SGG 4653
Advance Database System

Data Mining (Classification)

Contents

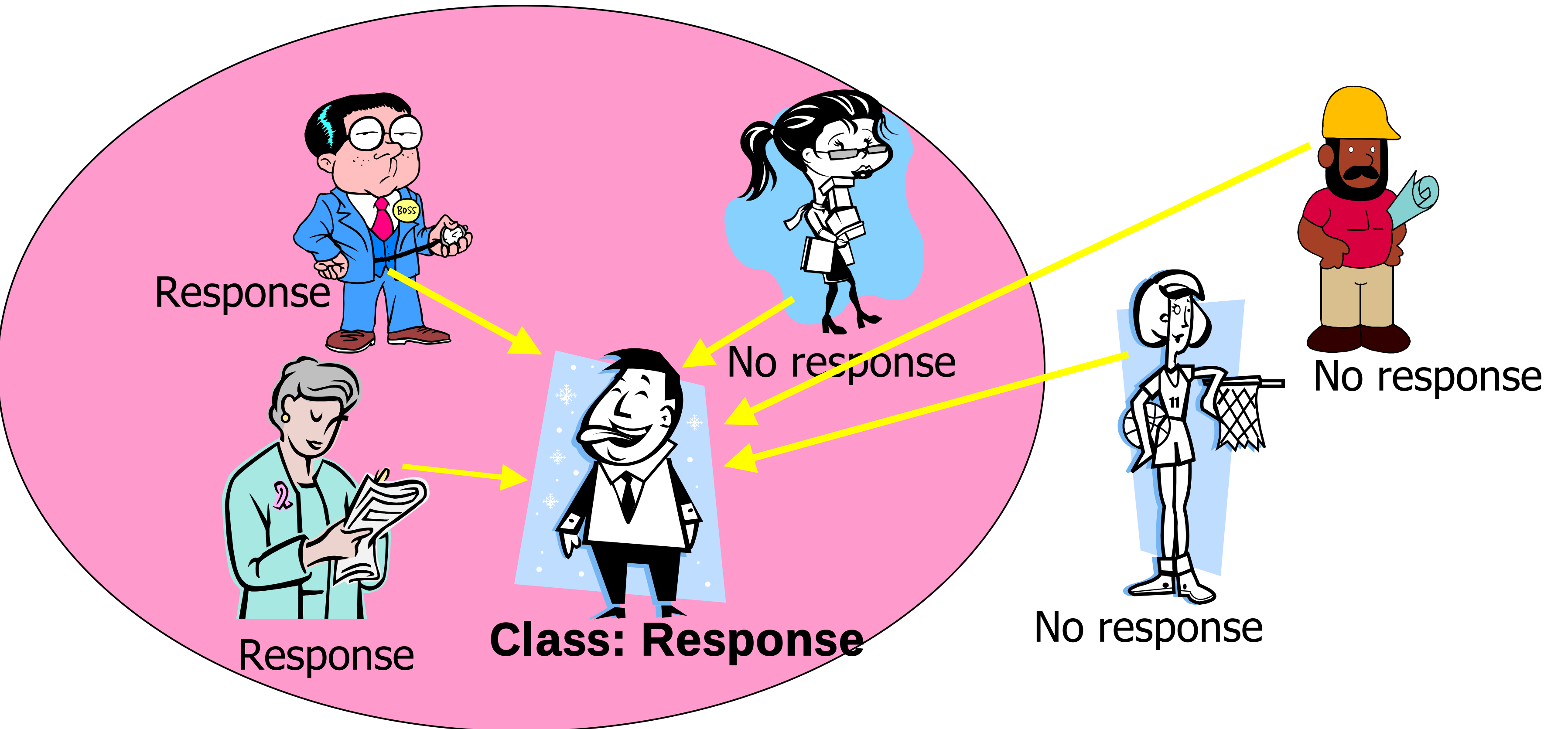
- Objective of this topic:
  - To extend understanding of classification technique
  - To understand classification technique based on k-Nearest Neighbour
- Contents of this topic:
  - K-Nearest Neighbor Classifiers

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.

