Mining and Processing Biomedical Data

Dr. rer. nat. Krisztian Buza
adiunkt naukowy
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw, Poland
chrisbuza@yahoo.com

Introduction to Classification

- Classification is the common denominator of various recognition tasks

Illustrative example: risk factors of smoking

- Which children are likely to start smoking later, after they have grown up?

Illustrative example: risk factors of smoking

- Which children are likely to start smoking later, after they have grown up?
- Assume that none of the children is smoking at the age of 10 years
- We ask them some questions at the age of 10
  - Do you have brothers or sister?
  - Do you have an older brother?
  - Does anyone smoke in the family?
  - …
- After 10 years: we ask them if they started smoking

Illustrative example: risk factors of smoking

Name	(Q1)	(Q2)	(Q3)	...
Anna	1	1	0	...
Peter	0	0	0	...
John	1	0	1	...
...	...	...	...	...

Data: answers to the original questions that were asked at the age of 10

Questions:
(Q1) – Do you have brothers or sisters?
(Q2) – Do you have an older brother?
(Q3) – Does anyone smoke in the family?

Answers are coded by 0 and 1,
1 = yes, 0 = no

Name	Did you start smoking?
Anna	1
Peter	0
John	1
...	...

Question that is asked 10 years after the obtaining the original data

Illustrative example: risk factors of smoking

Name	(Q1)	(Q2)	(Q3)	...
Anna	1	1	0	...
Peter	0	0	0	...
John	1	0	1	...
...	...	...	...	...

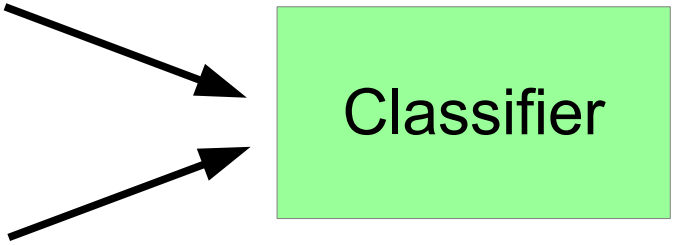
Name	Did you start smoking?
Anna	1
Peter	0
John	1
...	...

We construct a model, called classifier, that is able to predict, based on the answers to the original questions, who is likely to start smoking

Illustrative example: risk factors of smoking

Name	(Q1)	(Q2)	(Q3)	...
Anna	1	1	0	...
Peter	0	0	0	...
John	1	0	1	...
...	...	...	...	...

Name	Did you start smoking?
Anna	1
Peter	0
John	1
...	...



Name	(Q1)	(Q2)	(Q3)	...
Kate	0	0	1	...
...	...	...	...	...

Name	Prediction
Kate	1
...	...

The data obtained from some children who are at the age of 10 now

The model predicts if these children are likely to start smoking

Classification

- Classes: groups of instances (persons, objects, etc.); these groups are defined a priori, for example:
  - the group denoted by 1 is the group of children who start smoking,
  - the group denoted by 0 is the group of children who do not start smoking.
- Aim of classification:
  - understanding the data
  - predictions

Illustrative example: risk factors of smoking

Name	(Q1)	(Q2)	(Q3)	...
Anna	1	1	0	...
Peter	0	0	0	
John				
...	...	...	...	...

Train data: "historical" data used to construct the model

Name	Did you start smoking?
Anna	1
Peter	0
John	1
...	...

Classifier

Name	(Q1)	(Q2)	(Q3)	...
Kate	0	0	1	...
...				

Prediction data / test data: "new data"

Name	Prediction
Kate	1
...	...

The model predicts if these children are likely to start smoking

Illustrative example: risk factors of smoking

Name	(Q1)	(Q2)	(Q3)	...
Anna	1	1	0	...
Peter	0	0	0	
John				
...	...	...	...	...

Train data: "historical" data used to construct the model

Name	Did you start smoking?
Anna	1
Peter	
John	
...	

Class label gives which group the instance belongs to

Classifier

Name	(Q1)	(Q2)	(Q3)	...
Kate	0	0	1	...
..				

