Big data & Intelligence artificielle

Alexandre Gramfort
http://alexandre.gramfort.net

GitHub : @agramfort
Twitter : @agramfort
Big data & Intelligence artificielle
Quésaco?

Alexandre Gramfort
http://alexandre.gramfort.net
Objectives

• What is machine learning in 2017?
• What has changed since the 90’s? Why now?
• Data quality and data value
• What challenges do we face?
What is machine learning?
Google suggesting email reply
Facebook Face Recognition

Microsoft's Deep Learning Project Outperforms Humans In Image Recognition

Michael Thomsen, CONTRIBUTOR

I write about tech, video games, science and culture. FULL BIO ▼

Opinions expressed by Forbes Contributors are their own.
“AlphaGo seals 4-1 victory over Go grandmaster Lee Sedol” 2016
An AI Poker Bot Has Whipped the Pros

It's another seminal moment for machine learning, and a painful schooling for humans.

by Jamie Condliffe   January 31, 2017

https://www.technologyreview.com/s/603544/an-ai-poker-bot-has-whipped-the-pros/
JANUARY 25, 2017

Deep learning algorithm does as well as dermatologists in identifying skin cancer

In hopes of creating better access to medical care, Stanford researchers have trained an algorithm to diagnose skin cancer.

BY TAYLOR KUBOTA

It's scary enough making a doctor's appointment to see if a strange mole could be cancerous. Imagine, then, that you were in that situation while also living far away from the nearest doctor, unable to take time off work and unsure you had the money to cover the cost of the visit. In a scenario like this, an option to receive a diagnosis through your smartphone could be lifesaving.

Universal access to health care was on the minds of computer scientists at Stanford when they set out to create an artificially intelligent diagnosis algorithm for skin cancer. They made a database of nearly 130,000 skin disease images and trained their algorithm to visually diagnose potential cancer.
ARTIFICIAL INTELLIGENCE JOINS THE FIGHT AGAINST CARDIOVASCULAR DISEASES
What does it have in common?

**statistical machine learning**

**Definition:** Machine learning consists in teaching a computer to make decisions based on examples.
Types of input data

- **Structured:**
  - Tabular (SQL Database, CSV, Excel)

- **Unstructured:**
  - Images, videos
  - Text (comments, logs)
  - Graphs
  - Times series, signals
  - etc.
Industrial applications

- Forecasting: Sales, customer churn, traffic, price
- Predict click-through-rate (CTR) for online auctions
- Computer vision systems to monitor manufacturing (identify anomalies on products, …)
- Anomaly detection: fraud, network intrusion, predictive maintenance
- Build diagnostic systems from medical images or signals (ECG, etc.)
- etc.
Academic applications

• Learn to read people’s mind (brain computer interfaces)
• Decode the role of genes in regulation networks
• Learn to detect the Higgs boson following proton-proton collisions
• Learn to predict brain activity from image stimuli
• etc.
Predicting functional MRI activations

PhD work by M. Eickenberg at Inria
Predicting functional MRI activations

PhD work by
M. Eickenberg
at Inria
Predicting functional MRI activations

PhD work by
M. Eickenberg
at Inria
Predicting functional MRI activations

We predict brain maps without MRI scanner

PhD work by M. Eickenberg at Inria
How “big” is big data?

or what is enough data?

V.S.
Nate Silver predicts results in US presidential elections in 2012 for all 50 states

Barack Obama may have comfortably won re-election in the electoral college, and opened up a decisive lead (two million and counting) in the popular vote. But here is the absolute, undoubted winner of this election: Nate Silver and his running mate, big data.

Nate Silver predicts results in US presidential elections in 2012 for all 50 states

Barack Obama may have comfortably won re-election in the electoral college, and opened up a decisive lead (two million and counting) in the popular vote. But here is the absolute, undoubted winner of this election: Nate Silver and his running mate, big data.

