DM I: Block 'Classification'

Unit 'Underpinnings'

Myra Spiliopoulou

FACULTY OF
COMPUTER SCIENCE

1 Phases of Classification

2 Example datasets

3 Closing

The two phases of classification

Classification – Learning Phase

A classification algorithm takes as input a set of mutually exclusive classes $\mathscr{C} = \{C_1, \ldots, C_k\}$
and, for each $C_i \in \mathscr{C}, i = 1 \ldots k$, a set of instances D_i belonging to C_i.

The two phases of classification

Classification – Learning Phase

A classification algorithm takes as input a set of mutually exclusive classes $\mathscr{C} = \{C_1, \ldots, C_k\}$
and, for each $C_i \in \mathscr{C}, i = 1 \ldots k$, a set of instances D_i belonging to C_i.

The set $D = \cup_{i=1}^{k} D_i$ is used for *learning*, i.e. for training.
D must be representative of the population $\mathscr{D}$ under study.

The two phases of classification

Classification – Learning Phase

A classification algorithm takes as input a set of mutually exclusive classes $\mathscr{C} = \{C_1, \ldots, C_k\}$
and, for each $C_i \in \mathscr{C}, i = 1 \ldots k$, a set of instances D_i belonging to C_i.

The set $D = \cup_{i=1}^{k} D_i$ is used for *learning*, i.e. for training.
D must be representative of the population $\mathscr{D}$ under study.

The classification algorithm builds a *classifier* ξ,
i.e. a model that reflects what makes the instances in each class different from the instances in the other classes.

The two phases of classification

Classification – Learning Phase

A classification algorithm takes as input a set of mutually exclusive classes $\mathscr{C} = \{C_1, \ldots, C_k\}$
and, for each $C_i \in \mathscr{C}, i = 1 \ldots k$, a set of instances D_i belonging to C_i.

The set $D = \cup_{i=1}^{k} D_i$ is used for *learning*, i.e. for training.
D must be representative of the population $\mathscr{D}$ under study.

The classification algorithm builds a *classifier* ξ,
i.e. a model that reflects what makes the instances in each class different from the instances in the other classes.

Classification – Querying Phase

The classifier ξ is used as an ”oracle”.
For each instance $x \in \mathscr{D}$, it decides to which class from $\mathscr{C}$ does x belong.

Classification – Learning Phase: Model induction & deduction

Inducing and deducing a model on data of known class membership *before* applying the model on data of unknown class membership.

Example from Tan, Steinbach & Kumar, Ch.4 (2006)

