Introduction to Classification, Nearest Neighbors & Naïve Bayes

Aliridho Barakbah, Entin Martiana, Yuliana, Renovita

Knowledge Engineering Laboratory
Department of Information and Computer Engineering
Politeknik Elektronika Negeri Surabaya

Classification: Definition

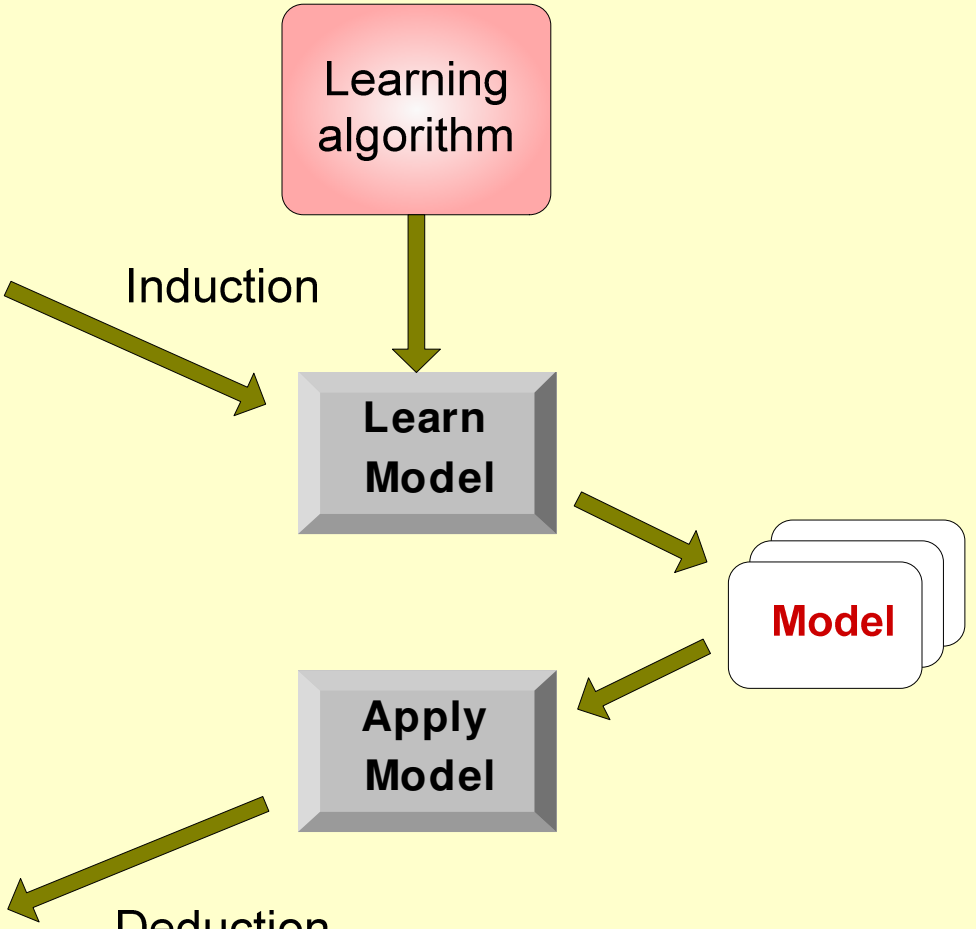
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Tan, Steinbach, Kumar, *Introduction to Data Mining*

Illustrating Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

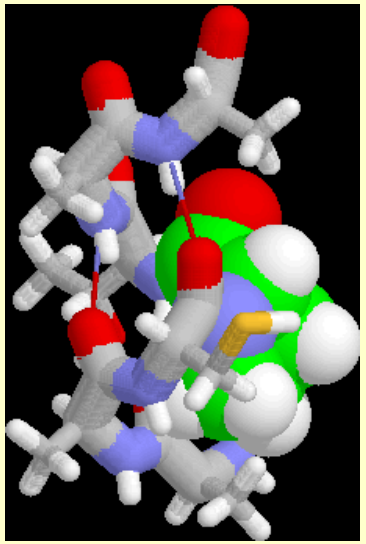
Test Set

Examples of Classification Task

- Predicting tumor cells as benign or malignant

- Classifying credit card transactions as legitimate or fraudulent

- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil



- Categorizing news stories as finance, weather, entertainment, sports, etc

Steps of Classification

- The classification process is divided into three steps: learning and testing.
  - Learning → Some data those have class label is induced to build a model of generalization
  - Validation → the model is examined with some of other data those have hidden class label, in order to determine accuracy of the model
  - Testing → the model is examined with new data those have no class label
- However, in the context classification performance, the validation step is commonly mentioned as testing
- If the accuracy is sufficient, this model can be used to predict unknown data classes.
- Classification is characterized by training data that has a label, based on this label the classification process obtains a pattern of attributes from a data.