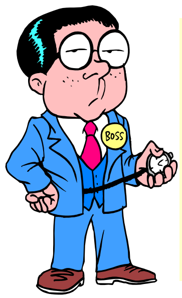
K-Nearest Neighbor Algorithm

- 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

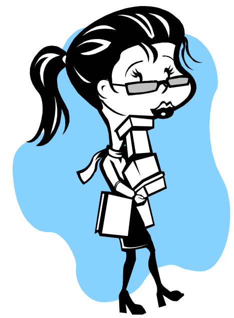


Distance Between Neighbors

- 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

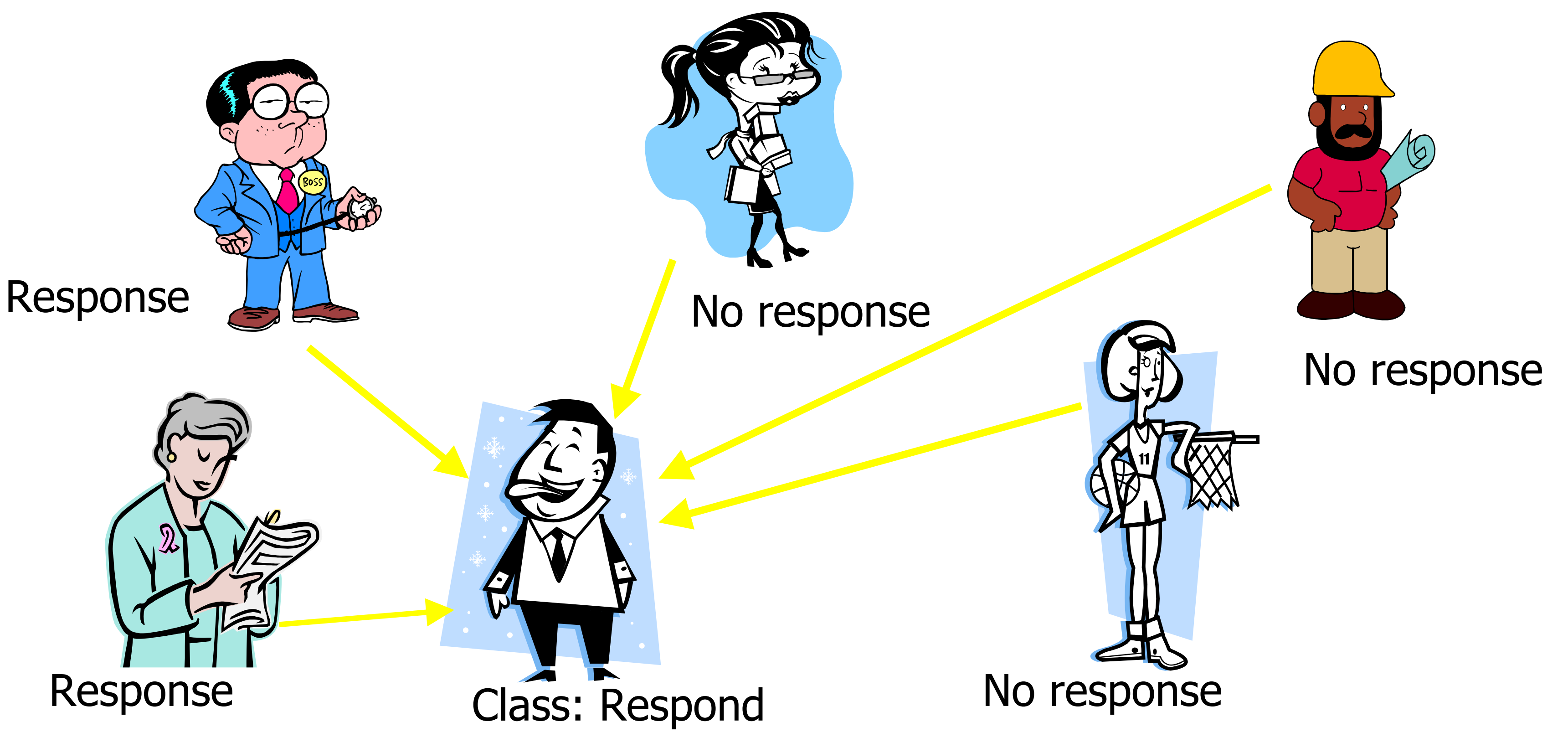
- "Closeness" is defined in terms of the *Euclidean* distance between two examples.
  - The Euclidean distance between $X=(x_1, x_2, x_3,\ldots x_n)$ and $Y=(y_1,y_2, y_3,\ldots y_n)$ is defined as:

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

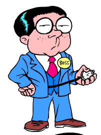
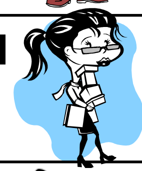
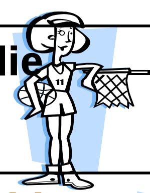
  - Distance (John, Rachel)=sqrt $[(35-41)^2+(95K-215K)^2+(3-2)^2]$

K-Nearest Neighbor: Instance Based Learning

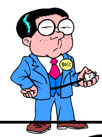
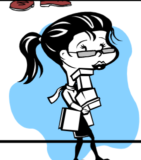
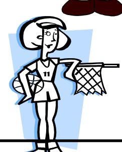
- No model is built: Store all training examples
- Any processing is delayed until a new instance must be classified.



