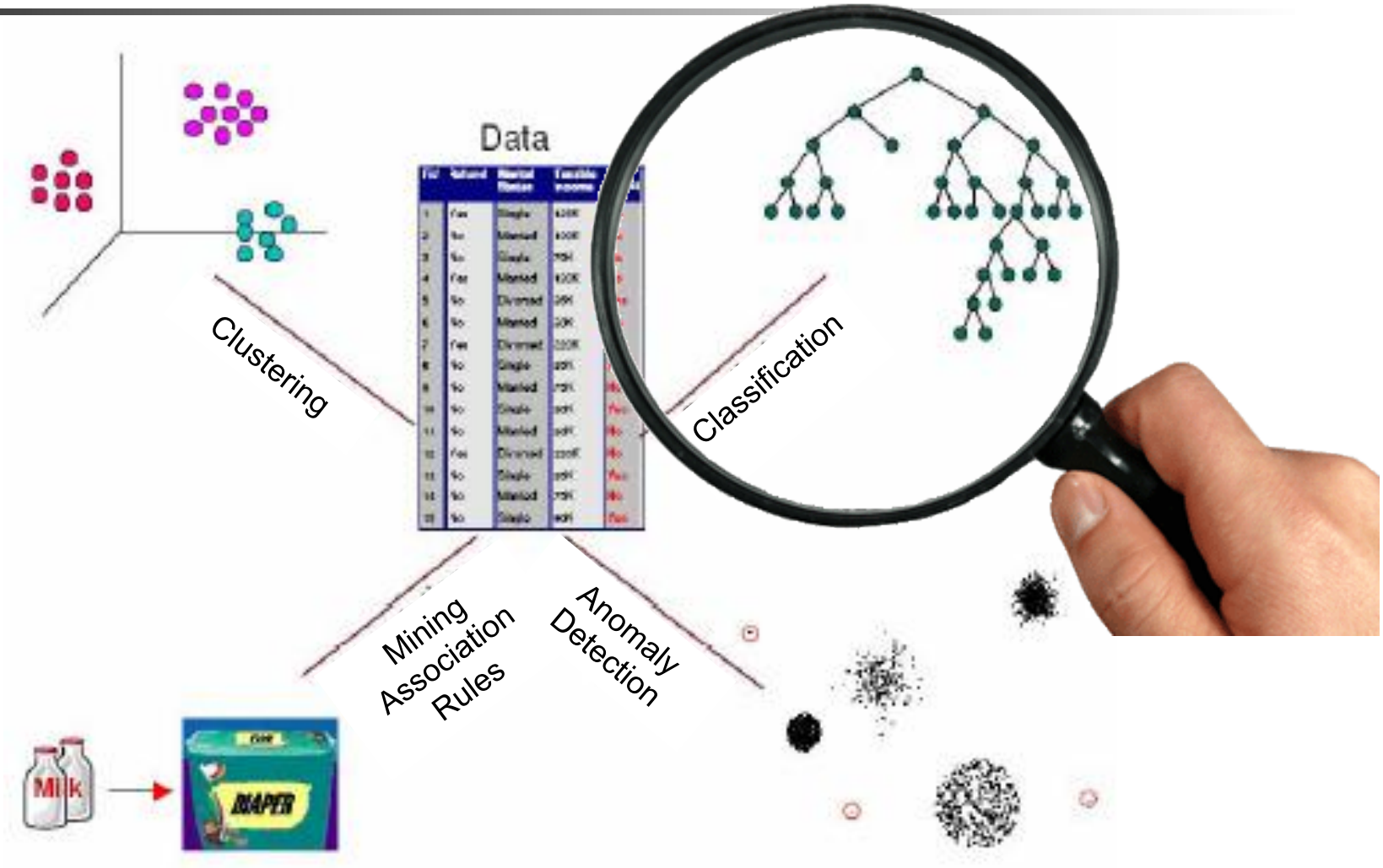
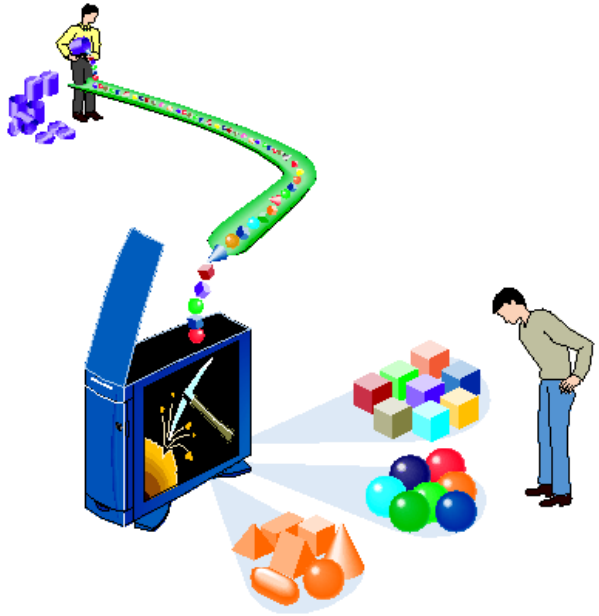
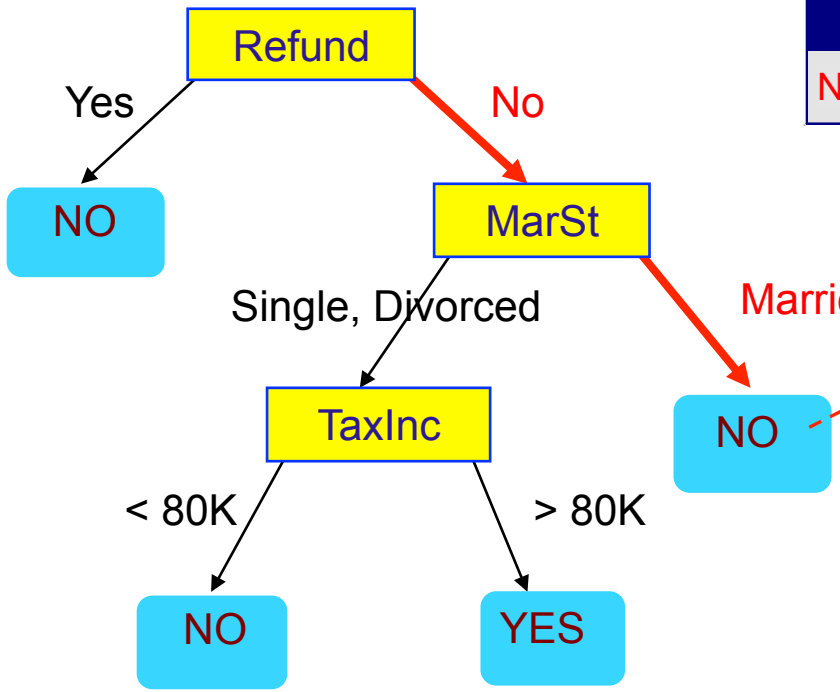
Data Mining Tasks





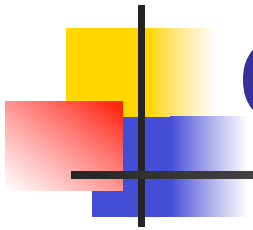
Classification Algorithms

Apply Model to New Data



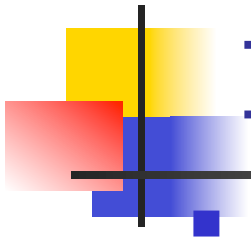
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Assign Cheat to “No”



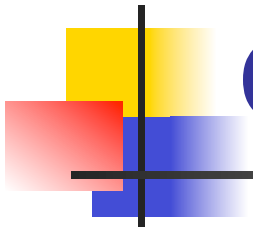
Classification Algorithms

- Nearest Neighbor
- Naïve Bayes
- Decision Tree
- …



Issues to consider..