Classification using Nearest Neighbors (NN)

- A simple method to classify a new data based on similarity with labeled data
- Similarity usually uses the distance metric
- The unit of distance generally uses the Euclidian
- Has several names: lazy algorithm, memory-based, instance-based, exemplar-based, case-based, experience-based

Types of NN

- 1-NN
  - Classification is based on 1 nearest training data
- *k*-NN
  - Classification is based on several (*k*) nearest training data. The classification result is determined with voting from class labels of *k* nearest training data
  - $k > 1$

1-NN Algorithm

- Calculate the distance between a new data to training data
- Determine 1 nearest training data
- Classify a new data into the label of the 1 nearest training data

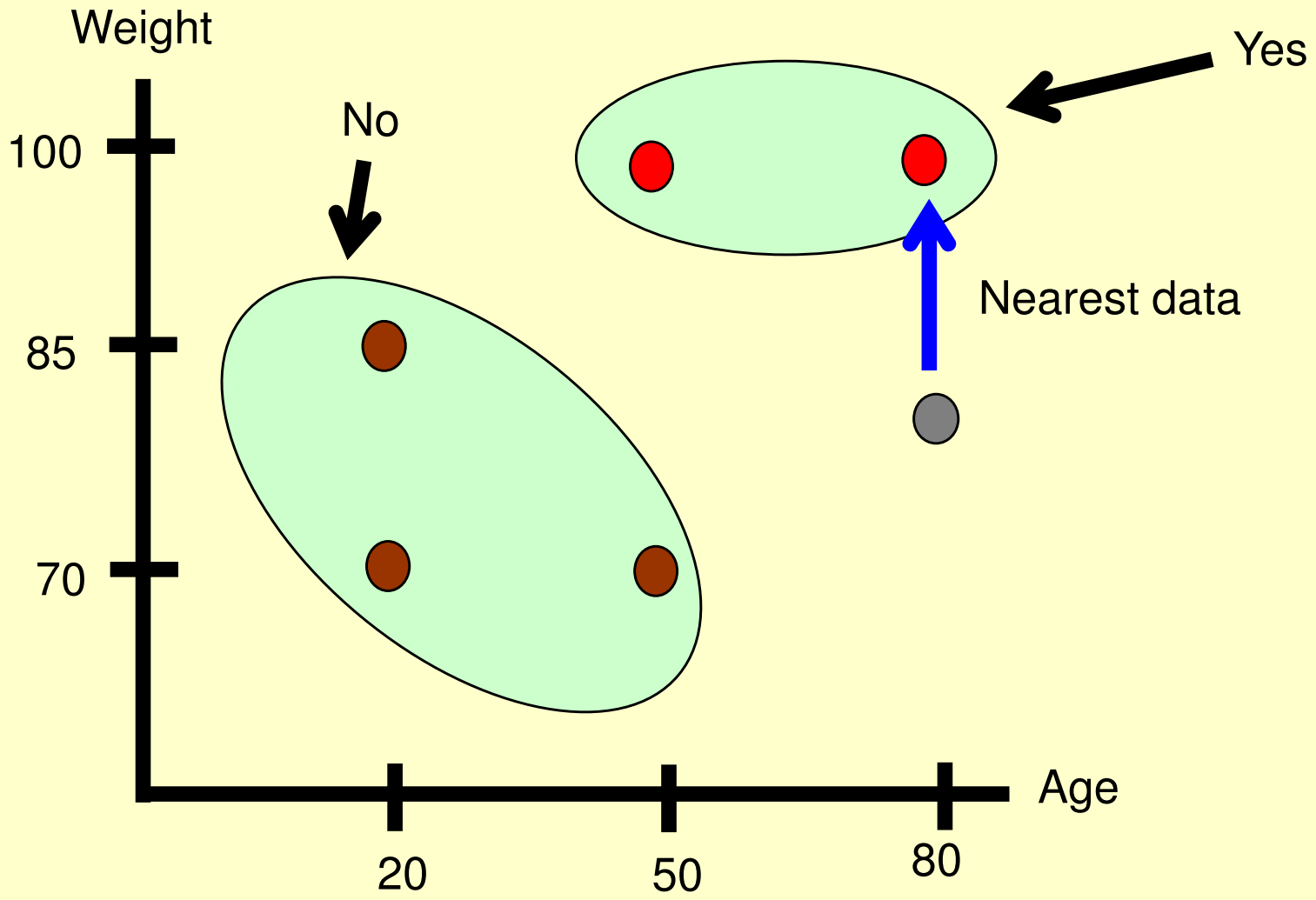
Example

Age	Weight	Hypertension
20	70	No
20	85	No
50	70	No
50	100	Yes
80	100	Yes

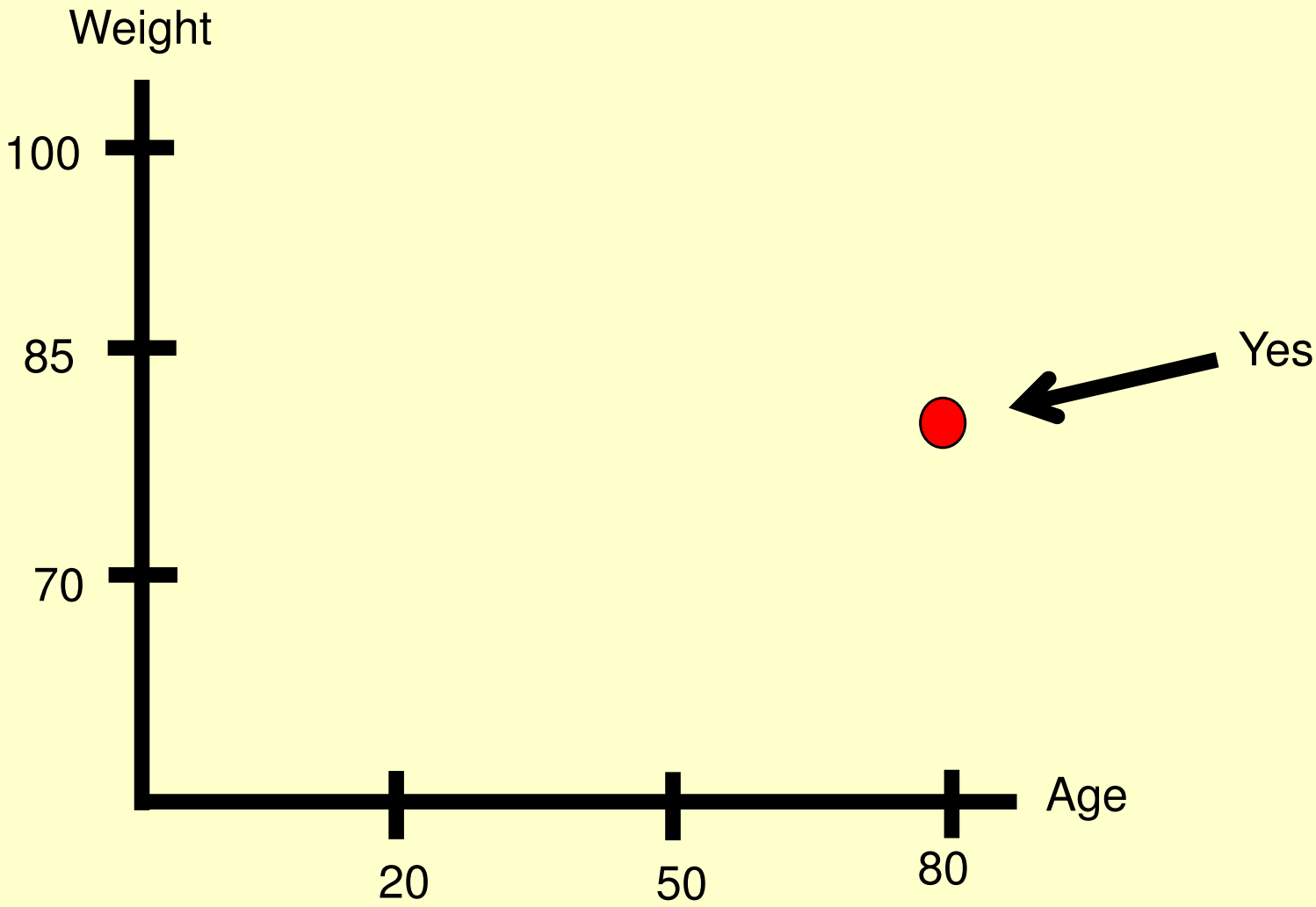
80	85	?