Prediction data / test data: "new data"

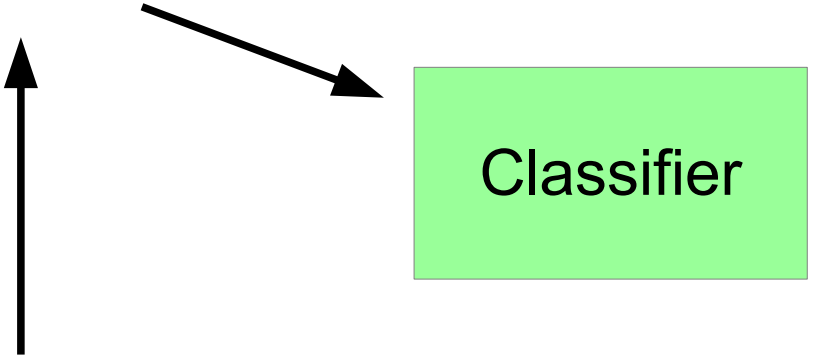
Name	Prediction
Kate	1
...	...

The model predicts if these children are likely to start smoking

Illustrative example: risk factors of smoking

Name	(Q1)	(Q2)	(Q3)	...
Anna	1	1	0	...
Peter	0	0	0	...
John	1	0	1	...
...	...	...	...	...

Name	Did you start smoking?
Anna	0
Peter	0
John	1
...	...



Train data: “historical” data used to construct the model

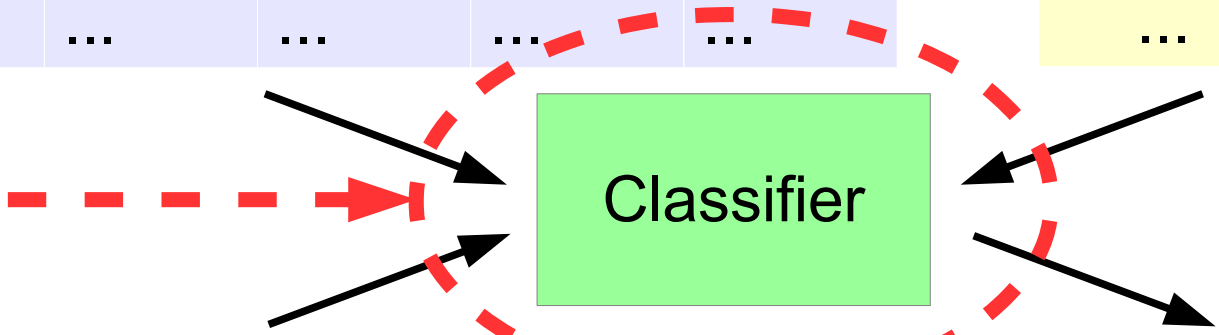
Class labels

Decision Tree and Nearest Neighbor

Illustrative example: risk factors of smoking

Name	(Q1)	(Q2)	(Q3)	...
Anna	1	1	0	...
Peter	0	0	0	...
John	1	0	1	...
...	...	...	...	...

Name	Did you start smoking?
Anna	1
Peter	0
John	1
...	...



Name	(Q1)	(Q2)	(Q3)	...
Kate	0	0	1	...
...	...	...	...	...

Name	Prediction
Kate	1
...	...