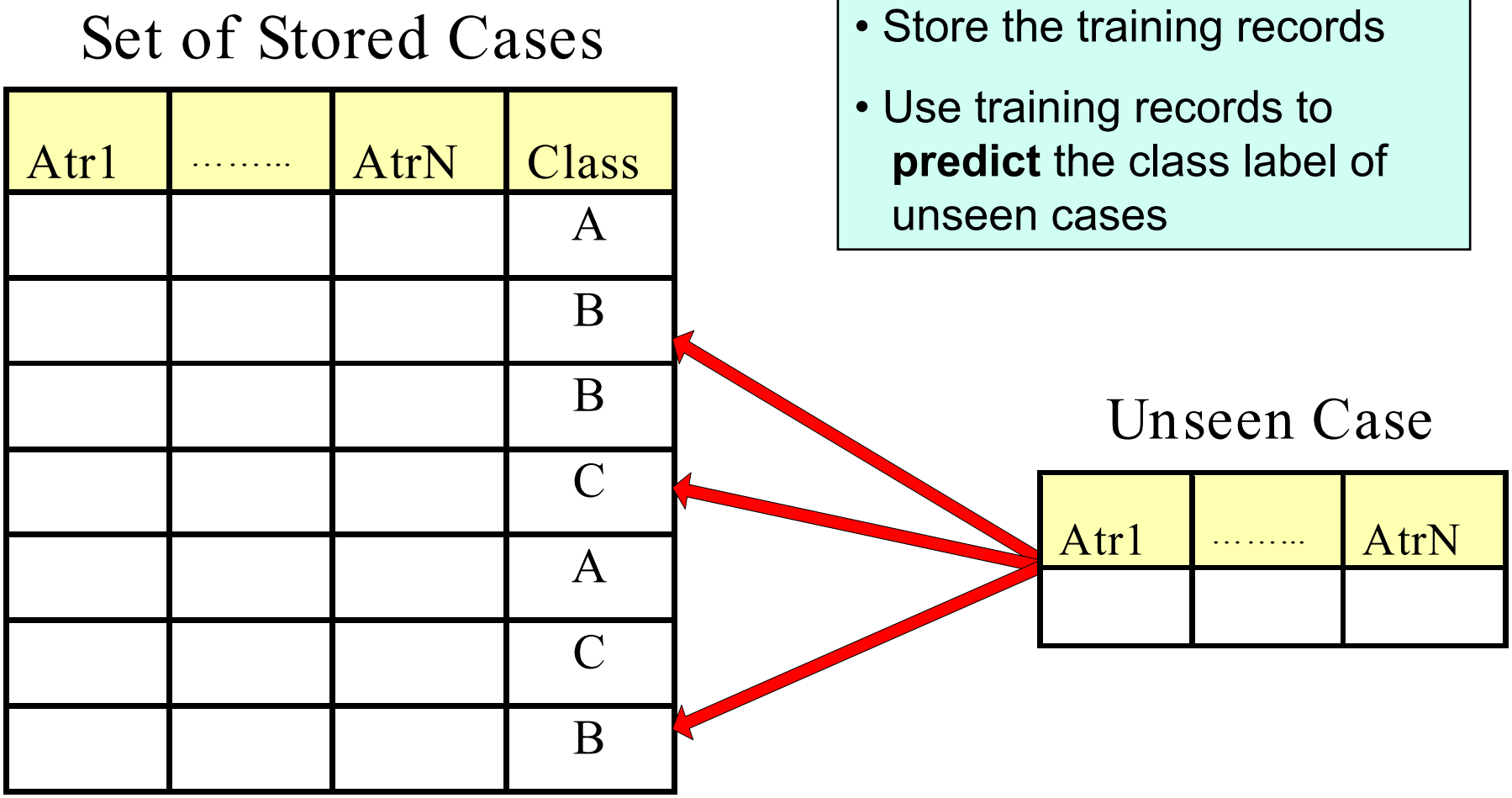
- Accuracy
  - classifier accuracy: predicting class label
- Speed
  - time to construct the model (training time)
  - time to use the model (classification/prediction time)
- Robustness:
  - handling noise and missing values
- Scalability:
  - Handling large amounts of data
- Interpretability
  - understanding the insight provided by the model