a new data

Classification with 1-NN



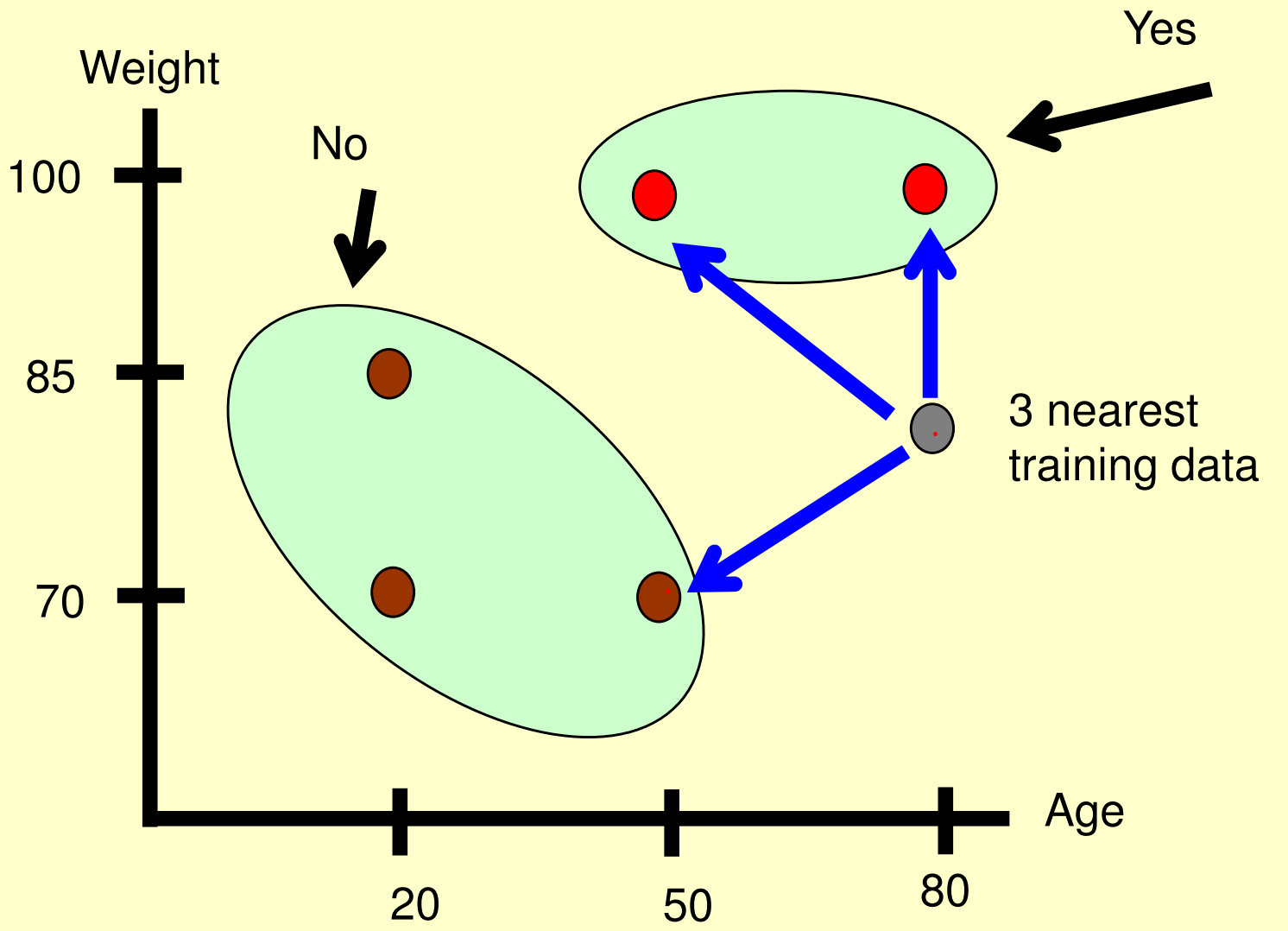
Classification with 1-NN



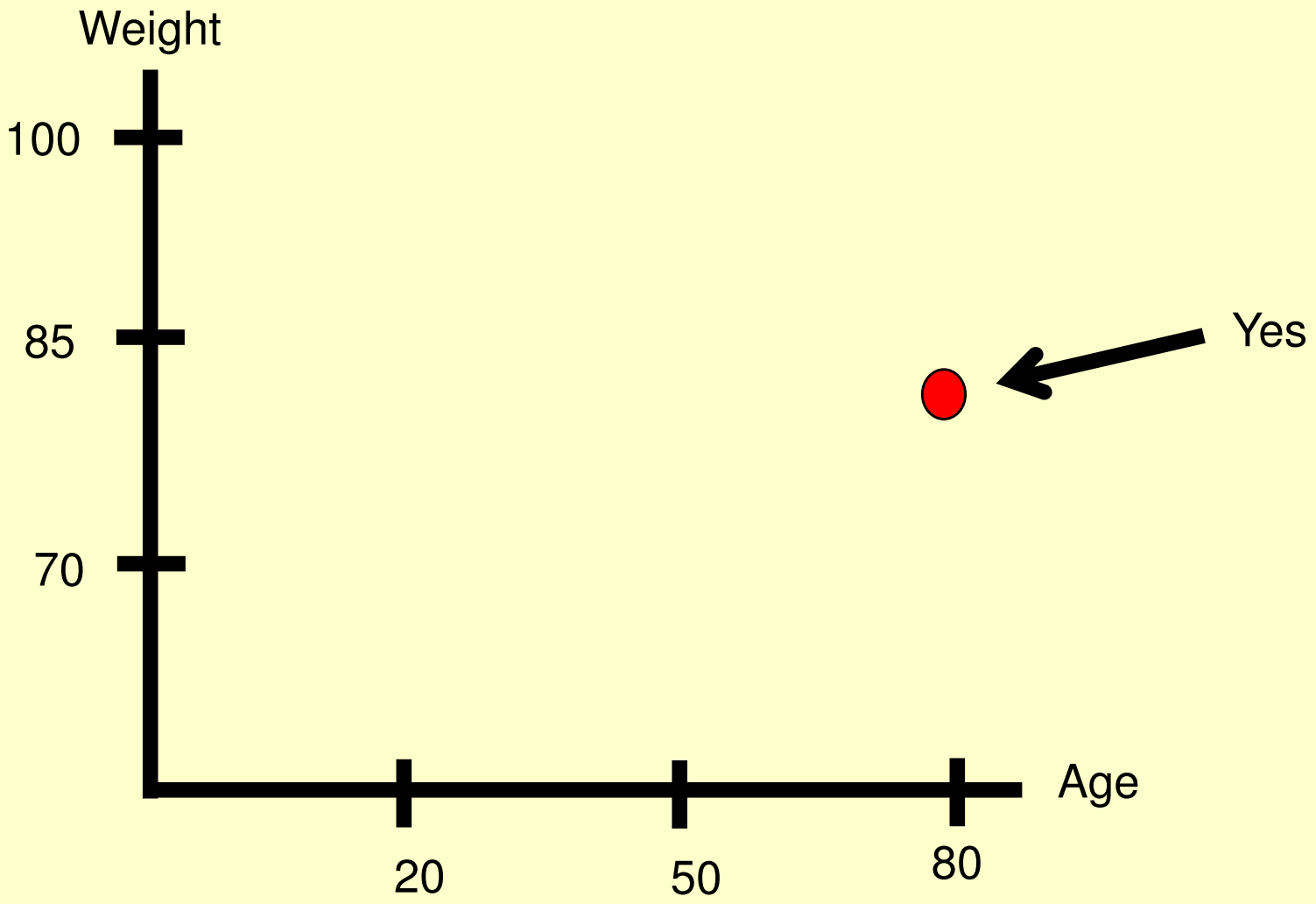
k-NN Algorithm

- Determine k
- Calculate the distance between a new data to all training data
- Find k nearest training data
- Vote class labels of the k nearest training data and classify a new data into the winning vote class label

