Machine Learning: Introduction

Marcin Sydow

Outline of this Lecture

- Data
- Motivation for Data Mining and Machine Learning
- Idea of Machine Learning
- Decision Table: Cases and Attributes
- Supervised and Unsupervised Learning
- Classification and Regression
- Examples

Data: Motivation for Data Mining

Observations:

1. Data is huge, interesting but hard to be analysed directly by humans
2. data is kept in digital form

Ergo: let algorithms and computers do the job!

Data is overwhelming

Each second, our civilisation produces huge amounts of data:

- WWW server logs (visits to Web pages, clicks)
- on-line shopping
- mobile phone connections
- credit card payments
- traffic sensors (cars, trains, planes)
- surveillance cameras and sensors
- stock market transactions
- scientific measurements (medicine, physics, astrophysics)
- etc...

Example tasks

Understanding the data:

- grouping similar cases
- recognising important patterns in data
- classifying new cases
- predicting the future based on previous observations
- detecting trends in data (e.g. early detection of economy crisis, etc.)

In machine learning we want the machines do the above, with limited or no support from humans

Taxonomy

- Supervised Learning
- Unsupervised Learning

Typical Scheme in Machine Learning

1. Data acquisition
2. Data cleaning and pre-processing
3. (supervised only) Division into training set and testing set
4. Learning on the data
5. Evaluating the performance (iterative)
6. Using the system

Supervised Learning

1. there is known “correct” answer provided to the system in the **training dataset** (teaching signal)
2. system learns on previously observed data (supervisor provides correct answers)
3. system applies what the learnt model to new cases

Unsupervised Learning

- no teaching signal (only “raw” data)
- task: detect some structure or mappings between the cases and attribute values (e.g. grouping, association rules)

Decision Table

Example: medical diagnostics (optical lenses)

age	prescription	astigmatism	tear prod.	DECISION
young	myope	no	reduced	NONE
young	myope	no	normal	SOFT
young	myope	yes	reduced	NONE
young	myope	yes	normal	HARD
young	hypermetrope	no	reduced	NONE
young	hypermetrope	no	normal	SOFT
young	hypermetrope	yes	reduced	NONE
young	hypermetrope	yes	normal	HARD
pre-presbyopic	myope	no	reduced	NONE
pre-presbyopic	myope	no	normal	SOFT
pre-presbyopic	myope	yes	reduced	NONE
pre-presbyopic	myope	yes	normal	HARD
pre-presbyopic	hypermetrope	no	reduced	NONE
pre-presbyopic	hypermetrope	no	normal	SOFT
pre-presbyopic	hypermetrope	yes	reduced	NONE
pre-presbyopic	hypermetrope	yes	normal	NONE
presbyopic	myope	no	reduced	NONE
presbyopic	myope	no	normal	NONE
presbyopic	myope	yes	reduced	NONE
presbyopic	myope	yes	normal	HARD
presbyopic	hypermetrope	no	reduced	NONE
presbyopic	hypermetrope	no	normal	SOFT
presbyopic	hypermetrope	yes	reduced	NONE
presbyopic	hypermetrope	yes	normal	NONE

Example: an outdoor game played only in specific but unknown weather conditions:

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

Example: Outdoor Game, cont.

Basic question:
"For which weather the game can be played, in general?"

Even if answer is unknown, there are recorded a couple of cases when the game was or was not played in particular weather conditions.

May be, it is possible to infer some "knowledge" about the good playing conditions so that it could be applied to a new case

Outdoor Game - a new case

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
overcast	cool	high	true	???