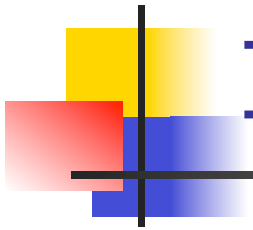


Classification Algorithms

- **Nearest Neighbor**
- Naïve Bayes
- Decision Tree
- …

Instance-Based Classifiers



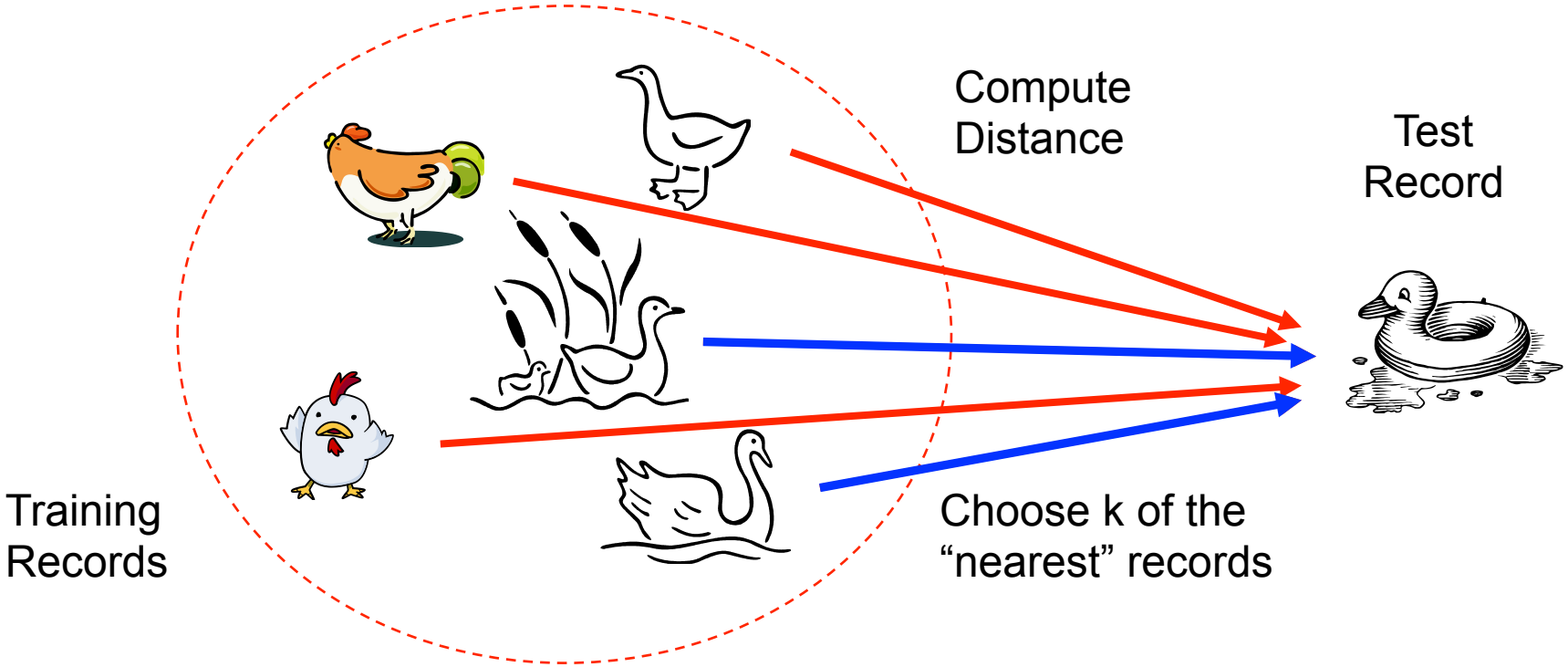


Instance-Based Classifiers

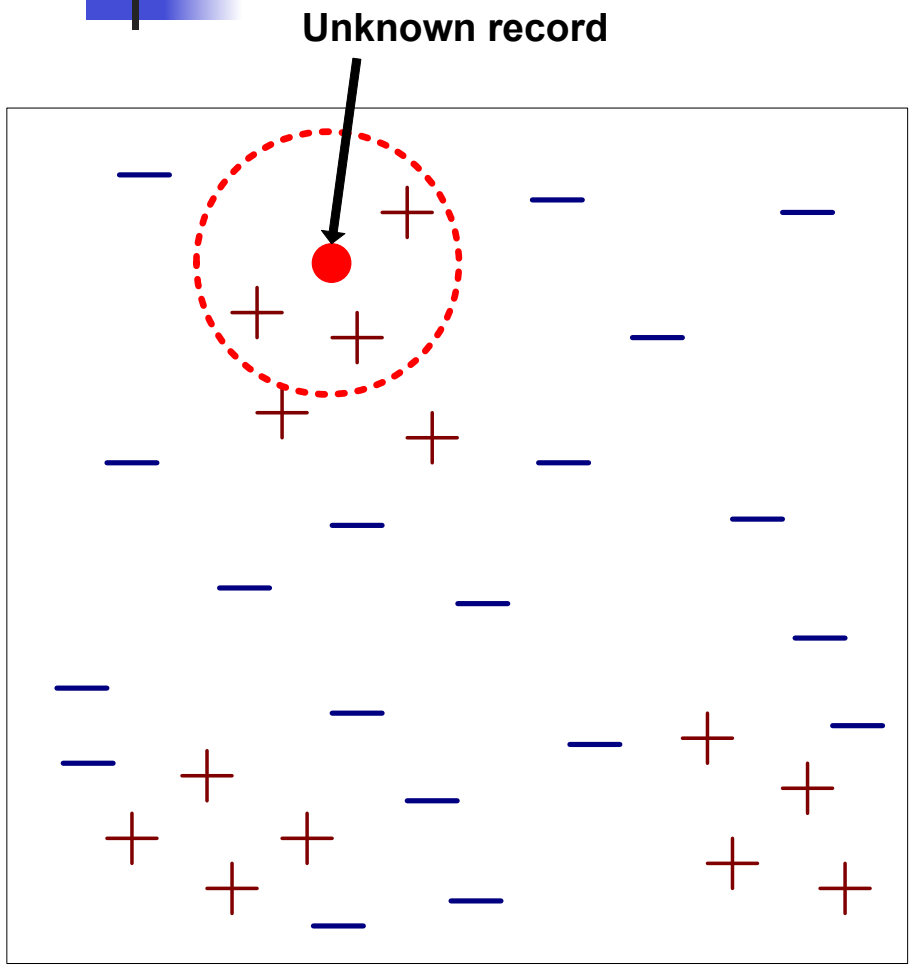
- Examples:
  - Rote-learner
    - Memorizes entire training data and performs classification only if attributes of record **match** one of the training examples exactly
  - **Nearest neighbor**
    - Uses k "closest" points (k nearest neighbors) for performing classification

Nearest Neighbor Classifiers

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's **probably** a duck ☺

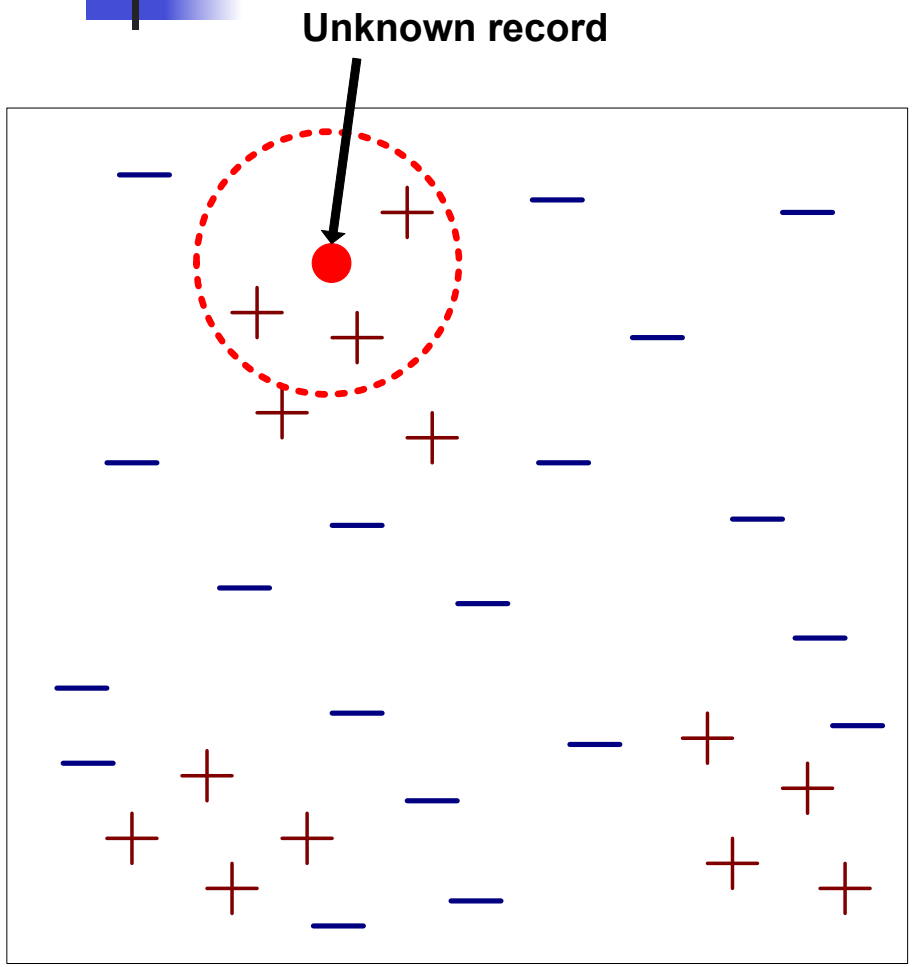


Nearest-Neighbor Classifiers



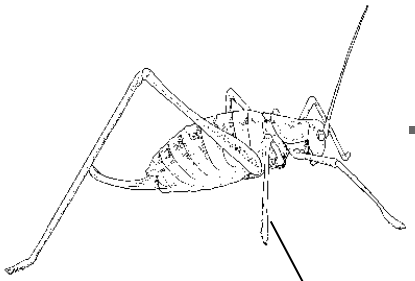
- Requires <u>three things</u>
  1. The set of stored records
  2. **Distance Metric** to compute distance between records
  3. **The value of *k***, the number of nearest neighbors to retrieve