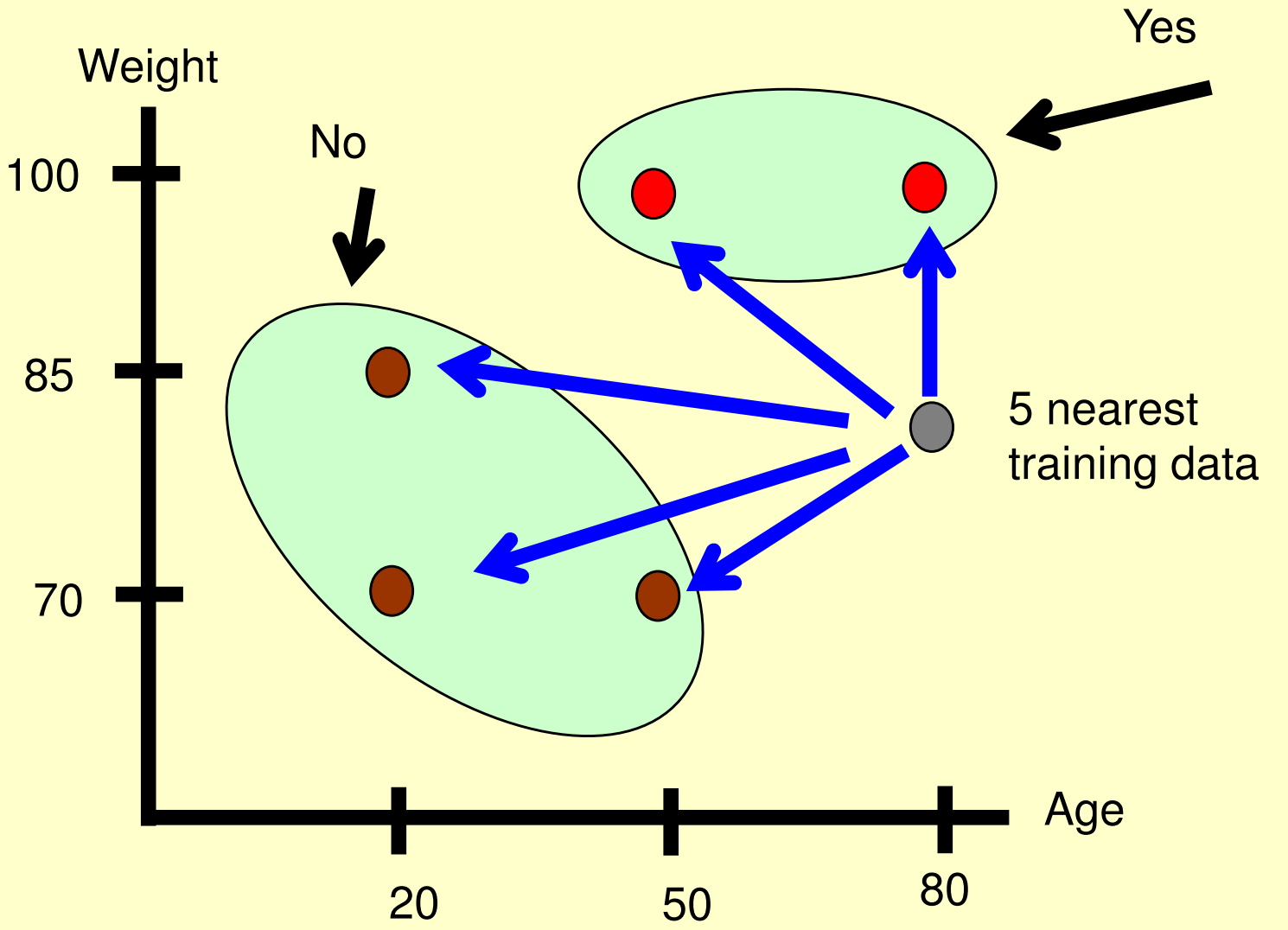
For example, k=3



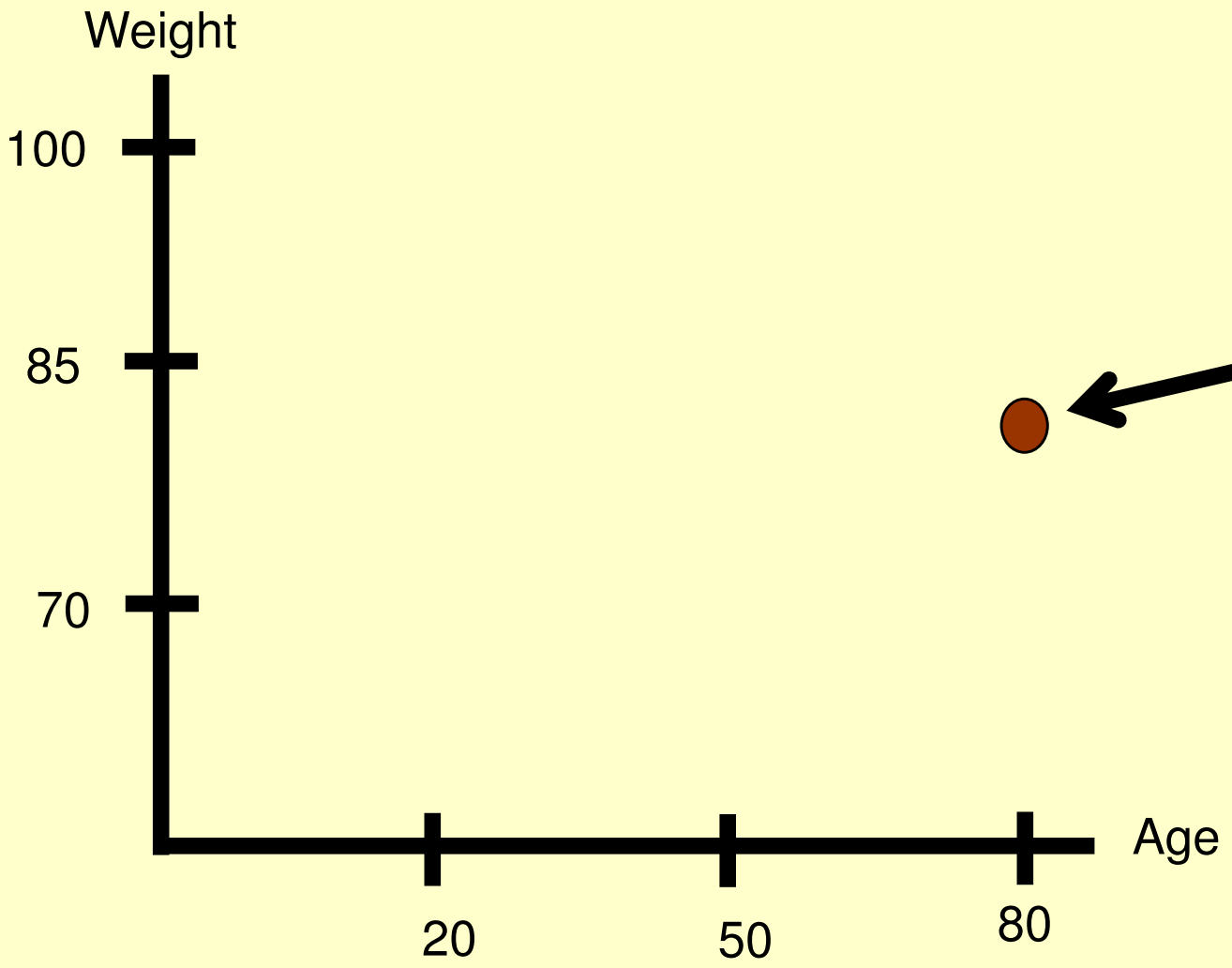
For example, k=3



For example, k=5



For example, k=5

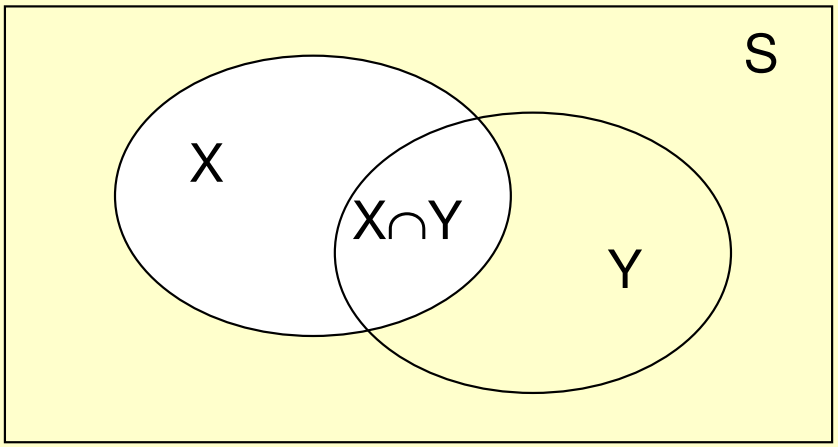


- Advantage:
  - Simple and analytically tractable
  - Possible to apply parallel implementation
  - Error rate: > bayesian, < 2xbayesian
- Disadvantage:
  - Needs huge memory and computation

Beyesian Learning

- Metode Find-S tidak dapat digunakan untuk data yang tidak konsisten dan data yang bias, sehingga untuk bentuk data semacam ini salah satu metode sederhana yang dapat digunakan adalah metode bayes.
- Metode Bayes ini merupakan metode yang baik di dalam mesin pembelajaran berdasarkan data training, dengan menggunakan probabilitas bersyarat sebagai dasarnya.

Probabilitas Bersyarat



$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$$

Probabilitas X di dalam Y adalah probabilitas interseksi X dan Y dari probabilitas Y, atau dengan bahasa lain P(X|Y) adalah prosentase banyaknya X di dalam Y