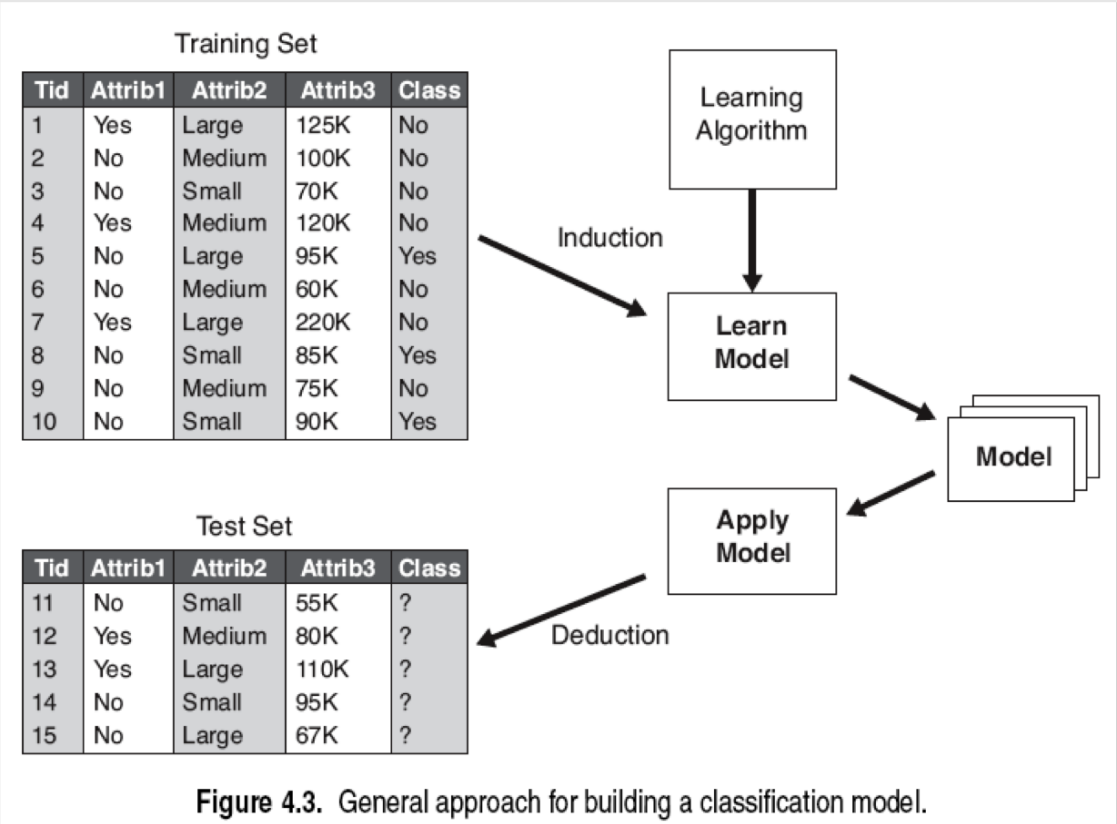

Figure 4.3. General approach for building a classification model.

The two phases of classification

Classification – Learning Phase

A classification algorithm takes as input a set of mutually exclusive classes $\mathscr{C} = \{C_1, \ldots, C_k\}$
and, for each $C_i \in \mathscr{C}, i = 1 \ldots k$, a set of instances D_i belonging to C_i.

The set $D = \cup_{i=1}^{k} D_i$ is used for *learning*, i.e. for training.
D must be representative of the population $\mathscr{D}$ under study.

The two phases of classification

Classification – Learning Phase

A classification algorithm takes as input a set of mutually exclusive classes $\mathscr{C} = \{C_1, \ldots, C_k\}$
and, for each $C_i \in \mathscr{C}, i = 1 \ldots k$, a set of instances D_i belonging to C_i.

The set $D = \cup_{i=1}^{k} D_i$ is used for *learning*, i.e. for training.
D must be representative of the population $\mathscr{D}$ under study.

We split D into a *training set* to be used for model induction
and a *test set* to be used for model deduction (includes: evaluation).

The two phases of classification

Classification – Learning Phase

A classification algorithm takes as input a set of mutually exclusive classes $\mathscr{C} = \{C_1, \ldots, C_k\}$
and, for each $C_i \in \mathscr{C}, i = 1 \ldots k$, a set of instances D_i belonging to C_i.

The set $D = \cup_{i=1}^{k} D_i$ is used for *learning*, i.e. for training.
D must be representative of the population $\mathscr{D}$ under study.

We split D into a *training set* to be used for model induction
and a *test set* to be used for model deduction (includes: evaluation).

The classification algorithm builds a *classifier* ξ,
i.e. a model that reflects what makes the instances in each class different from the instances in the other classes.

The two phases of classification

Classification – Learning Phase

A classification algorithm takes as input a set of mutually exclusive classes $\mathscr{C} = \{C_1, \ldots, C_k\}$
and, for each $C_i \in \mathscr{C}, i = 1 \ldots k$, a set of instances D_i belonging to C_i.

The set $D = \cup_{i=1}^{k} D_i$ is used for *learning*, i.e. for training.
D must be representative of the population $\mathscr{D}$ under study.

We split D into a *training set* to be used for model induction
and a *test set* to be used for model deduction (includes: evaluation).

The classification algorithm builds a *classifier* ξ,
i.e. a model that reflects what makes the instances in each class different from the instances in the other classes.

Classification – Querying Phase

The classifier ξ is used as an ”oracle”.
For each instance $x \in \mathscr{D}$, it decides to which class from $\mathscr{C}$ does x belong.

1 Phases of Classification

2 Example datasets

3 Closing

Classification – Example on verterbrate classification

Original dataset – Tan, Steinbach & Kumar, Ch.4 (2006)

- **Classes:** amphibian, bird, fish, mammal, reptile
- **Feature space:** Body Temperature, Skin Cover, Gives Birth, Aquatic Creature, Aerial Creature, Has Legs, Hibernates

Table 4.1. The vertebrate data set.

