldots y_n)$ is defined as:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

  - Distance (John, Rachel)=sqrt $[(35\text{-}41)^2+(95K\text{-}215K)^2 +(3\text{-}2)^2]$

5

Example : 3-Nearest Neighbors

Customer	Age	Income	No. credit cards	Response
John 	35	35K	3	No
Rachel 	22	50K	2	Yes
Hannah 	63	200K	1	No
Tom 	59	170K	1	No
Nellie 	25	40K	4	Yes
David 	37	50K	2	?

6

Example (3-NN)

Customer	Age	Income (K)	No. cards	Response	Distance from David
John 	35	35	3	No	sqrt $[(35-37)^2+(35-50)^2+(3-2)^2]$=**15.16**
Rachel 	22	50	2	Yes	sqrt $[(22-37)^2+(50-50)^2+(2-2)^2]$=**15**
Hannah 	63	200	1	No	sqrt $[(63-37)^2+(200-50)^2+(1-2)^2]$=**152.23**
Tom 	59	170	1	No	sqrt $[(59-37)^2+(170-50)^2+(1-2)^2]$=**122**
Nellie 	25	40	4	Yes	sqrt $[(25-37)^2+(40-50)^2+(4-2)^2]$=**15.74**
David 	37	50	2	**Yes**	

7

A Problem and its Solution

John:
Age=35
Income=95K
No. of credit cards=3

Rachel:
Age=41
Income=215K
No. of credit cards=2

Distance (John, Rachel)=sqrt $[(35-45)^2$**$+(95{,}000-215{,}000)^2$** $+(3-2)^2]$

- Distance between examples could be dominated by some attributes with relatively large numbers (e.g., income in our example).
- Important to normalize features (e.g., map numbers to numbers between 0-1)

  Example: Income

  If Highest income = 200K   Lowest income=0

  Davis’s income is normalized to 50/200, John’s income is normalized to 35/200, etc.)

8

k-NN with Normalization of Variables

Customer		Age	Income (K)	No. cards	Response
John		55/63= 0.55	35/200= 0.175	¾= 0.75	No
Rachel		22/63= 0.34	50/200= 0.25	2/4= 0.5	Yes
Hannah		63/63= 1	200/200 =1	¼= 0.25	No
Tom		59/63= 0.93	170/200 =0.85	¼= 0.25	No
Nellie		25/63= 0.39	40/200= 0.2	4/4= 1	Yes
David		37/63= 0.58	50/200= 0.25	2/4= 0.5	**Yes**

9

Another Problem with k-NN

- Distance works naturally with numerical attributes

D(Rachel, Johm)= sqrt $[(35-37)^2+(35-50)^2+(3-2)^2]=$**15.16**

What if we have nominal attributes?

Example: married

Customer	Married	Income (K)	No. cards	Response
John	Yes	35	3	No
Rachel	No	50	2	Yes
Hannah	No	200	1	No
Tom	Yes	170	1	No
Nellie	No	40	4	Yes
David	Yes	50	2	

10

K-NN with Nominal Attributes

- Method 1: Convert nominal attributes to numerical attributes
  - E.g., yes $\Rightarrow$ 1 and no $\Rightarrow$ 0
  - Blue $\Rightarrow$ 1, yellow $\Rightarrow$ 2, red $\Rightarrow$ 3, etc.
  - Problem?
- Method 2:

$$\text{Distance } (x, y) = \sum_{i=1}^{m} dist\,(x_i, y_i)$$

where

$$dist(x_i, y_i) = \begin{cases} 0 & \text{if } x_i \text{ and } y_i \text{ are nominal and } x_i = y_i \\ 1 & \text{if } x_i \text{ and } y_i \text{ are nominal and } x_i \neq y_i \\ |\,norm(x_i) - norm(y_i)\,| & \text{if } x_i \text{ and } y_i \text{ are continuous} \end{cases}$$

and m is the number of attributes

Example

- Distance between David and John:

D(David, John)= 0 + |0.25-0.175|+ |0.5-0.75|

Customer	Married	Income (K)	No. cards	Response
John	Yes	35/200=0.175	¾=0.75	No
Rachel	No	50/200=0.25	2/4=0.5	Yes
Hannah	No	200/200=1	¼=0.25	No
Tom	Yes	170/200=0.85	¼=0.25	No
Nellie	No	40/200=0.2	4/4=1	Yes
David	Yes	50/200=0.25	2/4=0.5	

K-Nearest Neighbor Classifier

Strengths and Weaknesses

Strengths:

- Simple to implement and use
- Comprehensible – easy to explain prediction
- Robust to noisy data by averaging k-nearest neighbors.
- Can learn complex target functions
- Can be used to do regression (how?)

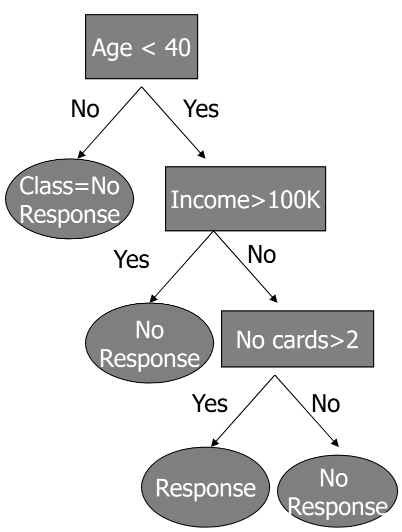
Weaknesses:

- Need a lot of space to store all examples.
- Takes more time to classify a new example than with a model (need to calculate and compare distance from new example to all other examples).

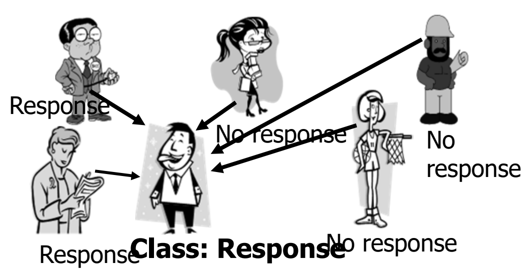
13

Decision Tree vs K-Nearest Neighbor Classifier

Classification Tree Model



K-Nearest Neighbors



14