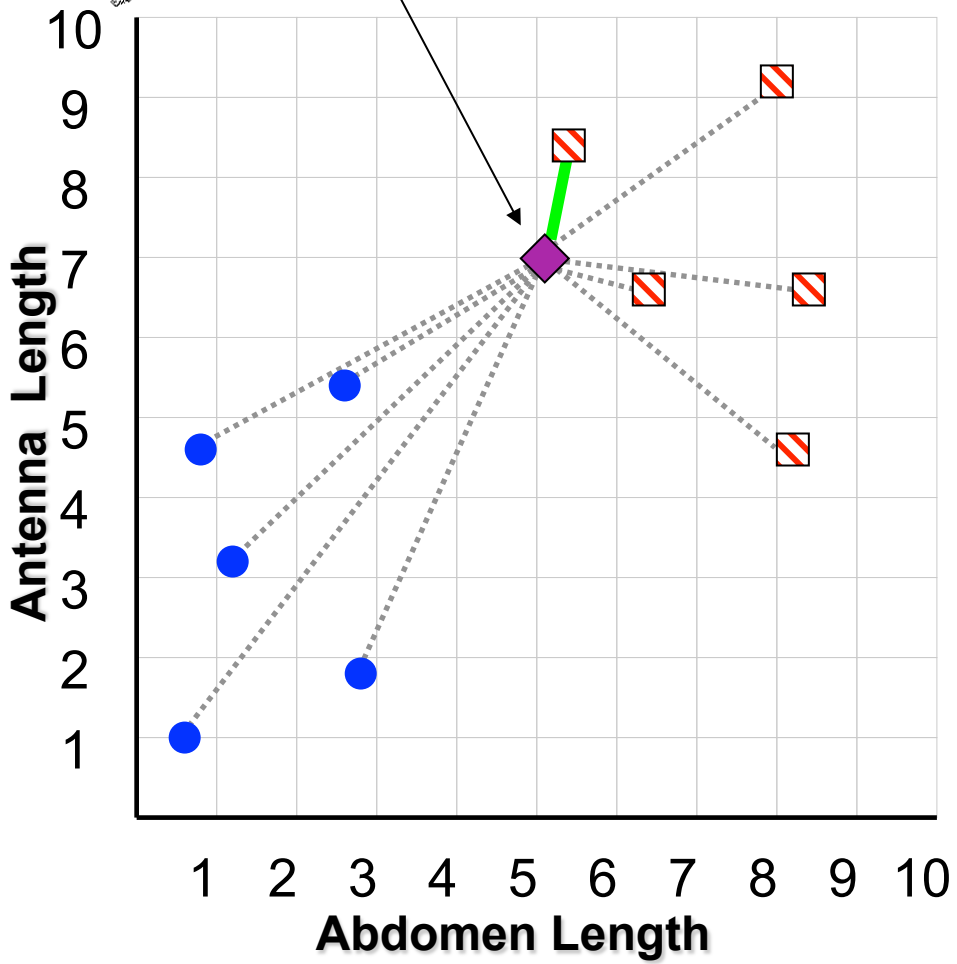
Nearest-Neighbor Classifiers



- To classify an unknown record:
  1. **Compute distance** to other training records
  2. **Identify *k*** nearest neighbors
  3. **Use** class labels of nearest neighbors to **determine** the class label of unknown record (e.g., by taking majority vote)

Nearest Neighbor Classifier





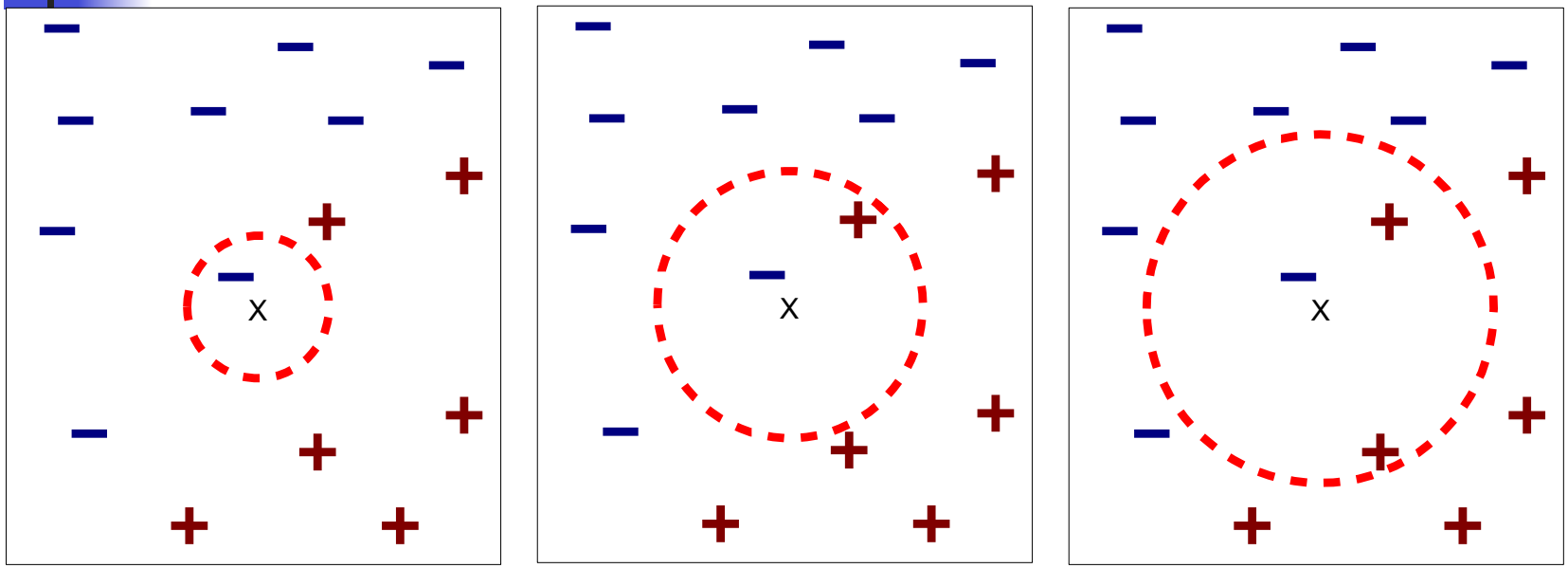
2-dimensional space

D1: Antenna Length
D2: Abdomen Length

If the **nearest** instance **(k=1)** to the previously unseen instance **is a Cricket**
class is **Cricket**
else
class is **Grasshopper**

