Introduction to
Data Science and Machine Learning

Tias Guns
VUB Brussels, Belgium
Data Science in Practice

Introduction

Hendrik Blockeel, Luc De Raedt

KU Leuven, Belgium
What is Data Science?

“The next new hype”?

Data Analytics…

Data Mining…

Statistics…

Predictive Analytics…

Big Data…

Machine learning…

Data Science
What is Data Science?

No single definition

Components:

- Data-driven (the more the better)
- Interdisciplinary (math, stat, CS, ...)
- Extract knowledge from observed data
Success stories 1/3

automatic image captioning

"man in black shirt is playing guitar."
"construction worker in orange safety vest is working on road."
"two young girls are playing with lego toy."
"girl in pink dress is jumping in air."
"black and white dog jumps over bar."
"young girl in pink shirt is swinging on swing."

Automatic Image Caption Generation
Sample taken from Andrej Karpathy, Li Fei-Fei
Success stories 2/3
product recommendation
Success stories 3/3

spam detection

[Example by Paolo Frasconi]
Common Data Mining Tasks

- Classification and class probability estimation
  - How likely is this consumer to respond to our campaign?
    CSP: Is this instance satisfiable or not?
- Regression
  - How much will she use the service?
    Alg. Sel: How long will the solver run to solve the instance?
- Similarity Matching
  - Can we find consumers similar to my best customers?
    Alg. Sel: What instances were similar to this one (& which solver was best)?
- Clustering
  - Do my customers form natural groups?
    Alg. Sel: What instances form natural subgroups (can configure separately)
# Common Data Mining Tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Supervised Methods</th>
<th>Unsupervised Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>*Classification</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>*Regression</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Causal Modeling</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Similarity Matching</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Link Prediction</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Data Reduction</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>*Clustering</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Co-occurrence Grouping</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Profiling</td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
### Terminology

This is one row (example).

Feature vector is: `<**Claudio**,115000,40,**no**>`

Class label (value of Target attribute) is **no**
Books

Most books: algorithmic or statistical focus

Focus on general principles

“In ten years’ time, technologies will likely have changed such that today’s choices seem quaint.”

...  

“general principles same for 20 years”
Principle 1:  
Data Science is a process
From collection to utilization

- Data collection
- Data storage
- Data retrieval
- Data analysis
Chronology

Collection
- 1950-: manual entry
- 1970-: ad-hoc databases
- 1990-: object-relational databases
- 2000-: automatic collection (microarrays, ...)

Storage
- 1950-: programming, query language
- 1970-: relational databases
- 1990-: information retrieval
- 2000-: social media, web, cloud

Retrieval
- 1950-: statistics
- 1970-: machine learning
- 1990-: data mining
- 2000-: collaborative filtering, social methods

Analysis
- 1950-: 
- 1970-: 
- 1990-: 
- 2000-: relational data mining, network analysis

“big data”

KU LEUVEN
CRISP-DM process
An example

Business understanding

SPAM email reduces productivity, automatically remove it
Given a text message, predict whether it is spam or not

→ text categorization, useful in general
→ we want a function from message to \{0,1\}
→ is called binary classification problem
An example

Data preparation: raw text

Modeling:
We could write a rule-based system, such as

if Title.contains("YOU HAVE WON!!!") then
return Spam

Does it work well? → evaluate
### Evaluation

#### An example

<table>
<thead>
<tr>
<th>Predicted as</th>
<th>Truth</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spam</td>
<td>150</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Legitimate</td>
<td></td>
<td>720</td>
</tr>
</tbody>
</table>

- False positives: 30
- False negatives: (720 - 30) = 690
- True positives: (150 - 30) = 120
- True negatives: (200 - 30) = 170
An example

Evaluation to Business understanding:

• How do we find good rules? Knowledge elicitation or formalization may be difficult

• How do we define good? Will depend on user?

We need a system that can adapt: self-learning
An example

- **Data understanding**: collect messages, in general and from the user, that are spam (negative) and legitimate (positive)

- **Data preparation**: bag-of-words representation

- **Modeling**: train a classifier (e.g. naïve bayes)

- **Evaluate**: on unseen emails

- **Deploy**: predict for new emails, retrain when user disagrees
Principle 2a: (Holger Hoos, yesterday)

Machine Learning is optimisation

it optimises loss functions
A formal task description

Function approximation!!!

• **Given:**
  - a space of possible instances \( X \)
  - an unknown target function \( f: X \rightarrow Y \)
  - a hypothesis space \( L \) containing functions \( X \rightarrow Y \)
  - a set of examples \( E = \{ (x, f(x)) \mid x \in X \} \)
  - a loss function \( \text{loss}(h, E) \rightarrow \mathbb{R} \)

• **Find:** \( h \in L \) that minimizes \( \text{loss}(h, E) \)

Matches many (not all) tasks
Linear Regression

Notations:

- Datapoints $X = \{x_1, x_2, \ldots, x_n\}$ $x_i \in \mathbb{R}^d$
- Labels $y = \{y_1, y_2, \ldots, y_n\}$ $y \in \mathbb{R}$
- Linear decision function $f(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$

$$f(x) = w^T x$$

- Parameter vector $w$
Linear Regression

- Goal: find a linear function $Xw$ that approximates the labels $y$.
- For a new test point $x$ the label $y$ can be estimated as $w^T x$.