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	?

Example: Distance from David

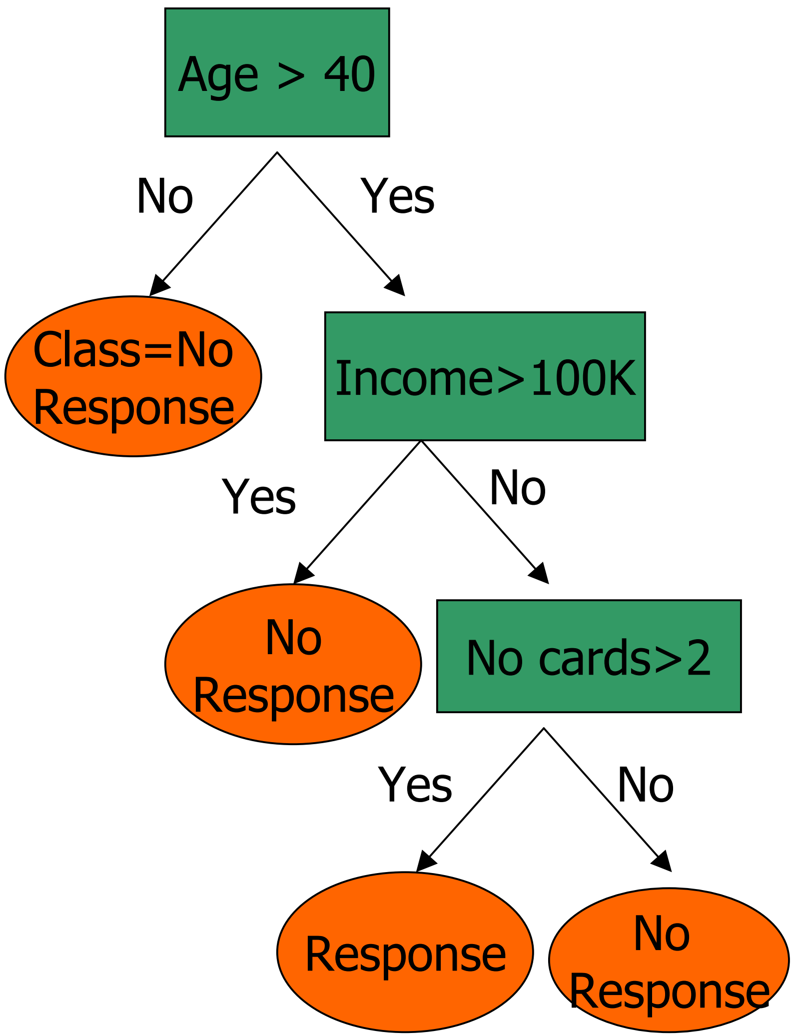
Customer	Age	Income (K)	No. cards	Response	Distance from David
John 	35	35	3	No	sqrt $[(35-37)^2+(35-50)^2+(3-2)^2]=15.16$
Rachel 	22	50	2	Yes	sqrt $[(22-37)^2+(50-50)^2+(2-2)^2]=15$
Hannah 	63	200	1	No	sqrt $[(63-37)^2+(200-50)^2+(1-2)^2]=152.23$
Tom 	59	170	1	No	sqrt $[(59-37)^2+(170-50)^2+(1-2)^2]=122$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