Probabilitas Bersyarat Dalam Data

#	Cuaca	Temperatur	Kecepatan Angin	Berolah-raga
1	Cerah	Normal	Pelan	Ya
2	Cerah	Normal	Pelan	Ya
3	Hujan	Tinggi	Pelan	Tidak
4	Cerah	Normal	Kencang	Ya
5	Hujan	Tinggi	Kencang	Tidak
6	Cerah	Normal	Pelan	Ya

Banyaknya data berolah-raga=ya adalah 4 dari 6 data maka dituliskan
P(Olahraga=Ya) = 4/6

Banyaknya data cuaca=cerah dan berolah-raga=ya adalah 4 dari 6 data maka dituliskan
P(cuaca=cerah dan Olahraga=Ya) = 4/6

$$P(cuaca = cerah|olahraga = ya) = \frac{4/6}{4/6} = 1$$

Probabilitas Bersyarat Dalam Data

#	Cuaca	Temperatur	Berolahraga
1	cerah	normal	ya
2	cerah	tinggi	ya
3	hujan	tinggi	tidak
4	cerah	tinggi	tidak
5	hujan	normal	tidak
6	cerah	normal	ya

Banyaknya data berolah-raga=ya adalah 3 dari 6 data maka dituliskan

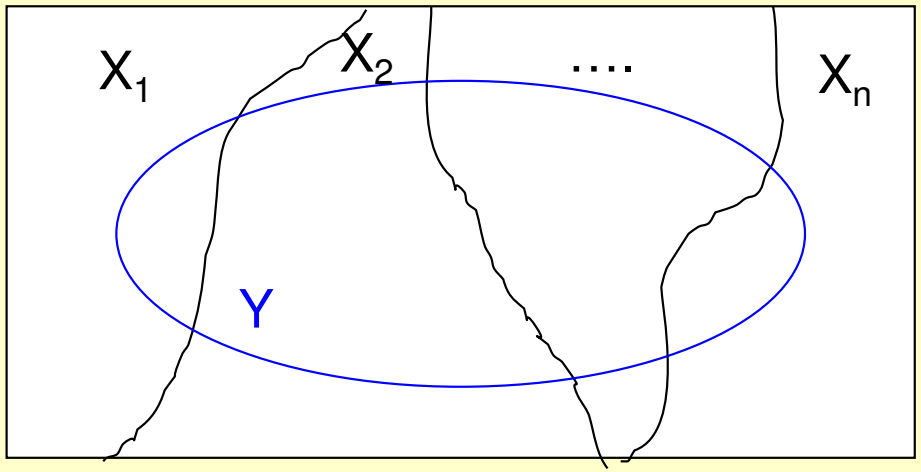
P(Olahraga=Ya) = 3/6

Banyaknya data cuaca=cerah, temperatur=normal dan berolah-raga=ya adalah 4 dari 6 data maka dituliskan