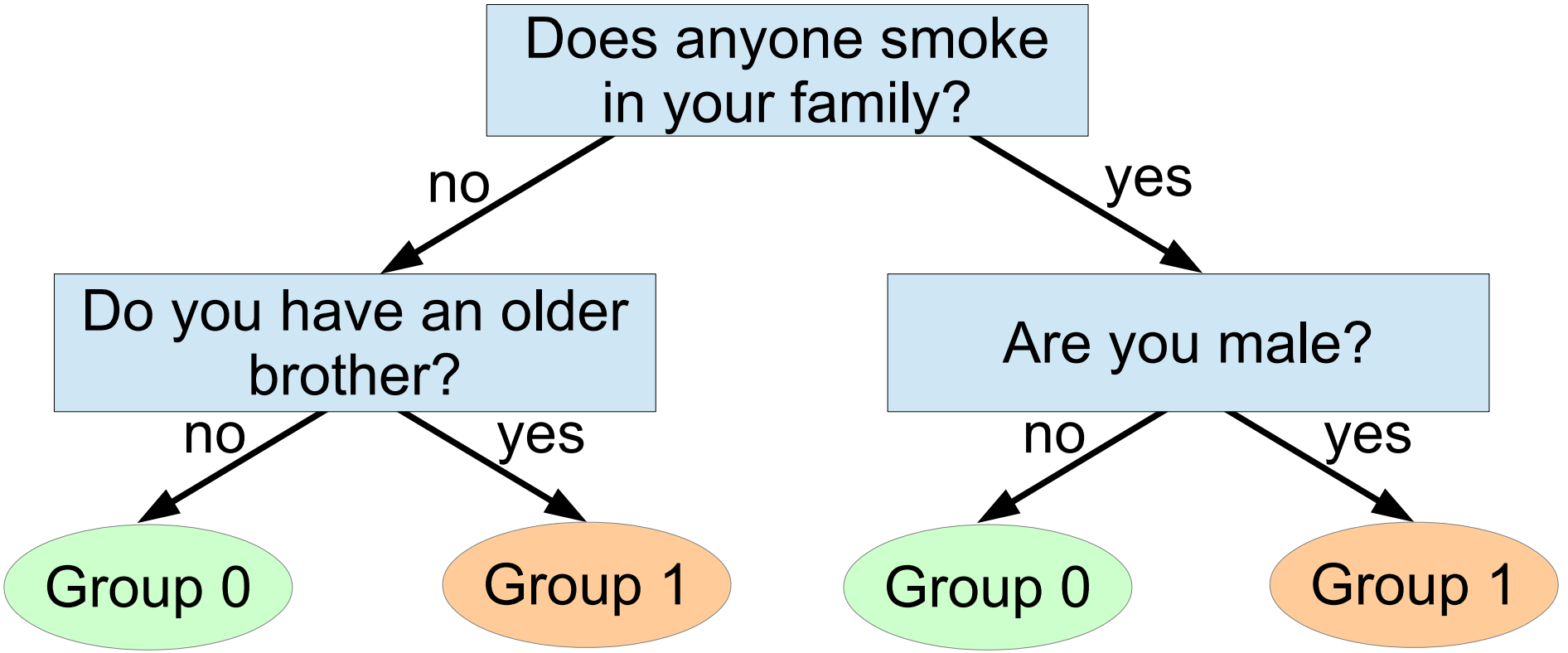
The data obtained from some children who are at the age of 10 now

The model predicts if these children are likely to start smoking

Classifiers

- Neural Networks
- Support Vector Machines (SVMs)
- Hidden Markov Models (HMMs)
- Rule-based classifiers
- Decision Trees
- Nearest Neighbor

Illustrative example: Decision Tree



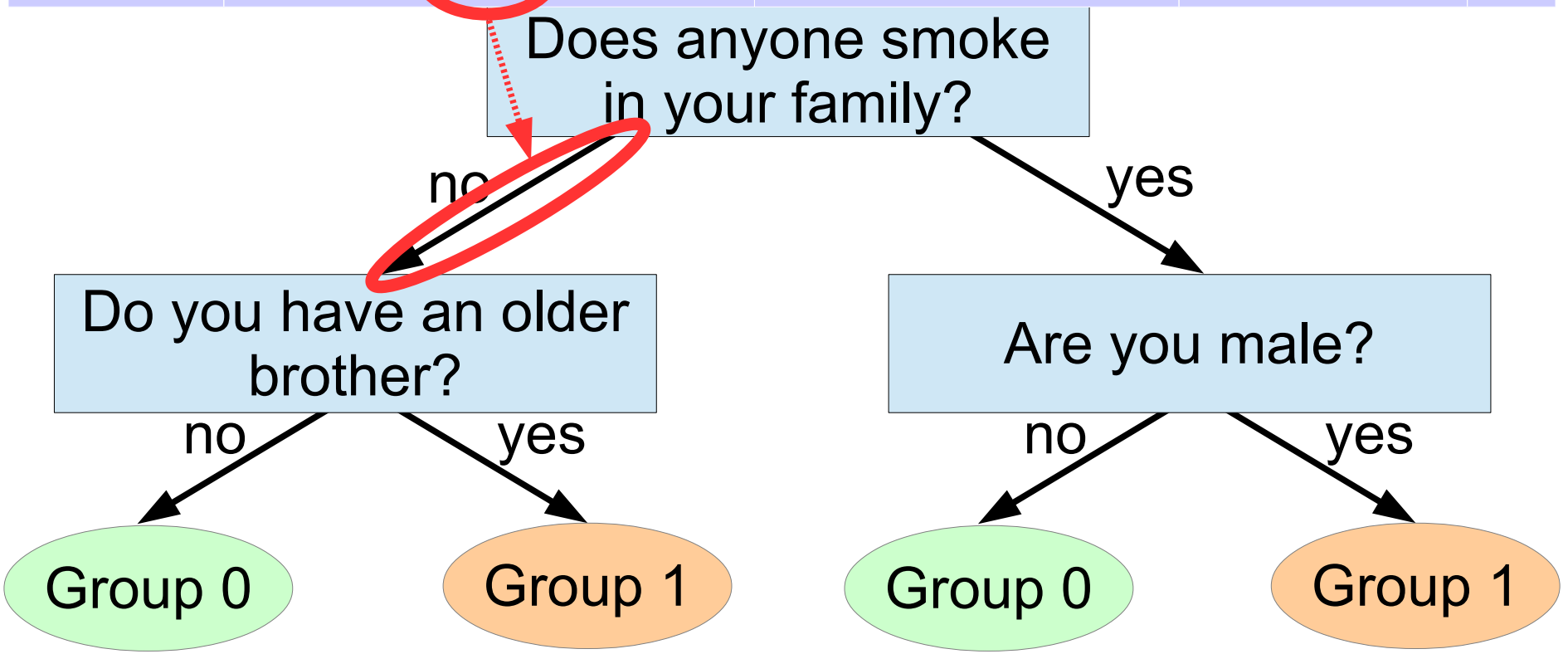
Legend
Is the person expected to start smoking?