Crickets
Grasshoppers

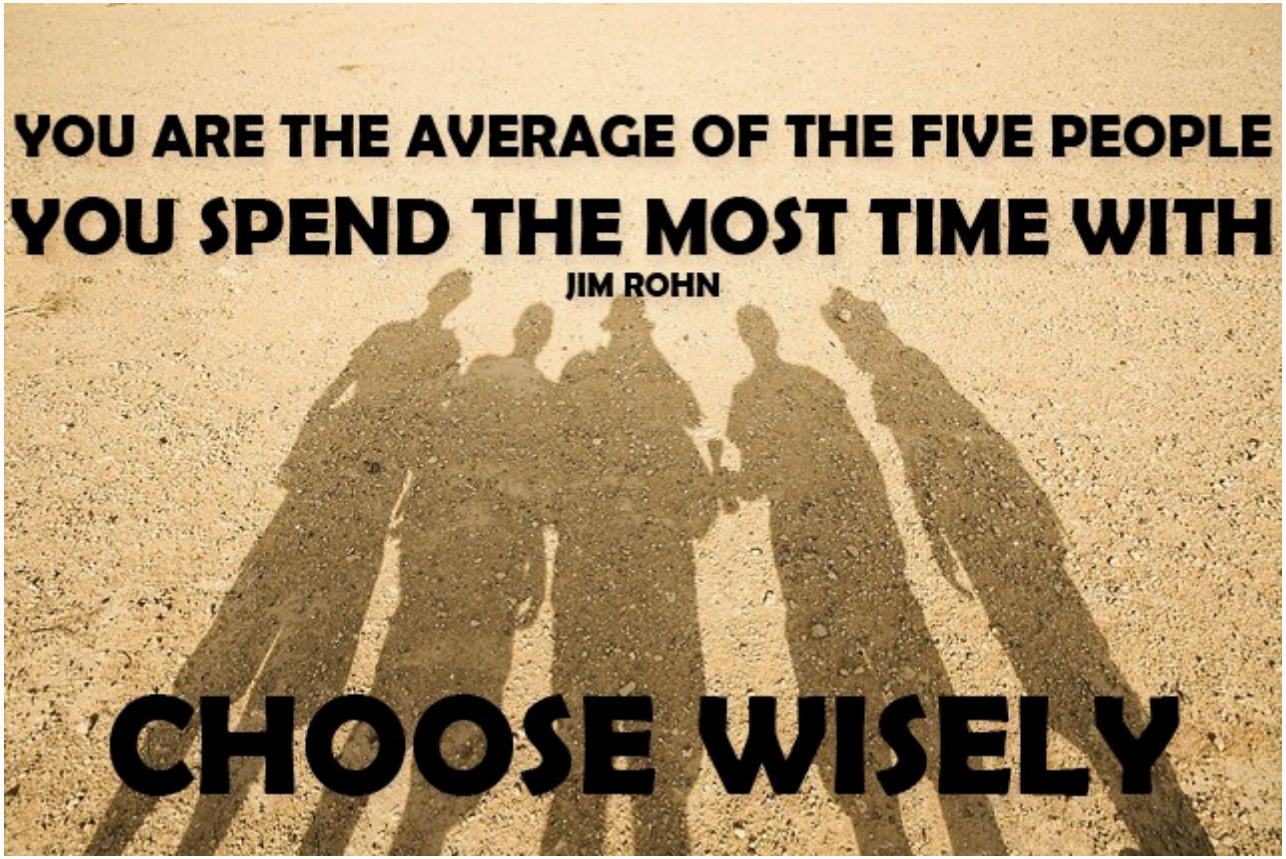
Definition of Nearest Neighbor



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

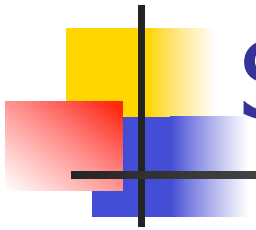
K-nearest neighbors of a record x are data points that have the k smallest **distance** to x

KNN in a quote



Similarity and Distance

- One of the fundamental concepts of data mining is the notion of **similarity** between data points
  - often formulated as a numeric distance between data points.
- Similarity is the basis for many data mining procedures.
  - Later we will see it again (for clustering)

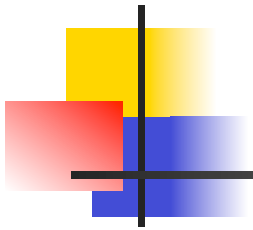


Similarity and Dissimilarity

- **Similarity**
  - Numerical measure of how alike two data objects are
  - Is higher when objects are more alike
  - Often falls in the range [0,1]

- **Dissimilarity**
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0

- Distance refers to a similarity or dissimilarity



Euclidean Distance

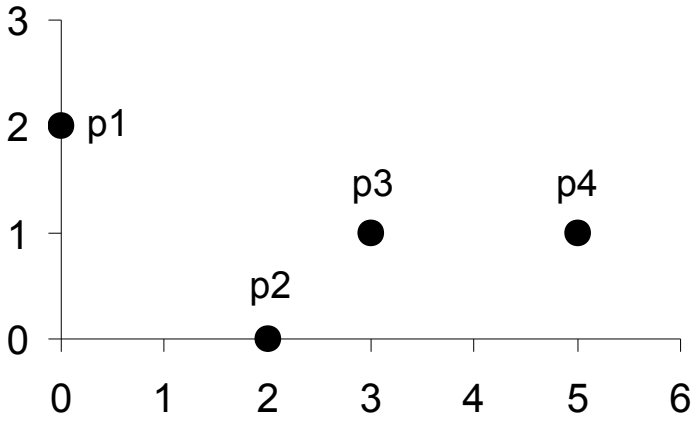
- **Euclidean Distance**

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where:

- *p* and *q* are data objects
- *n* is the number of dimensions (attributes)
- p_k and q_k are, respectively, the k^{th} attributes (dimensions) of data objects *p* and *q*

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Distance Between Data Items

- Each data item is represented with a set of attributes

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

Rachel:
Age = 22
Income = 50K
No. of credit cards = 2

- "Closeness" is defined in terms of the distance (Euclidean or some other distance) between two data items.
  - The Euclidean distance between John and Rachel

Distance(John, Rachel) = sqrt[$(35-22)^2+(35K-50K)^2+(3-2)^2$]