Decision Table: Cases and Attributes

Knowledge can be built on previously observed cases

Each case is described by some **attributes** of specified type (nominal or numeric)

For a given case, each of its attributes has some value (usually)

Decision Table:

- **cases** = rows
- **attributes** = columns

Attributes

- numerical or categorical
- ordered on not ordered
- scaling, attribute transformation
- quantisation (numerical -> categorical)

Outdoor Game Decision Table: Variant 1: Nominal Attributes

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

Outdoor Game Decision Table: Variant 2: Numeric Attributes

outlook	temperature	humidity	windy	PLAY?
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

Other Forms of Data

The data can be in a different form (than a rectangular matrix)

- logs (from servers, etc.)
- relational data (e.g. social networks)
- sequential (e.g. bio-informatics)
- graph data, etc.

Machine Learning

Task: to learn the relationships between the values of attributes in the given knowledge domain (i.e. decision table)

Moreover, this knowledge is to be discovered automatically by machine

There are two main paradigms:

1. Supervised Learning
2. Unsupervised Learning

Supervised Learning

1. Task: “predict” the correct value of one distinguished attribute (**decision attribute**) based on the values of other attributes
2. Learn it on an available set of observed cases (**training set**), for which the “correct” value of the decision attribute is known

It is called:

- **classification**, when the decision attribute is nominal
- **regression**, when the decision attribute is numeric

Supervised Learning, summarised

input: a new case, with unkown value of its decision attribute (values of other attributes are known)

output: the "correct" value of its decision attribute

The system can learn only on some limited number of cases (training set) for which the "correct" answer is known (provided by a supervisor)

Practical problems:

- missing values (how to reconstruct them?)
- incorrect values (how detect and correct them?)
- noisy data (how to de-noise them?)
- inconsistent data (what to do in this case?)

Classification, example 2
Botany: Species Recognition

Consider 3 different sub-species of Iris (flower)

- Iris-setosa
- Iris-versicolor
- Iris-virginica

Task: learn to identify sub-species of a plant based on sizes of its parts (attributes):

- sepal length (cm)
- sepal width (cm)
- petal length (cm)
- petal width (cm)

Species Recognition, cont.

Training set:
150 known cases (measured attributes and known correct classification to one of 3 subspecies)

The system learns on the training set

Subsequently, for any new case, classification is done based on the measurements of its sepals and petals.
The “knowledge” gained in the process of automatic learning is applied to new cases.