Group 0 =No

Group 1 =Yes

New instance to be classified

Name	Does anyone smoke in your family?	Do you have an older brother?	Do you…?	...
Kate	no	yes	...	...



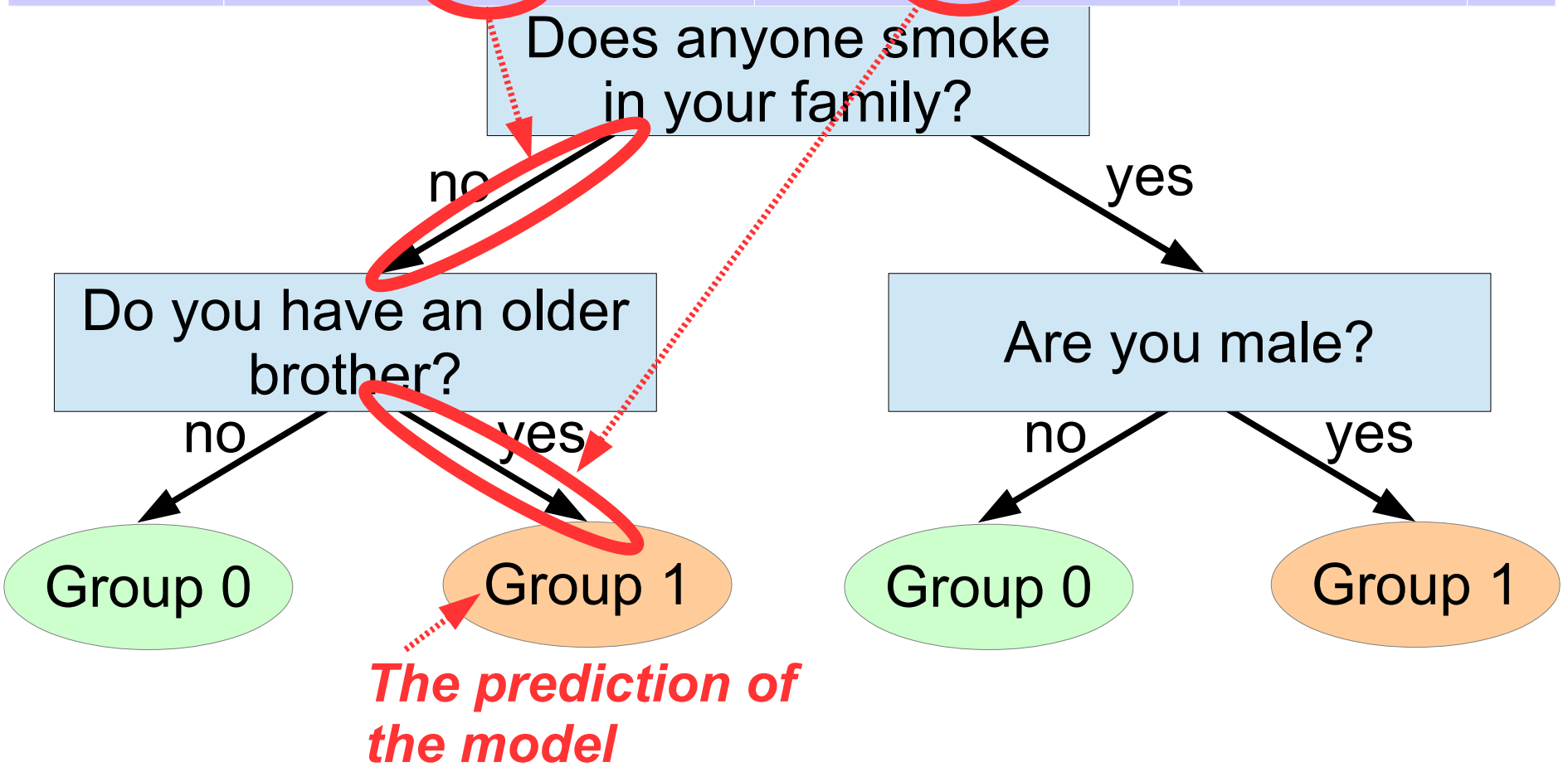
Legend
Is the person expected to start smoking?