The truth about Nate Silver

$ git clone gh:jseabold/538model

$ du -h 538model/data
188K  538model/data

15% of the capacity of a 3' floppy disk

https://github.com/jseabold/538model
Scaling to Very Very Large Corpora for Natural Language Disambiguation

Michele Banko and Eric Brill
Microsoft Research
1 Microsoft Way
Redmond, WA 98052 USA
{mbanko,brill}@microsoft.com

Abstract

The amount of readily available on-line text has reached hundreds of billions of words and continues to grow. Yet for most core natural language tasks, algorithms continue to be optimized, tested and compared after training on corpora consisting of only one million words or less. In this paper, we evaluate the performance of different learning methods on a prototypical natural language disambiguation task, potentially large cost of annotation for those learning methods that rely

The empirical NLP community has put a substantial effort into evaluating a large number of machine learning algorithms over fixed, and relatively small, domain-specific corpora. Since we now have access to significantly larger data sets, one has to wonder what conclusions have been drawn on small data sets and whether they extend over when these learning methods are applied in using much larger corpora.

In this paper, we present results on the effects of data size on machine learning for natural language disambiguation. In particular,

Figure 1. Learning Curves for Confusion Set Disambiguation
What is enough data?
What is enough data?

It depends and you need to try to know
What is enough data?

It depends and you need to try to know

Practical machine learning needs empirical studies
What has changed? Why now?
Some history

- Neural networks in the 60’s
- Back propagation to train neural nets in the 80’s
- Support vector machines in the 90’s
- Random forests (Breiman 2001)
Some history

- Neural networks in the 60’s
- Back propagation to train neural nets in the 80’s
- Support vector machines in the 90’s
- Random forests (Breiman 2001)

Why now?
Machine learning is now accessible to everyone
scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which set of categories a new observation belong.

Applications: Spam detection, Image recognition.

Algorithms: SVM, nearest neighbors, random forest, ...

Regression

Predicting a continuous value for a new example.

Applications: Drug response, Stock prices.

Algorithms: SVR, ridge regression, Lasso, ...

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes.

Algorithms: k-Means, spectral clustering, mean-shift, ...

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency.

Algorithms: PCA, Isomap, non-negative matrix factorization.

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning

Modules: grid search, cross validation, metrics.

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.

Modules: preprocessing, feature extraction.
In a Nutshell, scikit learn...

... has had 20,181 commits made by 650 contributors representing 179,355 lines of code

... is mostly written in Python with a well-commented source code

... has a well established, mature codebase maintained by a very large development team with stable Y-O-Y commits

... took an estimated 47 years of effort (COCOMO model) starting with its first commit in January, 2010 ending with its most recent commit 3 days ago

source: https://www.openhub.net/p/scikit-learn

Started in 2010 with colleagues at INRIA

Funding:

INRIA
Google
TELECOM LILLE
NYU CENTER FOR DATA SCIENCE
Criteo
Detecting negative comments from my customers

```bash
!head -2 train.csv
```

0,"""Imagine being able say, you know what, no sanctions, no forever hearings on IEAA regulations, no more hiding the pretense of friendly nuclear energy. You have 2 days to; i.e. let in the inspectors, quit killing the civilians, respect the border and rights of your neighboring country, or we ( whoever we are) will shut off your nuclear plant, your monitoring system and whatever else we fancy, like your water treatment plants and early warning sandstorm system and the traffic lights of all major cities...and yes..( pinky finger to lip edge) so your teenagers revolt and topple your regime... disconnect ... FACEBOOK.... buwhahjahahaha.""

0,""""But Jack from Raleigh wasn't done. He came back with this bit of furious grammatical genius:""""Holy hell, Jack. Calm down.""""GOD D@MN HILARIOUS!\nWho writes your material GraziD? MM never even acknowledged we were here (well accept when Uber ticked him off) GraziD not only interacts with us, he calls you dumb when you're being dumb... right beeaner?""""
Detecting negative comments from my customers

```python
>>> from sklearn.linear_model import LogisticRegression
>>> from sklearn.pipeline import make_pipeline, FeatureUnion
>>> from sklearn.feature_selection import SelectPercentile, chi2
>>> from sklearn.feature_extraction.text import TfidfVectorizer
>>> from sklearn.model_selection import cross_val_score

>>> # Define pipeline (text vectorizer, selection, logistic)
>>> select = SelectPercentile(score_func=chi2, percentile=16)
>>> lr = LogisticRegression(tol=1e-8, penalty='l2', C=10.,
                          intercept_scaling=1e3)
>>> char_vect = TfidfVectorizer(ngram_range=(1, 5), analyzer="char")
>>> word_vect = TfidfVectorizer(ngram_range=(1, 3), analyzer="word",
                             min_df=3)
>>> ft = FeatureUnion([("chars", char_vect), ("words", word_vect)])
>>> clf = make_pipeline(ft, select, lr)
```