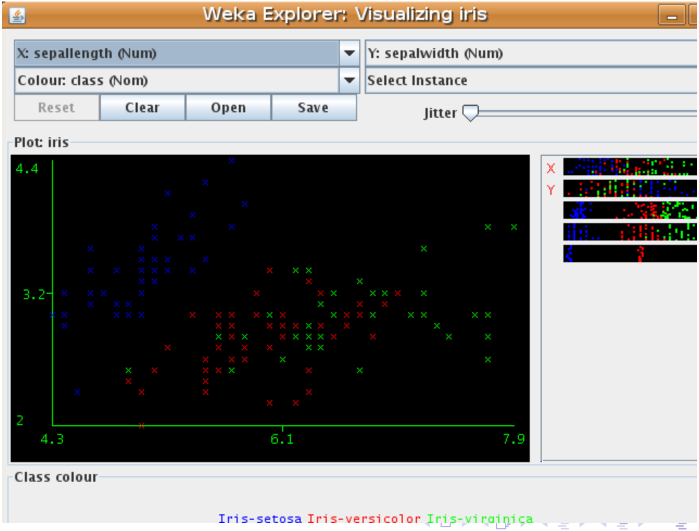
Iris sub-species recognition Dataset

S - iris setosa, V - iris versicolor, VG - iris virginica

ll	lw	pl	pw	?	ll	lw	pl	pw	?	ll	lw	pl	pw	?
5.1	3.5	1.4	0.2	S	7.0	3.2	4.7	1.4	V	6.3	3.3	6.0	2.5	VG
4.9	3.0	1.4	0.2	S	6.4	3.2	4.5	1.5	V	5.8	2.7	5.1	1.9	VG
4.7	3.2	1.3	0.2	S	6.9	3.1	4.9	1.5	V	7.1	3.0	5.9	2.1	VG
4.6	3.1	1.5	0.2	S	5.5	2.3	4.0	1.3	V	6.3	2.9	5.6	1.8	VG
5.0	3.6	1.4	0.2	S	6.5	2.8	4.6	1.5	V	6.5	3.0	5.8	2.2	VG
5.4	3.9	1.7	0.4	S	5.7	2.8	4.5	1.3	V	7.6	3.0	6.6	2.1	VG
4.6	3.4	1.4	0.3	S	6.3	3.3	4.7	1.6	V	4.9	2.5	4.5	1.7	VG
5.0	3.4	1.5	0.2	S	4.9	2.4	3.3	1.0	V	7.3	2.9	6.3	1.8	VG
4.4	2.9	1.4	0.2	S	6.6	2.9	4.6	1.3	V	6.7	2.5	5.8	1.8	VG
4.9	3.1	1.5	0.1	S	5.2	2.7	3.9	1.4	V	7.2	3.6	6.1	2.5	VG
5.4	3.7	1.5	0.2	S	5.0	2.0	3.5	1.0	V	6.5	3.2	5.1	2.0	VG
4.8	3.4	1.6	0.2	S	5.9	3.0	4.2	1.5	V	6.4	2.7	5.3	1.9	VG
4.8	3.0	1.4	0.1	S	6.0	2.2	4.0	1.0	V	6.8	3.0	5.5	2.1	VG
4.3	3.0	1.1	0.1	S	6.1	2.9	4.7	1.4	V	5.7	2.5	5.0	2.0	VG
5.8	4.0	1.2	0.2	S	5.6	2.9	3.6	1.3	V	5.8	2.8	5.1	2.4	VG
5.7	4.4	1.5	0.4	S	6.7	3.1	4.4	1.4	V	6.4	3.2	5.3	2.3	VG
5.4	3.9	1.3	0.4	S	5.6	3.0	4.5	1.5	V	6.5	3.0	5.5	1.8	VG
5.1	3.5	1.4	0.3	S	5.8	2.7	4.1	1.0	V	7.7	3.8	6.7	2.2	VG
5.7	3.8	1.7	0.3	S	6.2	2.2	4.5	1.5	V	7.7	2.6	6.9	2.3	VG
5.1	3.8	1.5	0.3	S	5.6	2.5	3.9	1.1	V	6.0	2.2	5.0	1.5	VG
5.4	3.4	1.7	0.2	S	5.9	3.2	4.8	1.8	V	6.9	3.2	5.7	2.3	VG
5.1	3.7	1.5	0.4	S	6.1	2.8	4.0	1.3	V	5.6	2.8	4.9	2.0	VG
5.0	3.0	1.6	0.2	S	6.6	3.0	4.4	1.4	V	7.2	3.2	6.0	1.8	VG
5.0	3.4	1.6	0.4	S	6.8	2.8	4.8	1.4	V	6.2	2.8	4.8	1.8	VG

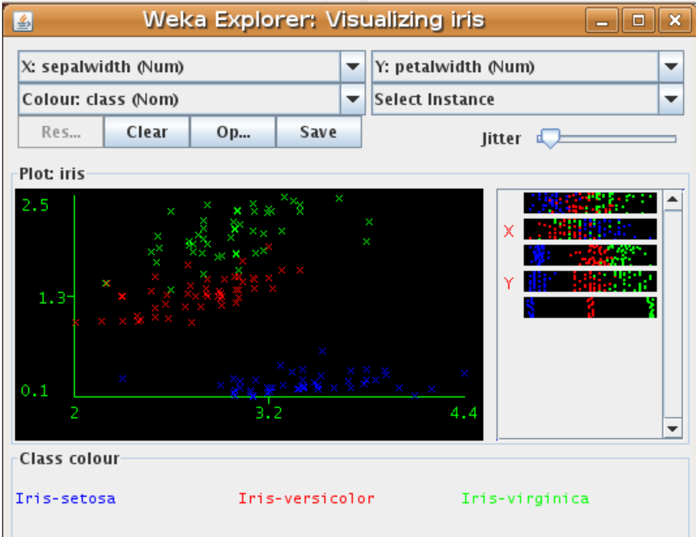
Iris classification - Visualisation

sepal width / sepal length - not very good separation



Iris classification - Visualisation 2

sepal width / petal width - quite good separation



How to achieve supervised machine learning?

