Why data-driven?

Hongxin Zhang
zhx@cad.zju.edu.cn

State Key Lab of CAD&CG, ZJU
2015-03-10
Outline

- Background
- What is data-driven about?
- Is it really useful for computer science and technology?
The largest challenge of Today’s CS

- Big Data

- Big companies are collecting data!!!
  - Google, Apple, Facebook, IBM, Microsoft, Amazon, ...

- In china, Baidu, Alibaba, Tecent, Sina
The largest challenge of Today’s CS

- Data, Data, Data …
  - The tedious effort required to create digital worlds and digital life.
    - Finding new ways to communicate and new kinds of media to create.
    - Experts are expensive: scientists, engineers, filmmakers, graphic designers, fine artists, and game designers.

- Process existing data and then create new ones from them.
Computers are really fast

• If you can create it, you can render it
How do you create it?

Digital Michaelangelo Project

Steven Schkolne
Pure procedural synthesis vs. Pure data

- Creating motions for a character in a movie
  - Pure procedural synthesis.
    - compact, but very artificial, rarely used in practice.
  - “By hand” or “pure data”.
    - higher quality but lower flexibility.

- the best of both worlds: hybrid methods?!
Everything but Avatar
Bayesian Reasoning

- Principle modeling of uncertainty.
- General purpose models for unstructured data.
- Effective algorithm for data fitting and analysis under uncertainty.

- But currently it is always used as a black box.

Belief v.s. Probability
Data driven modeling
Data-driven vocabulary

- Data
  - data-driven, data mining
- Learning
  - machine learning, statistical learning
- Uncertainty
  - probability, likelihood
- Intelligent
  - Inference, decision, detection, recognition
Data-driven related techniques

- Artificial Intelligence
- Machine Learning
- Data mining (KDD)
- Control and information Theory
- Statistics and Bayesian methods
- Computer Vision
- Multi-media
- Bio-informatics
- Computer Graphics
- Information retrieval

ML ≠ AI
Data-driven system

- Learning systems are not directly programmed to solve a problem, instead develop own program based on:
  - examples of how they should behave
  - from trial-and-error experience trying to solve the problem

Different from standard CS: want to implement unknown function, only have access to sample input-output pairs (training examples)
Main categories of learning problems

Learning scenarios differ according to the available information in training examples

- **Supervised**: correct output available
  - **Classification**: 1-of-N output (speech recognition, object recognition, medical diagnosis)
  - **Regression**: real-valued output (predicting market prices, temperature)

- **Unsupervised**: no feedback, need to construct measure of good output
  - **Clustering**: Clustering refers to techniques to segmenting data into coherent “clusters.”
  - **Novelty-detection**: detecting new data points that deviate from the normal.

- **Reinforcement**: scalar feedback, possibly temporally delayed
Main class of learning problems

Learning scenarios differ according to the available information in training examples

- **Supervised**: correct output available
  - ...
- **Semi-Supervised**: only a part of output available
  - **Ranking**:
- **Unsupervised**: no feedback, need to construct measure of good output
  - ...
- **Reinforcement**: scalar feedback, possibly temporally delayed
And more …

- Time series analysis.
- Dimension reduction.
- Model selection.
- Generic methods.
- Graphical models.
Why data driven methods?

- **Develop enhanced computer systems**
  - automatically adapt to user, customize
  - often difficult to acquire necessary knowledge
  - discover patterns offline in large databases (*data mining*)

- **Improve understanding of human, biological learning**
  - computational analysis provides concrete theory, predictions
  - explosion of methods to analyze brain activity during learning

- **Timing is good**
  - growing amounts of data available
  - cheap and powerful computers
  - suite of algorithms, theory already developed
Is it really useful for computer science and technology?

- **Con:** Everything is machine learning or everything is human tuning?
  - Sometimes, this may be true.

- **Pro:** more understanding of learning, but yields much more powerful and effective algorithms.
  - Problem taxonomy.
  - General-purpose models.
  - Reasoning with probabilities.

- I believe the mathematic magic.
What will be a successful D-D algorithm?

- Computational efficiency
- Robustness
- Statistical stability
The First Example: Google!

- 每天过滤 200 亿个网页
- 每天追踪 300 亿个的独立 URL
- 每月接受 1000 亿次搜索请求
Object detection and recognition - the power of DD

The image is copied from
http://vismod.media.mit.edu/vismod/demos/facerec/
Object detection and recognition

Face [Vaillant et al IEE 1994] [Garcia et al PAMI 2005] [Osadchy et al JMLR 2007]
Pedestrian: [Kavukcuoglu et al. NIPS 2010] [Sermanet et al. CVPR 2013]
Speech recognition

Modern architecture for pattern recognition

Speech recognition: early 90's – 2011

- MFCC (fixed)
- Mix of Gaussians (unsupervised)
- Classifier (supervised)

Object Recognition: 2006 - 2012

- SIFT, HoG (fixed)
- K-means, Sparse Coding (unsupervised)
- Pooling
- Classifier (supervised)

Low-level Features
Mid-level Features
Speech recognition

- Hierarchy of representations with increasing level of abstraction
- Each stage is a kind of trainable feature transform
- Image recognition
  - Pixel → edge → texton → motif → part → object
- Text
  - Character → word → word group → clause → sentence → story
- Speech
  - Sample → spectral band → sound → ... → phone → phoneme → word →

Diagram:

```
Trainable Feature Transform
Trainable Feature Transform
Trainable Feature Transform
Trainable Feature Transform
```
Director Ang Lee Takes Risks with Mean Green 'Hulk'

LOS ANGELES (Reuters) - Taiwan-born director Ang Lee, perhaps best known for his Oscar-winning "Crouching Tiger, Hidden Dragon," is taking a big risk with the splashy summer popcorn flick......

FAMILY DRAMA, BIG ACTION

For loyal comic book fans who may think Lee's "Hulk" will be too touchy-feely, think again. "This is a drama, a family drama," said Lee, "but with big action." His slumping shoulders twitch and he laughs......
Mesh Processing – Data clustering/segmentation

- Hierarchical Mesh Decomposition using Fuzzy Clustering and Cuts.
  By Sagi Katz and Ayellet Tal, SIGGRAPH 2003
Texture synthesis and analysis – Hidden Markov Model


Reflectance texture synthesis – Dimension reduction


- More recent papers…
Human shapes - Dimension reduction


Image processing and synthesis - Graphical model


A pirouette and promenade in five synthetic styles drawn from a space that contains ballet, modern dance, and different body types. The choreography is also synthetic. Streamers show the trajectory of the left hand and foot.

- **Style Machines.** M. Brand and A. Hertzmann. SIGGRAPH 2000.

- **A Data-Driven Approach to Quantifying Natural Human Motion.** L. Ren, A. Patrick, A. Efros, J. Hodgins, J. Rehg. SIGGRAPH 2005
VideoTextures -
Reinforcement Learning

Summary

- Learning (from Data) is a nut-shell, :-)D

Keywords

- Noun: data, models, patterns, features;
- Adj.: probabilistic, statistical;
- Verb: fitting, reasoning, mining.
Homework

- Try to find potential learning based (data driven) applications in your research area
Reference

- Reinforcement learning: A survey
The End

新浪微博: @浙大张宏鑫

微信公众号: [QR Code]