K-Nearest Neighbor Classifier

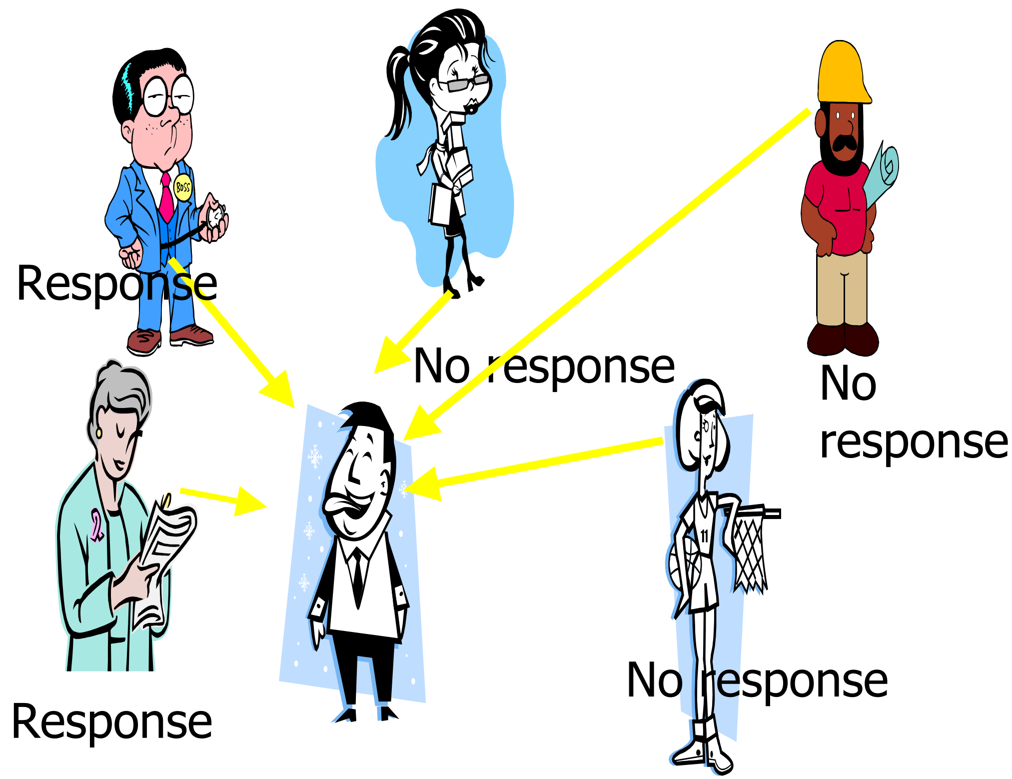
- Strengths
  - Simple to implement and use
  - Comprehensible – easy to explain prediction
  - Robust to noisy data by averaging k-nearest neighbors.
  - Some appealing applications
- Weaknesses
  - Need a lot of space to store all examples.
  - Takes more time to classify a new example than with a model (need to calculate and compare distance from new example to all other examples).

K-Nearest Neighbor Classifier

Classification Tree Modes

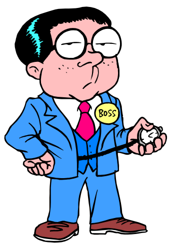


K-Nearest Neighbors



Class: Response

Strengths and Weaknesses K-Nearest Neighbor Classifier



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

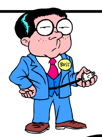
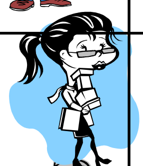
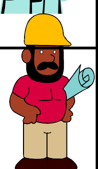
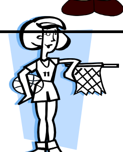
  Example: Income

  Highest income = 500K

  Davis's income is normalized to 95/500, Rachel income is normalized to 215/500, etc.)

Strengths and Weaknesses K-Nearest Neighbor Classifier

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

Strengths and Weaknesses K-Nearest Neighbor Classifier

- **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	

Issues regarding classification (1): Data Preparation

- **Data cleaning**
  - **Preprocess data in order to reduce noise and handle missing values**
- **Relevance analysis (feature selection)**
  - **Remove the irrelevant or redundant attributes**
- **Data transformation**
  - **Generalize and/or normalize data**

Issues regarding classification (2): Evaluating Classification Methods

- Predictive accuracy
- Speed and scalability
  - time to construct the model
  - time to use the model
- Robustness
  - handling noise and missing values
- Scalability
  - efficiency in disk-resident databases
- Interpretability:
  - understanding and insight provided by the model
- Goodness of rules
  - decision tree size
  - compactness of classification rules

Conclusions

- K-Nearest Neighbor Classifier is Learning by analogy
- K-Nearest Neighbor algorithm:
  - 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
- **Strengths**
  - **Simple to implement and use**
  - **Comprehensible – easy to explain prediction**
  - **Robust to noisy data by averaging k-nearest neighbors.**
  - **Some appealing applications**
- **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).**