Group 0 =No

Group 1 =Yes

New instance to be classified

Name	Does anyone smoke in your family?	Do you have an older brother?	Do you…?	...
Kate	no	yes	...	...



Legend
Is the person expected to start smoking?

Group 0 =No

Group 1 =Yes

Classifiers

- Neural Networks
- Support Vector Machines (SVMs)
- Hidden Markov Models (HMMs)
- Rule-based classifiers
- Decision Trees
- Nearest Neighbor

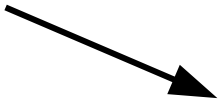
Illustrative example: risk factors of smoking – classification via nearest neighbor

Train data

Name	(Q1)	(Q2)	(Q3)	...	Class label
Anna	1	1	0	...	1
Peter	0	0	0	...	0
John	1	0	1	...	0
Tom	0	0	1	...	1

Instance to be classified

Name	(Q1)	(Q2)	(Q3)	...
Kate	0	0	1	...



The most similar instance ("Tom") has the class label "1" → the prediction of the model is "1"

Name	Prediction
Kate	1

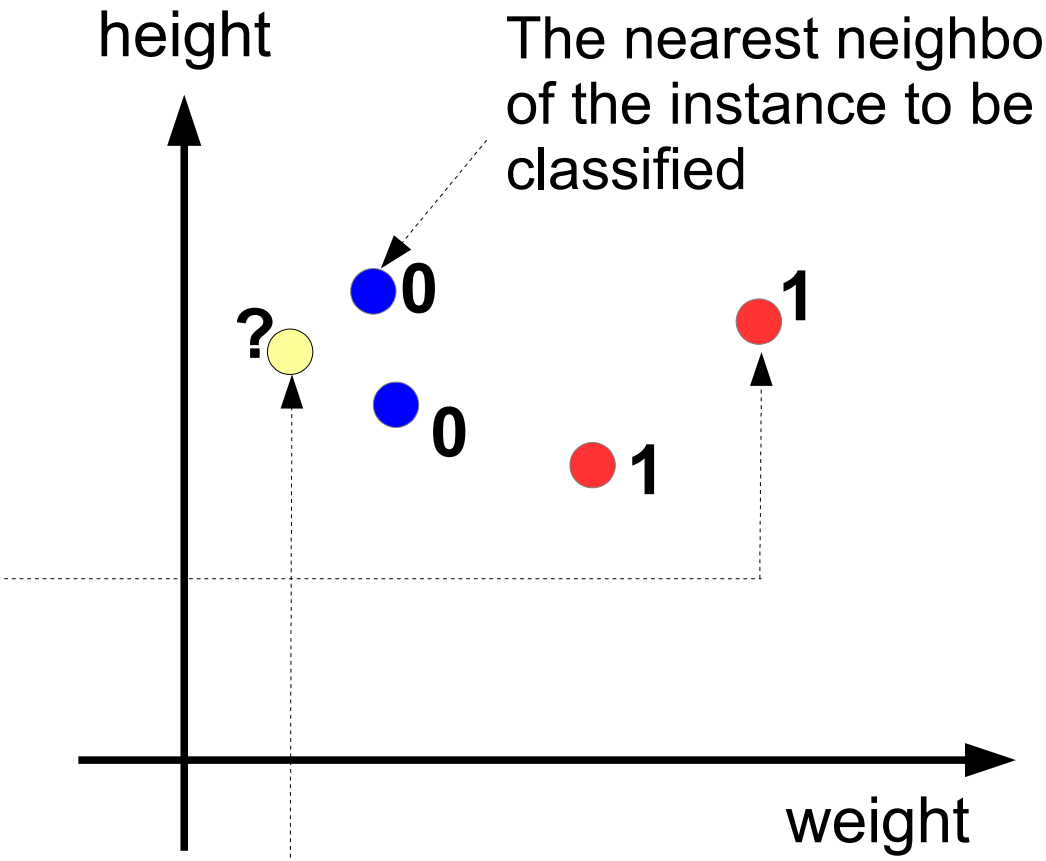
Nearest Neighbor Classification of Vector Data (Example)