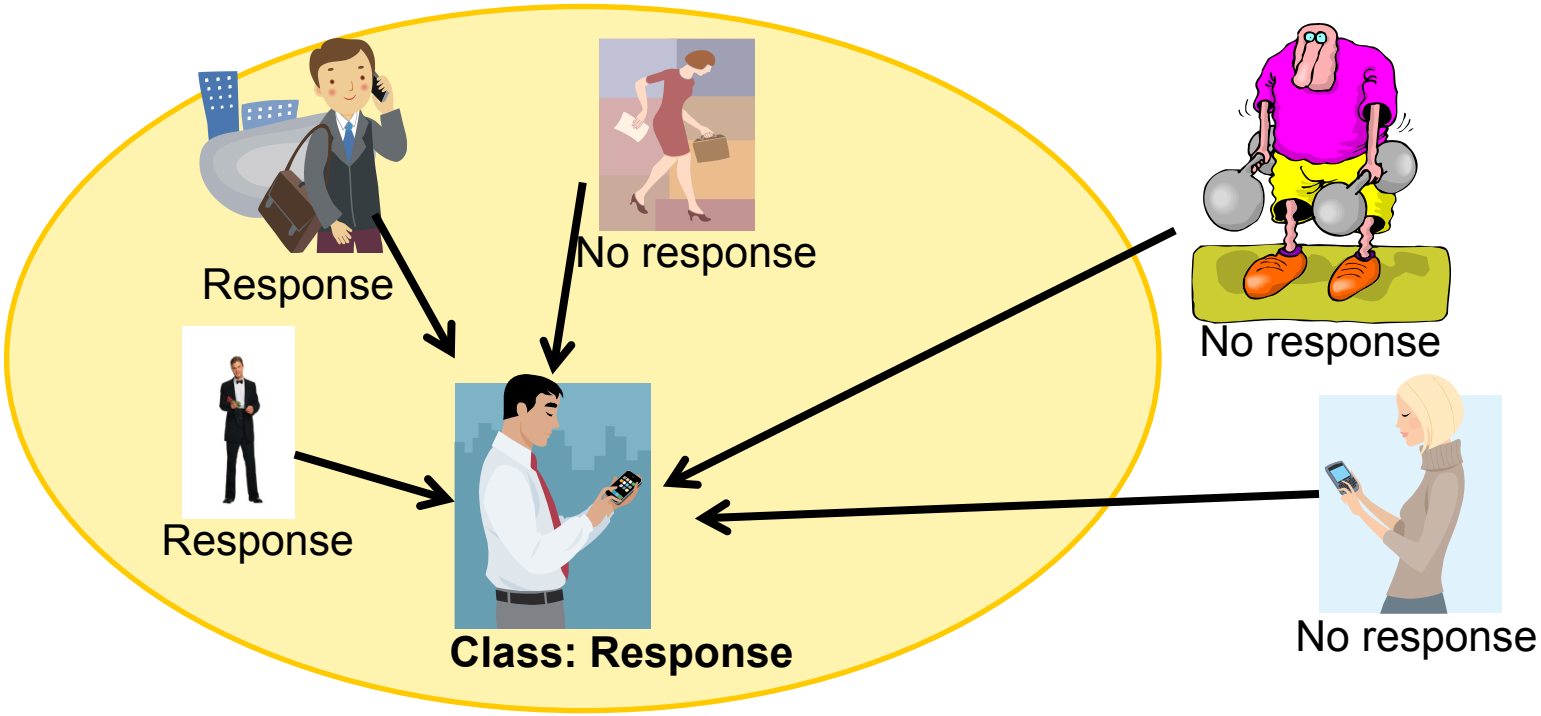
Nearest Neighbor Classification

- Compute distance between two points:
  - Euclidean distance

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

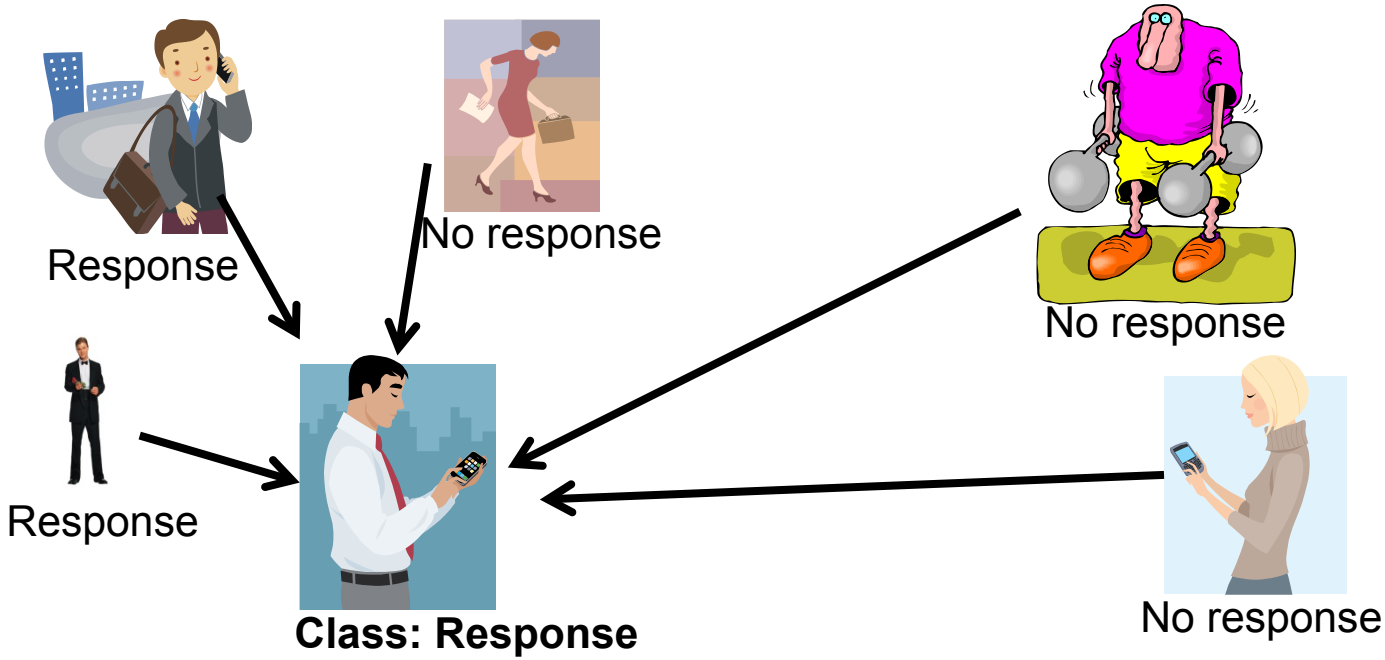
- Determine the class from nearest neighbors
  - take **the majority vote** of class labels among the k-nearest neighbors, or
  - **Weigh** the vote according to distance
    - weight factor, $w = 1/d^2$

Nearest Neighbors for Classification



k-Nearest Neighbor Algorithms

- No model is built: Store all training examples
  - Memory-based learning
- Any processing is delayed until a new instance must be classified
  → **lazy** classification technique



k-Nearest Neighbor Classifier

Example (*k*=3)

Customer	Age	Income (K)	No. of cards	Response
John 	35	35	3	Yes
Rachel 	22	50	2	No
Ruth 	63	200	1	No
Tom 	59	170	1	No
Neil 	25	40	4	Yes
David 	**37**	**50**	**2**	?

Distance from David

k-Nearest Neighbor Classifier

Example (*k*=3)

Customer	Age	Income (K)	No. of cards	Response	Distance from David
John 	35	35	3	Yes	sqrt $[(35-37)^2+(35-50)^2+(3-2)^2]$=**15.16**
Rachel 	22	50	2	No	sqrt $[(22-37)^2+(50-50)^2+(2-2)^2]$=**15**
Ruth 	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**
Neil 	25	40	4	Yes	sqrt $[(25-37)^2+(40-50)^2+(4-2)^2]$=**15.74**
David 	**37**	**50**	**2**	?	

k-Nearest Neighbor Classifier

Example (*k*=3)

Customer	Age	Income (K)	No. of cards	Response	Distance from David
John 	35	35	3	Yes	sqrt $[(35-37)^2+(35-50)^2+(3-2)^2]$=**15.16**
Rachel 	22	50	2	No	sqrt $[(22-37)^2+(50-50)^2+(2-2)^2]$=**15**
Ruth 	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**
Neil 	25	40	4	Yes	sqrt $[(25-37)^2+(40-50)^2+(4-2)^2]$=**15.74**
David 	**37**	**50**	**2**	?	