11 lines of code...
>>> # run classification
>>> scores = cross_val_score(clf, X, y, cv=2)
>>> print(np.mean(scores))
0.819479193344
```python
>>> # run classification
>>> scores = cross_val_score(clf, X, y, cv=2)
>>> print(np.mean(scores))
0.819479193344
```
All models available off the shelf ... for classification

… and for clustering

Leading ML software are open-source.
THE LONG JOURNEY ACROSS TECHNO SLAVIA

Source: https://blog.dataiku.com/2015/11/19/technoslavia
software and **hardware** revolution

*Google Tensor Processor Unit (TPU)*
Hardware revolution

Example: Google Tensor Processor Unit (TPU)

Hardware revolution too too

Apple working on AI chip to keep pace with Google and Amazon

May 26, 2017, 2:50pm PDT

INDUSTRIES & TAGS Technology, Venture Capital

Evolution of data-human relationship
Evolution of data-human relationship
There is an ocean of data but most of it is undrinkable...
Data quality & Data value
Google Voice data duration system

New voicemail from (617) 427-3523 at 9:34 AM

Google Voice <voice-noreply@gmail.com>

to me

Categorize this message as: Updates

Never show this again

Voicemail from: (617) 427-3523 at 9:34 AM

Sorry dear / dialed the wrong number.

Play message

Click here to Reply or Forward
Data is not static
On data is the new oil...

but “more data” isn’t always “more value”

Model performance vs. sample size
(actual production system)

Source: https://technocalifornia.blogspot.fr/2012/07/more-data-or-better-models.html
A bit of data is usually enough to know if it can work
Epilogue
Machine Learning Taxonomy

Machine Learning
Machine Learning Taxonomy

- Machine Learning
  - Unsupervised learning
    - Clustering
    - Dimensionality Reduction
    - Anomaly Detection
Machine Learning Taxonomy

Unsupervised learning
- Clustering
- Dimensionality Reduction
- Anomaly Detection

Supervised learning
- Regression
- Classification
- Ranking
(some) Books
Some hard challenges remain
Some hard challenges remain

- The **stationarity** assumption: Assumption is that the distribution of training data match the distribution of test data
Some hard challenges remain

- The **stationarity** assumption: Assumption is that the distribution of training data match the distribution of test data

- The problem of **interpretability**: The more we measure the less we understand
Some hard challenges remain

• The **stationarity** assumption: Assumption is that the distribution of training data match the distribution of test data

• The problem of **interpretability**: The more we measure the less we understand

• How to **learn from small data**? How to **transfer** knowledge?
Some hard challenges remain

- The **stationarity** assumption: Assumption is that the distribution of training data match the distribution of test data.

- The problem of **interpretability**: The more we measure the less we understand.

- How to **learn from small data**? How to **transfer** knowledge?

- **Privacy**: How to control the access to data?
Some hard challenges remain

- The **stationarity** assumption: Assumption is that the distribution of training data match the distribution of test data.

- The problem of **interpretability**: The more we measure the less we understand.

- How to **learn from small data**? How to **transfer** knowledge?

- **Privacy**: How to control the access to data?

- **Veracity**: How to guarantee truthfulness? Can we fool ML systems?
Some hard challenges remain

- The **stationarity** assumption: Assumption is that the distribution of training data match the distribution of test data.
- The problem of **interpretability**: The more we measure the less we understand.
- How to **learn** from small data? How to **transfer** knowledge?
- **Privacy**: How to control the access to data?
- **Veracity**: How to guarantee truthfulness? Can we fool ML systems?
- Sometimes **model is wrong** and **data is partial** (think recent elections US & UK).
THMs*

*Take Home Messages not Theorems!
THMs*

• You can achieve great results with small data

*Take Home Messages not Theorems!
THMs*

• You can achieve great results with small data
• Your need to “nurture” your database to increase data quality

*Take Home Messages not Theorems!
THMs*

- You can achieve great results with small data
- Your need to “nurture” your database to increase data quality
- Open-source technology is leading the ML revolution

*Take Home Messages not Theorems!
THMs* 

- You can achieve great results with small data
- Your need to “nurture” your database to increase data quality
- Open-source technology is leading the ML revolution
- The best way to know if data science can help you is to try (small first)

*Take Home Messages not Theorems!