Sum Squared Error

$$E = \sum_{i=1}^{n} (y_i - w^T x_i)^2 = \|y - Xw\|^2$$
**Linear Regression**

- Goal: find a linear function $\mathbf{Xw}$ that approximates the labels $\mathbf{y}$.
- For a new test point $\mathbf{x}$ the label $y$ can be estimated as $\mathbf{w}^T\mathbf{x}$.

Loss function is Sum of Squared errors (L2 norm) 
Is convex $\rightarrow$ optimise it (least squares regr.)

$$E = \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 = \|\mathbf{y} - \mathbf{Xw}\|^2$$
Deep learning

handwritten number recognition:

backpropagation: (stochastic) gradient descent
Deep learning

backpropagation: (stochastic) gradient descent
Principle 2b:

If you look too hard at a dataset, you'll find things that don't generalize to unseen data.
Principle 2b:

If you look too hard at a dataset, you'll find things that **don't generalize** to unseen data

**Overfitting**
A formal task description

Function approximation!!!

• **Given:**
  - a space of possible instances \( X \)
  - an unknown target function \( f: X \rightarrow Y \)
  - a hypothesis space \( L \) containing functions \( X \rightarrow Y \)
  - a set of examples \( E = \{ (x, f(x)) \mid x \in X \} \)
  - a loss function \( \text{loss}(h,E) \rightarrow \mathbb{R} \)

• **Find:** \( h \in L \) that minimizes \( \text{loss}(h,E) \)

**model**

“supervised”

Matches many (not all) tasks
A toy example

Example from: Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000
Linear Separation

![Graph showing salmon and sea bass with a linear decision boundary and some misclassified fishes.]

Some misclassified fishes
Too simple decision?
Overfitting

No misclassifications but... very complex decisions can overfit the training data
Reasonable Solution

The optimal tradeoff might look like this unless a tradeoff of this kind is carefully defined the learning problem is ill-posed!
Fitting Graph

![Graph showing error rate vs complexity of model with two lines: one for Training and one for Holdout.](chart.png)
Over-fitting in tree induction

![Graph showing the relationship between accuracy and tree size (nodes). The graph illustrates the performance on training data and holdout data, with a "sweet spot" indicated where the two curves intersect.]
Need for holdout evaluation

Under-fitting
- In sample evaluation is in favor or “memorizing”
- On the training data the right model would be best
- But on new data it would be bad

Good

Over-fitting
Intermezzo:

Is clause learning machine learning?
Principle 2b:

If you look too hard at a set of data, you'll find things that don't generalize to unseen data

=> Clause learning is NOT machine learning
Principle 3:
Data science needs to be evaluated in the context of operation.
Data Mining versus Use of the Model

“Supervised” modeling:

“Training” data have all values specified

New data item has some value unknown (e.g., will she leave?)
Pitfalls in DM

- Training data is not consistent with actual use
- Bad sample
- Bad features

Ex: “survivorship issues”

Lending agency wants to use ML to screen applications and accept/reject them

- data of given loans + outcome
- BAD: use this to learn a predictive model
Context of operation: opportunities

Algorithm selection: SATzilla evolution (H.Hoos)

2003-2004:
- regression of runtime per solver (training)
- select solver with lowest prediction (use)

> 2011:
- **cost-sensitive** classification of solver pairs (training)
- majority voting over all predicted winners (use)
Principle 4:
Entities that are similar on some attributes often are similar on unseen attributes.
Similarity

• Ex: clustering

• Also optimisation, e.g. min. distances to cluster center
Similarity

Key concept: distance between objects

Euclidean, manhattan, edit distances (strings), dynamic time warping (temporal sequences), ...

Ex. algorithm configuration (ISAC, Kadioglu et al): cluster instances, do configuration for each cluster
Principle 5:

To draw causal conclusions,

one must pay very close attention to the presence of (possibly unseen) confounding factors.
Causality?

Machine models exploit correlation, NOT causality

Very tempting to inspect model and see “what causes things to be true/false”
Causality?

Machine models exploit correlation, NOT causality

Very tempting to inspect model and see “what causes things to be true/false”

E.g. coefficients of linear regression

\[ Y = 20X_1 - 12X_2 + 300X_3 + 99X_4 - 299X_5 \]

Which feature has most impact?
Principles

1. Data Science is a process
2. ML is optimisation of loss functions
3. ML must generalize to unseen data
4. Evaluate data science in its operational context
5. Similar entities can have similar unseen attributes
6. Correlation, not causation
Popular practical tools

- Exploratory analysis of high dim. tabular data: **tableau** (web only, not open source)
- Classification and regression on tabular data: **scikitlearn**
- Non-linear regression on large tabular data: **xdgboost**
- Deep learning on sensory data (images, audio, ...): **keras**
Connections 1/2

→ Constraint Solving for ML & DM
  - Constraint-based itemset mining
  - Constrained clustering
  - ML with hard constraints
    • structured input/output
    • side-constraints
    • with constraint solver
ML & DM for Constraint Solving

- Learning to solve better
  - Algorithm configuration / hyperparameter tuning
  - Algorithm selection
  - Dynamic search heuristics
- Learning to model better
  - Learning constraints
  - Learning the objective function
Questions?

Slides available: http://homepages.vub.ac.be/~tiasguns/ (soon)