P(cuaca=cerah, temperatur=normal, Olahraga=Ya) = 2/6

$$P(cuaca = cerah, temperatur = normal | olahraga = ya) = \frac{2/6}{3/6} = \frac{2}{3}$$

Bayes Theorem



$$P(X_k \mid Y) \quad \frac{P(X_k \cap Y)}{P(Y)}$$

$$P(Y \mid X_k) = \frac{P(X_k \cap Y)}{P(X_k)}$$

$$P(X_k \cap Y) = P(Y \mid X_k)\, P(X_k)$$

$$P(Y) = \sum_i P(Y \mid X_i)$$

Keadaan Posterior (Probabilitas X_k di dalam Y) dapat dihitung dari keadaan prior (Probabilitas Y di dalam X_k dibagi dengan jumlah dari semua probabilitas Y di dalam semua X_i)

$$P(X_k \mid Y) = \frac{P(Y \mid X_k)\, P(X_k)}{\sum_i P(Y \mid X_i)}$$

Naïve Bayes Classifier

- Let each instance *x* of a training set *D* be described by a conjunction of *n* attribute values $<a_1, a_2, .., a_n>$ and let *f(x),* the target function, be such that $f(x) \in V$, a finite set.

- **Bayesian Approach:**

$$vMAP = argmax_{vj \in V} P(v_j | a_1, a_2, .., a_n)$$
$$= argmax_{vj \in V} [P(a_1, a_2, .., a_n | v_j) P(v_j) / P(a_1, a_2, .., a_n)]$$
$$= argmax_{vj \in V} [P(a_1, a_2, .., a_n | v_j) P(v_j)$$

- **Naïve Bayesian Approach:** We assume that the attribute values are conditionally independent so that $P(a1,a2,..,an|vj) = \prod_i P(a_i|v_j)$ [and not too large a data set is required.]

- ***Naïve Bayes Classifier:***

$$v_{NB} = argmax_{vj \in V} P(v_j) \prod_i P(a_i|v_j)$$

Naïve Bayes Classifier

#	Cuaca	Temperatur	Kecepatan Angin	Berolah-raga
1	Cerah	Normal	Pelan	Ya
2	Cerah	Normal	Pelan	Ya
3	Hujan	Tinggi	Pelan	Tidak
4	Cerah	Normal	Kencang	Ya
5	Hujan	Tinggi	Kencang	Tidak
6	Cerah	Normal	Pelan	Ya

Apakah bila cuaca cerah dan kecepatan angin kencang, orang akan berolahraga?

Fakta: $P(X1=cerah|Y=ya) = 1$, $P(X1=cerah|Y=tidak) = 0$
$P(X3=kencang|Y=ya) = 1/4$, $P(X3=kencang|Y=tidak) = 1/2$

$P(X1=cerah,X3=kencang \mid Y=ya)$
$= \{ P(X1=cerah|Y=ya).P(X3=kencang|Y=ya) \} . P(Y=ya)$
$= \{ (1) . (1/4) \} . (4/6) = 1/6$
$P(X1=cerah,X3=kencang \mid Y=tidak)$
$= \{ P(X1=cerah|Y=tidak).P(X3=kencang|Y=tidak) \} . P(Y=tidak)$
$= \{ (0) . (1/2) \} . (2/6) = 0$

KEPUTUSAN ADALAH BEROLAHRAGA = YA

Kelemahan Metode Bayes

- Metode Bayes hanya bisa digunakan untuk persoalan klasifikasi dengan *supervised learning*.
- Metode Bayes memerlukan pengetahuan awal untuk dapat mengambil suatu keputusan. Tingkat keberhasilan metode ini sangat tergantung pada pengetahuan awal yang diberikan.

Beberapa Aplikasi Metode Bayes

- Menentukan diagnosa suatu penyakit berdasarkan data-data gejala (sebagai contoh hipertensi atau sakit jantung).
- Mengenali buah berdasarkan fitur-fitur buah seperti warna, bentuk, rasa dan lain-lain
- Mengenali warna berdasarkan fitur indeks warna RGB
- Mendeteksi warna kulit (*skin detection*) berdarkan fitur warna chrominant
- Menentukan keputusan aksi (olahraga, art, psikologi) berdasarkan keadaan.
- Menentukan jenis pakaian yang cocok untuk keadaan-keadaan tertentu (seperti cuaca, musim, temperatur, acara, waktu, tempat dan lain-lain)
- Menentukan ekspresi (sedih, gembira, dll) dari kalimat yang diucapkan