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard shark	cold-blooded	scales	yes	yes	no	no	no	fish
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

Classification Example Dataset 1 – Verterbrates_binary

Modified dataset, based on Tan, Steinbach & Kumar, Ch.4 (2006): two classes 'Mammal yes/no', 6 variables + Id variable

Id	Body Temperature (1)	Skin Cover	Gives Birth (2)	Aquatic (3)	Aerial (4)	Legs (5)	Hiber-nates (6)	Mammal
human	warm-blooded	hair	yes	no	no	two	no	yes
python	cold-blooded	scales	no	no	no	zero	yes	no
salmon	cold-blooded	scales	no	yes	no	zero	no	no
whale	warm-blooded	hair	yes	yes	no	zero	no	yes
frog	cold-blooded	none	no	semi	no	four	no	no
komodo dragon	cold-blooded	scales	no	no	no	four	no	no
bat	warm-blooded	hair	yes	no	yes	four	yes	yes
pigeon	warm-blooded	feathers	no	no	yes	two	no	no
cat	warm-blooded	fur	yes	no	no	four	no	yes
leopard shark	cold-blooded	scales	yes	yes	no	zero	no	no
turtle	cold-blooded	scales	no	semi	no	four	no	no
penguin	warm-blooded	feathers	no	semi	no	two	no	no
porcupine	warm-blooded	quills	yes	no	no	four	yes	yes
eel	cold-blooded	scales	no	yes	no	zero	no	no
salamander	cold-blooded	none	no	semi	no	four	yes	no

Classification – Another verterbrate dataset example

Verterbrate classification on

- **Classes:** mammal, non-mammal
- **Feature space:** Body Temperature, Gives Birth, Four-legged, Hibernates

from Tan, Steinbach & Kumar, Ch.4 (2006)

Training Set (cf. Table 4.3, with modifications in two instances!)

Name	Body Temperature	Gives Birth	Four-legged	Hiber-nates	Class label
porcupine	warm-blooded	yes	yes	yes	mammal
cat	warm-blooded	yes	yes	no	mammal
bat	warm-blooded	yes	no	yes	mammal
whale	warm-blooded	yes	no	no	mammal
salamander	cold-blooded	no	yes	yes	non-mammal
komodo dragon	cold-blooded	no	yes	no	non-mammal
python	cold-blooded	no	no	yes	non-mammal
salmon	cold-blooded	no	no	no	non-mammal
eagle	warm-blooded	no	no	no	non-mammal
guppy	cold-blooded	yes	no	no	non-mammal

Classification – Another verterbrate dataset example

Verterbrate classification on

- **Classes:** mammal, non-mammal
- **Feature space:** Body Temperature, Gives Birth, Four-legged, Hibernates

from Tan, Steinbach & Kumar, Ch.4 (2006)

Test Set (cf. Table 4.4 – one instance modified!)

Name	Body temperature	Gives Birth	four legged	Hiber-nates	Class label
human	warm-blooded	yes	no	no	mammal
pigeon	warm-blooded	no	no	no	non-mammal
elephant	warm-blooded	yes	no	no	mammal
leopard shark	cold-blooded	yes	no	no	non-mammal
turtle	cold-blooded	no	no	no	non-mammal
penguin	warm-blooded	no	no	no	non-mammal
eel	cold-blooded	no	no	no	non-mammal
dolphin	warm-blooded	yes	no	no	mammal
spiny anteater	warm-blooded	no	yes	yes	mammal
gila monster	cold-blooded	no	yes	yes	non-mammal

Classification Example Dataset 2 – Weather for playing golf

Public domain dataset ”Golf” – Witten & Eibe, Book on Mining with Java

- **Classes:** Yes, No
- **Feature space:** Outlook, Temperature [1], Humidity, Windy

Golf dataset as running example

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

[1] also: ”Temp”

. . . and the fictive dataset on patient responses to treatment

Id	I2	I15	I30	I22	I31	I9	I26	Response
#1	f	VH	yes	better	no	r	no	yes
#2	m	M	no	better	no	b	yes	no
#3	m	M	no	worse	no	b	no	no
#4	f	VH	yes	worse	no	b	no	yes
#5	m	L	no	no effect	no	l	no	no
#6	m	M	no	better	no	l	no	no
#7	f	VH	yes	better	yes	l	yes	yes
#8	f	H	no	better	yes	r	no	no
#9	f	H	yes	better	no	l	no	yes
#10	m	M	yes	worse	no	b	no	no
#11	m	M	no	no effect	no	l	no	no
#12	f	H	no	no effect	no	r	no	no
#13	f	L	yes	better	no	l	yes	yes
#14	m	M	no	worse	no	b	no	no
#15	m	L	no	no effect	no	l	yes	no

1 Phases of Classification

2 Example datasets

3 Closing

Phases, Data and . . .

- How to build a classifier? → Classification algorithms
- How to assess the quality of a classifier? → Model evaluation
- After building many classifiers:
  How to figure out which one is best? → Model comparison

Thank you very much!

Questions?