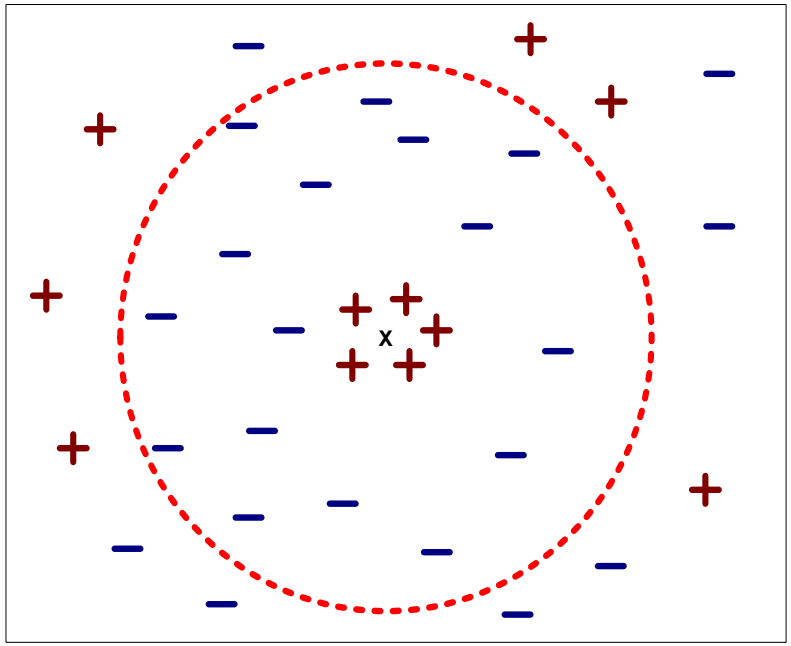
k-Nearest Neighbor Classifier

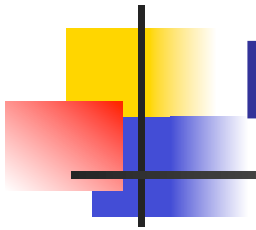
Example (*k*=3)

Customer	Age	Income (K)	No. of cards	Response	Distance from David
John 	35	35	3	Yes	sqrt $[(35-37)^2+(35-50)^2+(3-2)^2]$=**15.16**
Rachel 	22	50	2	No	sqrt $[(22-37)^2+(50-50)^2+(2-2)^2]$=**15**
Ruth 	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**
Neil 	25	40	4	Yes	sqrt $[(25-37)^2+(40-50)^2+(4-2)^2]$=**15.74**
David 	**37**	**50**	**2**	**Yes**	

Nearest Neighbor Classification…

- Choosing the value of k:
  - If k is **too small**
    - sensitive to noise points
  - If k is **too large**
    - neighborhood may include points from other classes

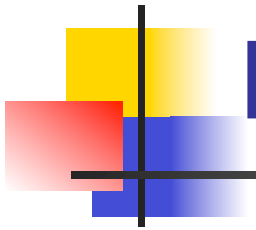




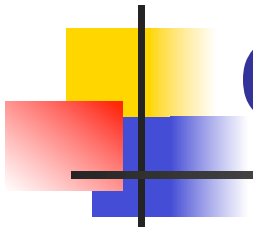
Nearest Neighbor Classification…

- Scaling issues
  - Attributes may have to be **scaled:**
    - to prevent distance measures from being **dominated** by one of the attributes
  - Example:
    - height of a person may vary from 1.5m to 1.8m
    - weight of a person may vary from 90lb to 300lb
    - Income
      - Highest income = 500K
      - John's income is normalized to 35/500, Rachel's income is normalized to 50/500, etc.)

Nearest neighbor Classification…

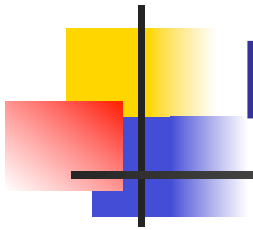


- k-NN classifiers are **lazy** learners
  - It does not build models explicitly
    - Unlike **eager** learners such as decision tree induction
  - Classifying unknown records is relatively expensive
- **Robust** to noisy data:
  - by averaging *k*-nearest neighbors



Classification Algorithms

- Nearest Neighbor
- **Naïve Bayes**
- Decision Tree
- …



Bayes Classifier

- A probabilistic model for the classification problem
- What is the probability that:
  - An applicant will default on loan?
  - An email is a spam?
  - …
- Conditional Probability:
  - Example: **P(spam | keywords)**

Bayes Classifier

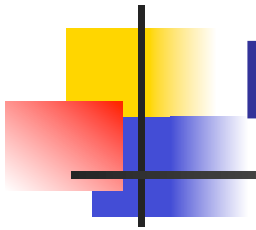
- Joint Probability: P(A,C)

$$P(A,C) = P(A|C) \times P(C) = P(C|A) \times P(A)$$

Example: *P(burger, coke)=P(coke|burger) × P(burger)=..*

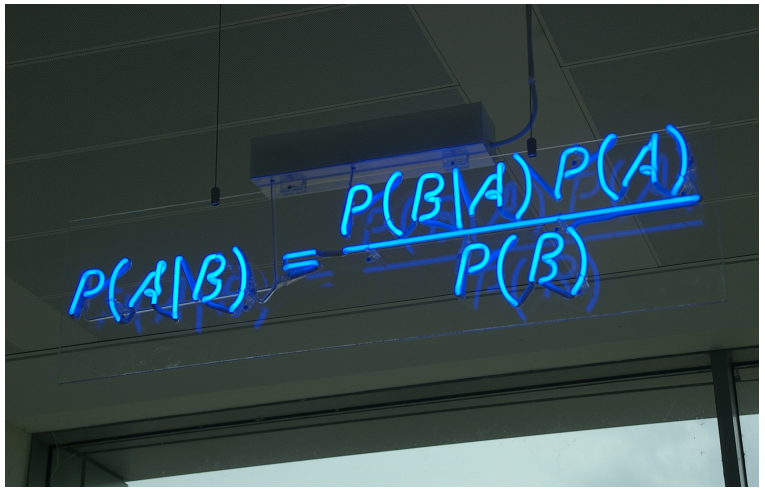
- **Bayes theorem:**

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$



Bayes Classifier

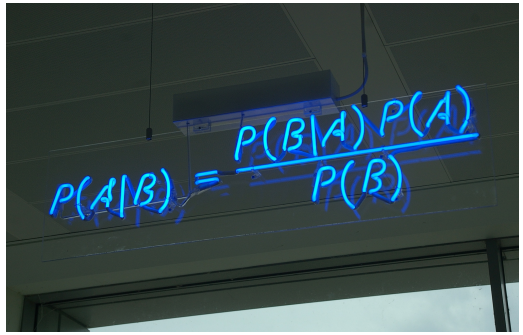
- Bayes' Theorem is a way of flipping around conditional probabilities
- To find P(spam|keywords):
  - use P(keywords|spam)!



