A GUIDED TOUR OF AI

HUST-Berkeley Science Forum

Laurent El Ghaoui
UC Berkeley
Logos I relate to…
China Vows to Become an Artificial Intelligence World Leader

China launches a grand plan for AI industries and sets the goal for next dozen years

By Charlotte Gao
July 21, 2017

China is betting big on artificial intelligence (AI). On July 20, China published a new grand plan on developing its AI industries, claiming that the development of AI has been raised up to the level of
The hype

- Game Competitions
- Autonomous Flight
- Chatbots
- Humanoid Robotics
- Driverless Cars
- Even AI Citizens!
It all started 2,500 years ago...

Linear equations have been around for thousands of years. The above shows a 17th century Chinese text that explains the ancient art of “fangcheng” (rectangular arrays).

AI in the industrial age

Early names of AI:

- Automation
- Control
- Optimization
- Operations research
- Statistics
What changed?

LABELLED Data + Computing Power

i.e. a LOT of human input

and, a LOT of trial-and-error
Outline

• Tech dive:
  • Unsupervised learning
  • Supervised learning
  • Optimization and Control

• Applications
• Challenges
What is data science?

Data Science

= Machine Learning, Statistics

+ Optimization, control

(Predict, diagnose)

(Act)

Analogy: driving
Tech dive: machine learning, optimization and control

- Machine learning, statistics:
  - Unsupervised learning: represent and understand the structure of data
    E.g. clustering
  - Supervised learning: predict by learning from examples
- Optimization / Control / Decision-making
  - Based on predictions about the system, decide which actions to take
Data: labelled or not

We are good at collecting data

Some of it is labelled

By Joe Watson - December 14, 2014
There were no wolves in the movie.
0 of 3 people found this review helpful

“Cat”

But most is not labelled

?
Representing data

Data is converted to a matrix of numbers
Example: from text documents to a matrix

Sentence: *Gold drops as China tightens, down 2 percent on week.*

Dictionary: gold, silver, china, u.s., market, tightens

Numerical form of sentence: \( x = (1,0,1,0,0,1) \)

Any collection of documents can be represented in tabular form:

- A column represents a single document.
- A row represents the “score” of a particular term across documents.
- This is a VERY CRUDE representation of text (but, seems effective!)
Example: from text documents to a matrix
A matrix is a cloud of points in high dimensions

Each row represents the vote of a Senator on 650 bills

We can represent each Senator as a point in a 650-dimensional space

Each dimension represents a particular bill
Unsupervised learning: principal component analysis

PCA algorithm:
- Find direction of highest variance
- Project data orthogonal to that direction
- Repeat on projected points
- Stop until satisfactory level of cumulative variance
How to summarize a cloud?

We can “summarize” a cloud in 3D by approximating it by a line---or a plane!

In higher dimensions we can use the concept of subspace (of, say 20 dimensions)

How to choose a “good” line?
Projecting data on a line

One data point

origin

projected data point

Score of (projected) point

Score of Senators projected on random line (with party affiliation shown)
Line with maximal variance

We can choose a line so that the scores of the projected points have maximal variance (spread)

It turns out that the line agrees exactly with the party affiliation

Note that the party affiliation was not known to the algorithm!

Take-aways:
• Validates the algorithm (automatically learns the presence of two parties)
• We can rank Senators (are they extreme or more close to the other party)
Compressing to a low-dimensional subspace

We can iterate on the “maximum variance line” idea:

• Project points on a line
• Then project points on the (plane) orthogonal to line
• Find a new line of maximum variance
• Iterate k times to get a k-dimensional compression, a.k.a. “low-rank approximation”
Low-rank compression of images
Low-rank compression of other data sets

Market price time-series: 80% of the total variance in data contained in a 10-dimensional subspace.

Likewise most text data sets can be accurately approximated by very low-rank matrices.
Low-rank compression: use cases

• Extracting interesting features from high-dimensional data points
• In the low-dimensional space, algorithms run better:

Clustering / Outlier detection / Similarity between data points / etc..

Advanced versions:
• Auto-encoders (known as “word embeddings” when applied to text)
• Robust PCA
• Sparse PCA
• …
Beyond PCA: learning network structure

Correlation graph: All assets are correlated

Conditional independence graph: Discovers structure

Source: Interest rate data for various financial instruments having different maturities.
Supervised learning

Goal: given data points AND labels or numbers associated with them, learn a prediction rule that allows to assign a label or number to a new (test) point.

Methods:
• Linear regression: least-squares, logistic regression, …
• Binary / multi-class classification: SVM, logistic regression, …
• Nonlinear models: neural networks
Supervised learning: least-squares

Procedure:
• Fit a linear function through data $P_i = (x_i, y_i), i=1,…,m$
• For a new point $x$, set prediction $y$ according to what the line says
Supervised learning: binary classification

Procedure:
• Fit a (hyper-) plane that is “as far as possible from the two clouds”
• For a new point $x$, set prediction label according to which side the point falls

In binary classification each data point comes with a (binary) label (color)

Goal is to be able to predict the label of a new point

Predicted label for new data point depends on which side of the line it falls
Supervised learning: neural networks

Classical least-squares:

$$\min_{w} \| X \ast w - y \|^2_2$$

Two-layer neural network:

$$\min_{W_1, W_2} \mathcal{L}(F(W_1 \ast F(W_2 \ast X)) - Y)$$

- L is a loss function (depends on the task)
- W1, W2 are (matrix) weights
- X is “input” data, Y is “output”

- Can be extended to many layers
- Training can be difficult (time-consuming, fail to converge, etc)
- Works well with LOTS of data
Optimization and control

Measure → Predict → Control
Linear program:

\[
\min_{x} c^T x \quad \text{subject to} \quad Ax \geq b
\]