Train data

Name	weight (kg)	height (cm)	Class label
Anna	95	162	1
Peter	83	173	0
John	79	181	0
Tom	108	180	1

Instance to be classified

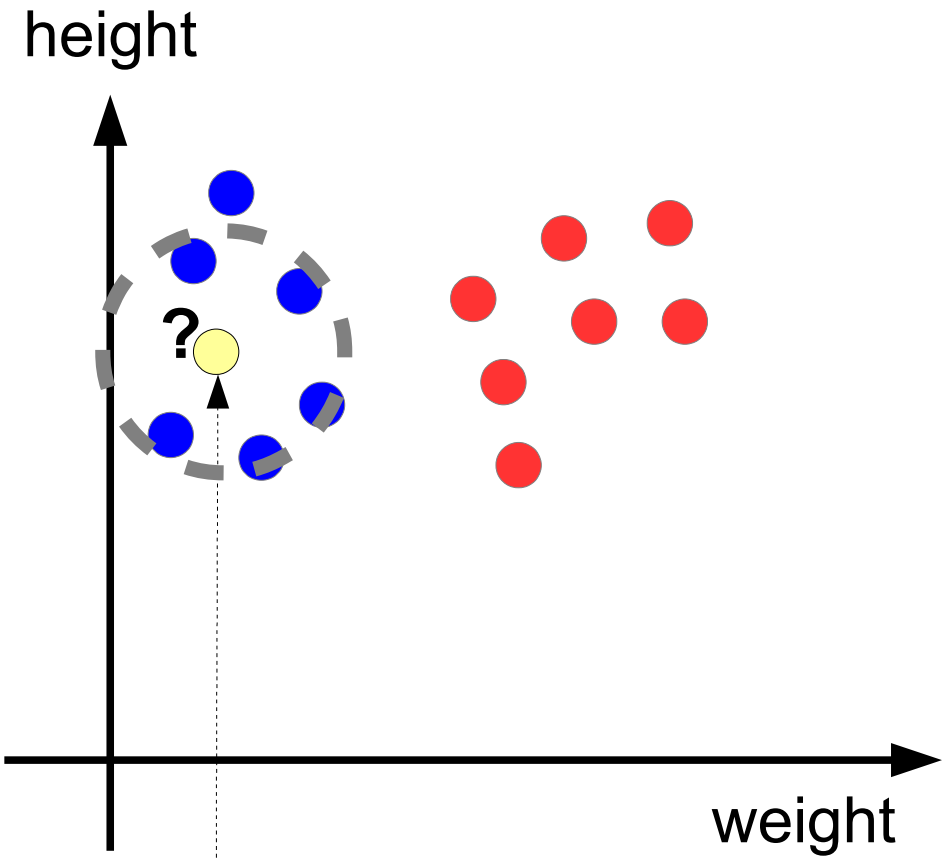
Name	weight	height
Kate	65	167



Instances belonging to class 0 and 1 are denoted by blue and red circles respectively. The yellow circle denotes the instance to be classified.

Nearest Neighbor Classification of Vector Data (Example)

In order to classify an instance, we may consider several nearest neighbors (i.e., several most similar train instances), not only one.



Instance to be classified

Name	weight	height
Kate	65	167

Instances belonging to class 0 and 1 are denoted by blue and red circles respectively. The yellow circle denotes the instance to be classified.