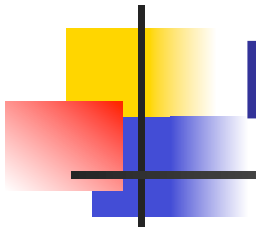
Example of Bayes Theorem

- Given:
  - A doctor knows that migraines cause stiff neck 50% of the time
  - Prior probability of any patient having migraines is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20

- If a patient has stiff neck, what's the probability she has migraines?



$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$



Bayesian Classifiers

- Consider each attribute and class label as random variables
- Given a record with attributes (A_1, A_2,…,A_n)
  - **Goal:** is to predict class C
  - Specifically, we want to find the value of C that **maximizes** P(C| A_1, A_2,…,A_n)
- Can we estimate P(C| A_1, A_2,…,A_n) from data?

Bayesian Classifiers

- Approach:
  - **compute** the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}$$

  - **Choose** value of C that maximizes
    $P(C \mid A_1, A_2, \ldots, A_n)$
  - Equivalent to choosing value of C that maximizes
    $P(A_1, A_2, \ldots, A_n \mid C)\, P(C)$

- How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Bayesian Classifiers - Example

- Approach:
  - **compute** the posterior probability P(C | A_1, A_2, …, A_n) for all values of C using the Bayes theorem

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}$$

  - Example

$$P(spam \mid transfer, money) = \frac{P(transfer, money \mid spam) P(spam)}{P(transfer, money)}$$

$$P(no\_spam \mid transfer, money) = \frac{P(transfer, money \mid no\_spam) P(no\_spam)}{P(transfer, money)}$$

  - How to estimate P(A_1, A_2, …, A_n | C)?

Naïve Bayes Classifier

- Assume **independence** among attributes A_i
  - $P(A_1, A_2, \ldots, A_n \mid C) = P(A_1 \mid C_j)\, P(A_2 \mid C_j) \ldots P(A_n \mid C_j)$
  - Can estimate $P(A_i \mid C_j)$ for all A_i and C_j ☺

- New record is classified to C_j :
  - if Π **$P(A_i \mid C_j) \times P(C_j)$** is maximal

Estimating Probabilities from Data

$$\Pi\, P(A_i \mid C_j) \times P(C_j)$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- **Classes:**
- $P(C) = N_c/N$
- Examples
  - P(No) = 7/10
  - P(Yes) = 3/10

Estimating Probabilities from Data

$$\Pi\, P(A_i|\, C_j) \times P(C_j)$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- **Attributes:**
- $P(A_i \mid C_j) = |A_{ij}|/ N_c$
  - where $|A_{ij}|$ is number of instances having attribute A_i that belongs to class C_j
  - Examples:
    - P(Status=Married|No) = 4/7
    - P(Refund=Yes|Yes) =0

Example

Classes:

C1:buys_computer = ‘yes’

C2:buys_computer = ‘no’

Record X = (age <=30,

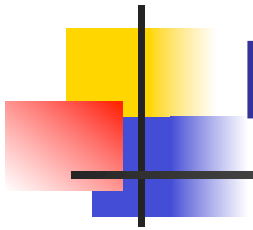
Income = medium,

Student = yes

Credit_rating = Fair)

Classify X

Attribute1	Attribute2	Attribute3	Attribute4	Class label
age	**income**	**student**	**credit_rating**	**buys_computer**
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



Example

X = (age <= 30 , income = medium, student = yes, credit_rating = fair)

- **compute and compare:**
  - P(buys_computer = "yes"|**X**)
  - P(buys_computer = "no"|**X**)

- **Equivalent to:**
  - P(**X**|buys_computer = "yes") × P(buys_computer = "yes")
  - P(**X**|buys_computer = "no") × P(buys_computer = "no")

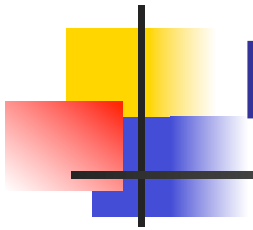
Example

P(**X**|buys_computer = “yes”) × **P(buys_computer = “yes”)**
P(**X**|buys_computer = “no”) × **P(buys_computer = “no”)**

- P(buys_computer = “yes”) = 9/14 = 0.643

- P(buys_computer = “no”) = 5/14= 0.357

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



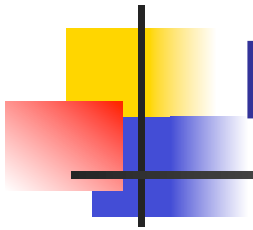
Example

P(X|buys_computer = "yes") × P(buys_computer = "yes")
P(X|buys_computer = "no") × P(buys_computer = "no")

X = (age <= 30 , income = medium, student = yes, credit_rating = fair)

- **P(X|buys_computer = "yes")**
  - P(age = "<=30" | buys_computer = "yes") = 2/9 = 0.222
  - P(income = "medium" | buys_computer = "yes") = 4/9 = 0.444
  - P(student = "yes" | buys_computer = "yes) = 6/9 = 0.667
  - P(credit_rating = "fair" | buys_computer = "yes") = 6/9 = 0.667
  - **P(X|buys_computer = "yes") = 0.222 x 0.444 x 0.667 x 0.667 = 0.044**

- **P(X|buys_computer = "no")**
  - P(age = "<= 30" | buys_computer = "no") = 3/5 = 0.6
  - P(income = "medium" | buys_computer = "no") = 2/5 = 0.4
  - P(student = "yes" | buys_computer = "no") = 1/5 = 0.2
  - P(credit_rating = "fair" | buys_computer = "no") = 2/5 = 0.4
  - **P(X|buys_computer = "no") = 0.6 x 0.4 x 0.2 x 0.4 = 0.019**



Example

P(X|buys_computer = “yes”) × P(buys_computer = “yes”) = **0.028**
P(X|buys_computer = “no”) × P(buys_computer = “no”)= **0.007**

Therefore, **X** belongs to class **(“buys_computer = yes”)**

Avoiding the Zero-Probability Problem

- Naïve Bayesian prediction requires each conditional prob. be **non-zero**. Otherwise, the predicted prob. will be zero

$$\Pi\ P(A_i|\ C_j) \times P(C_j)$$

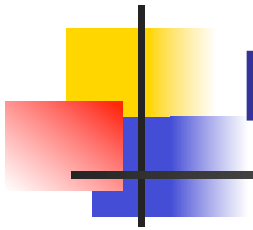
- Ex. Suppose a dataset with 1000 tuples, income=low (0), income= medium (990), and income = high (10)

- Use **Laplacian correction** (or Laplacian estimator)
  - *Adding 1 to each case*

    Prob(income = low) = 1/1003

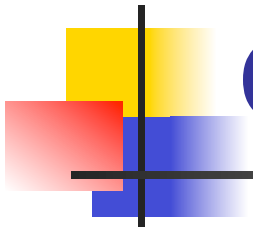
    Prob(income = medium) = 991/1003

    Prob(income = high) = 11/1003
  - The "corrected" prob. estimates are close to their "uncorrected" counterparts



Naïve Bayesian Classifier: Comments

- **Advantages**
  - Easy to implement
  - Good results obtained in most of the cases
- **Disadvantages**
  - Independence assumption may not hold for some attributes
  - Practically, dependencies exist among variables
  - Use other techniques…
  -



Classification Algorithms

- Nearest Neighbor
- Naïve Bayes
- **Decision Tree**
- …