• \(x\) is a vector of “decision variables”
• Constraints are linear on \(x\)

LPs and variants can be used to describe many decision problems, e.g. energy management or optimal design of engineering systems.
Application: text analytics

70% of information is TEXT

What the computer sees

Real time information extraction

- Topics and Subtopics
- Summarization
- Trends, Sentiment and Consensus
- Outlier detection
- Streaming analysis
- Multilingual analysis

Hello 你好  Здравствуйте

مرحبا твуйте
Just counting: good but not enough

<table>
<thead>
<tr>
<th>Nicholas Kristof</th>
<th>Roger Cohen</th>
</tr>
</thead>
<tbody>
<tr>
<td>mr</td>
<td>obama</td>
</tr>
<tr>
<td>people</td>
<td>iran</td>
</tr>
<tr>
<td>obama</td>
<td>said</td>
</tr>
<tr>
<td>said</td>
<td>american</td>
</tr>
<tr>
<td>president</td>
<td>president</td>
</tr>
<tr>
<td>world</td>
<td>iranian</td>
</tr>
<tr>
<td>new</td>
<td>israel</td>
</tr>
<tr>
<td>american</td>
<td>states</td>
</tr>
<tr>
<td>years</td>
<td>new</td>
</tr>
<tr>
<td>united</td>
<td>united</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nicholas Kristof</th>
<th>Roger Cohen</th>
</tr>
</thead>
<tbody>
<tr>
<td>videos</td>
<td>olmert</td>
</tr>
<tr>
<td>darfur</td>
<td>persian</td>
</tr>
<tr>
<td>antibiotics</td>
<td>chemical</td>
</tr>
<tr>
<td>facebook</td>
<td>mohammad</td>
</tr>
<tr>
<td>sudanese</td>
<td>ali</td>
</tr>
<tr>
<td>janjaweed</td>
<td>dialogue</td>
</tr>
<tr>
<td>youtube</td>
<td>cease</td>
</tr>
<tr>
<td>sudan</td>
<td>iranian</td>
</tr>
<tr>
<td>sweatshops</td>
<td>tehran</td>
</tr>
<tr>
<td>invite</td>
<td>holocaust</td>
</tr>
</tbody>
</table>

Co-occurrence:  uses only “positive” samples

Statistical method:  uses all samples

``Climate change’’ in The New York Times

Summarize

Analyze a time slice

Frequency of term

Study a topic vs. time

Back to data

``Climate change'' in China’s People Daily

Source: 90,000 news articles from China’s People Daily, English version, 2000-2012.
``Climate change”': findings

In the NYT:
• Recognition of climate change by international organizations like the UN.
• Links to science terms points to tangible effects noted in scientific journals, and impels increased coverage.

In the PD:
• Discussion of green partnerships with Africa indicates China's multilateral approach to foreign relations.
Text analytics for safety

ASRS data:

A collection of ~25K reports on flight safety written by commercial pilots in the US, maintained by NASA.

Goals:

• Understand and diagnose issues.
• If possible, predict incidents.
Unsupervised learning: sparse PCA of ASRS data

Highest variance directions correspond to four main pilot tasks:
- Navigate (fly)
- Aviate (on runway)
- Communicate (with tower)
- Manage

Communications / runway issues predominant in big airports
Goal: Analyze the relevant features in classifying reports from one airport against all others

- At DFW we find the terms “Rwy36R” and “TwxwF”.
- This corresponds to an intersection with lots of near-miss collisions, due to lack of visibility from Tower.
Other societal-scale applications

- Energy
- Public Health
- Security
- Financial System
- Retail

More Data -> Pattern Detection -> Optimization
Optimal pricing for retail

Optimizing price based on uncertain product demand:

- A large online retailer wishes to price millions of items, based on estimated demand, in real-time.
- Goal: maximize revenue under profit margin (profit/revenue ratio) bound, inventory and price constraints.
- Challenge: demand estimates are noisy.

Results:
- Custom algorithm 100 times faster than current one.
- Enables scaling up to billions of items (ie, allows bundles).
Energy production

Optimize production based on demand prediction for next day

Optimizing based on uncertain demand:
- Energy demand follows patterns, some predictable some not
- Output often does not match demand
- Energy costs can be greatly reduced via AI

Predict energy mismatch
Case study: combined heat and power (CHP) plant

Basic problem: adjust 24-hour production variables so as to minimize operational costs, under operational and demand constraints, with demand not exactly known in advance.

Such plants are currently driven manually... Can we do driverless energy production?

CHP generation:
• Cheap
• Environmentally friendly
Driverless energy: results

- State-of-the-art optimization methods reduce production costs by 4% in average.
- More sophisticated methods enable up to an average 12% cost reduction.
Case study: predictive maintenance

Using human knowledge:

- In the future, “Internet-of-Things” (IoT) networks will capture sensor data from machinery across production systems
- We can capture now the vast human knowledge in technician or safety engineers
- Allows to diagnose failure trends, predict events

Real-time collection of human-generated maintenance / safety reports
Case study: health, government and insurance

Analyse Reports from Medical Visits → Identify What Illnesses are Occurring → Rapidly Locate Affected Areas

Early detection of potential issues
Case study: health, government and insurance

- **Government**: Coordinate Fast and Targeted Counter-Measures
- **Public Health**: Increase Specialized Healthcare Staffing & Medical Supplies
- **Insurance**: Anticipate Surge in Claims & Hedge Economic Impact
We will get more and more data

And AI can help us work with this data

AI is a powerful tool to help its users

- Like any tool, there are dangers
- But also an opportunity to improve societal-scale processes
- AI helps people become super-human
- Don’t focus on the hype! Work on novel stuff
Thank you!