There are many known successful approaches/algorithms:

- Instance-based (Nearest Neighbours - kNN)
- Rule-based
- Decision trees
- Bayesian
- Linear Regression
- (Feed-forward) Neural Networks (Perceptron, one-layer, multi-layer)
- Recursive NN
- Kernel-based approaches (e.g. Support Vector Machines)
- other...

Examples of Applications of Classification

- Hand-written digit recognition (classify each written symbol to one of 10 digits)
- e-mail spam detection (classify each e-mail to spam or normal)
- automatic language recognition in text documents (English, Polish, German, etc...)
- Web document topic classification
- face recognition
- oil leakage detection based on satellite images
- search engine spam detection (spam, normal, borderline)

Regression

When the decision attribute (to be predicted) is nominal we call it classification.
When it is numeric we call it Regression
Examples:

- predict the price of some good tomorrow based on complex economical and political context
- predict country-level energy consumption tomorrow
- predict temperature tomorrow

Example: regression

Predicting CPU performance based on its parameters

Attributes (to base the prediction on):

- MYCT cycle time (ns)
- MMIN main memory min
- MMAX main memory max
- CACH cache
- CHMIN channels min
- CHMAX channels max

Decision Attribute (to be predicted):

performance (real number)

Example: regression

MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	performance
125	256	6000	256	16	128	199
29	8000	32000	32	8	32	253
29	8000	16000	32	8	16	132
26	8000	32000	64	8	32	290
23	16000	32000	64	16	32	381
23	16000	32000	64	16	32	381
23	16000	64000	64	16	32	749
23	32000	64000	128	32	64	1238
400	1000	3000	0	1	2	23
400	512	3500	4	1	6	24
60	2000	8000	65	1	8	70
50	4000	16000	65	1	8	117
167	524	2000	8	4	15	23
143	512	5000	0	7	32	29
143	1000	2000	0	5	16	22
110	5000	5000	142	8	64	124
143	1500	6300	0	5	32	35
143	3100	6200	0	5	20	39
143	2300	6200	0	6	64	40

Unsupervised Learning

There is no distinguished decision attribute for which we want to know the “correct answer”

The system has to automatically “learn” or “discover” some patterns and dependencies in the data

Main unsupervised learning tasks:

- clustering (group cases into “similar” categories)
- association rule mining (knowledge discovery)
- outlier detection (identify atypical/strange cases)

Clustering

Automatically detect groups of cases so that

- cases in the same group are “similar”
- cases in different groups are “dissimilar”

Some notion of similarity must be provided to the system prior to clustering process

Often, an important pre-processing step in data analysis. (Afterwards, different approaches could be applied to different clusters)

Recently, importance of the clustering increases with the increasing amounts of data (no resources to tag the data with “correct” answers)

Outlier detection

Automatically detect cases which are **atypical** in the whole population based on the values of attributes

Applications:

- hack detection in large computer networks
- fraud detection in e-commerce
- erroneous data detection
- data cleaning
- early epidemy detection
- etc.

Recap

- Motivation for data mining
- decision tables, attributes
- Supervised and unsupervised learning
- Classification vs regression
- Approaches to supervised learning (only names)
- Examples of applications of supervised/unsupervised learning
- Main tasks of unsupervised learning

Questions/Problems:

- list (other than machine learning) aspects of AI
- what is the motivation for data mining
- what is a Decision Table
- what is Supervised Learning
- classification vs regression
- list 4+ approaches to supervised learning

Thank you for attention