Inf2b Learning and Data

Lecture 1: Introcution to Learning and Data

Hiroshi Shimodaira
(Credit: Iain Murray and Steve Renals)

Centre for Speech Technology Research (CSTR)
School of Informatics
University of Edinburgh

Jan-Mar 2014

Welcome to Inf2b!

Today's Schedule:

1. Course structure
2. What is (machine) learning? (and why should you care?)
3. Administrative stuff
   - How to do well
4. Setting up a learning problem

(time allowing)

Course structure

Website:
http://www.inf.ed.ac.uk/teaching/courses/inf2b/

Constituents:

- 30 lectures (including review)
- Tutorials starting in week 2
- 2 assessed assignments

Equal split into two threads:

- Algorithms and Data Structures – KK (Kyriakos Kalorkoti)
- Learning and Data – Hiroshi

1. Course structure
2. What is machine learning
3. Administrative stuff
4. Setting up a learning problem

Face detection

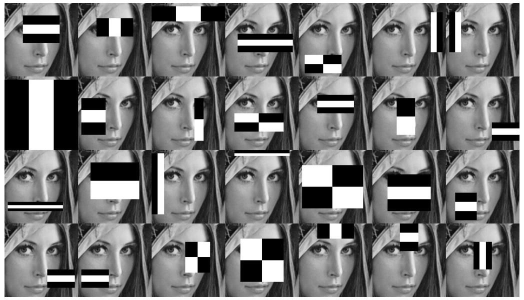
How would you detect a face?

(R. Vaillant, C. Monrocq and Y. LeCun, 1994)

http://demo.pittpatt.com/

How does album software tag your friends?

Viola–Jones Face detection (2001)



Taken from: http://ahprojects.com/projects/cv-dazzle
A nice demo: http://vimeo.com/12774628

A neat algorithm & data structure

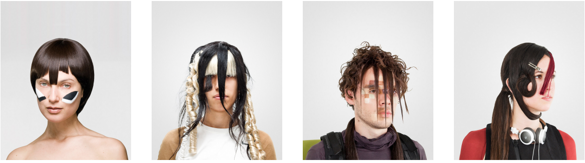
Rectangle intensity: naively need to add 10^3 to 10^6 pixels

Pre-computation: *Integral Image*,

add/subtract 4 values ⇒ rectangle intensity

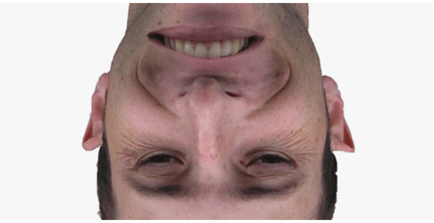
http://en.wikipedia.org/wiki/Summed_area_table

Hiding from the machines



Taken from: http://ahprojects.com/projects/cv-dazzle

How does human vision work?



Intro summary

- Fit numbers in a program to data
- More robust than hand-fitted rules
- Can't approach humans at some tasks (e.g., vision)
- Machines make better predictions in many other cases

Applications of machine learning

Within informatics:

- **Vision:** as we've seen
- **Graphics:** increasingly data driven
- **Natural Language Processing (NLP):** text search/summarisation, speech recognition/synthesis
- **Robotics:** vision, planning, control, . . .
- **Compilers:** learning how to optimise

and beyond: data analysis across the sciences

Every day:

- Adverts / recommendations all over the web · · · Big Data
- Discounts in Tescos http://www.mathworks.co.uk/discovery/big-data-matlab.html
- Speech recognition, Machine Translation, . . . with self-driving cars 'soon'?

1. Course structure
2. What is machine learning
3. Administrative stuff
4. Setting up a learning problem

Private study

∼2 hours private study per lecture,
in addition to tutorials & assignments!

No required textbook for Inf2b
There are notes. See those for recommended books.

Come to lectures! (really, skipping lectures is a *bad* idea)

Class reps

Wanted: Inf2b class reps (for ADS and & learning)

Email: h.shimodaira@ed.ac.uk
your name, degree, email address.

Two hours study this week?

Start to familiarise yourself with Matlab (or Octave)
Introductory worksheet on the course website
Many others at the end of a web search

Love Python? Learn NumPy+SciPy+Matplotlib
(instead, or as well)

Vital skills:

- add, average, multiply vectors and matrices
- plot data stored in vectors
- save/read data to/from files

1. Course structure
2. What is machine learning
3. Administrative stuff
4. Setting up a learning problem

The Netflix Prize

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

"We're quite curious, really. To the tune of one million dollars.

It's "easy" really. We provide you with a lot of anonymous rating data, and a prediction accuracy bar that is 10% better than what Cinematch can do on the same training data set."

http://www.netflixprize.com, October 2006.

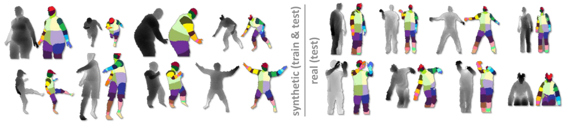
Kaggle

Crowd-sourcing data-science solutions:
http://kaggle.com/

Creating training data

Microsoft Kinect (Shotton et al., CVPR 2011)

http://research.microsoft.com/apps/pubs/default.aspx?id=145347



Random forest applied to fantasies

Summary of problem setting-up

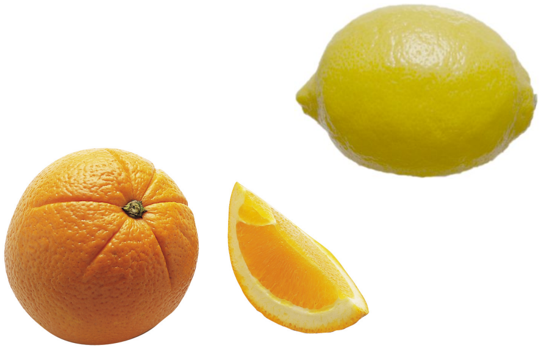
Each challenge has:

- A measure of success Objective function, cost function, metric, . . .
- Data is useful (but needs to be available)
- Nothing is certain

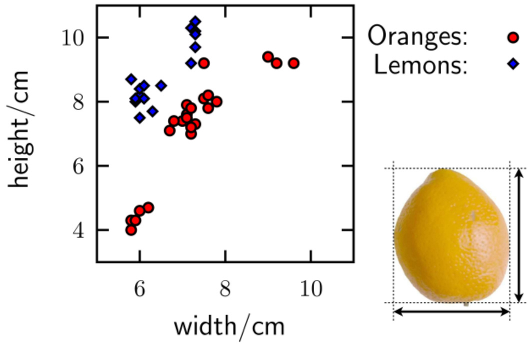
we will use probability a lot

How does a machine use the data?

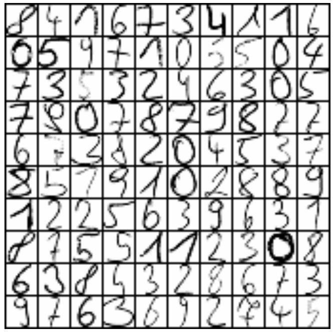
Oranges and Lemons



A two-dimensional space

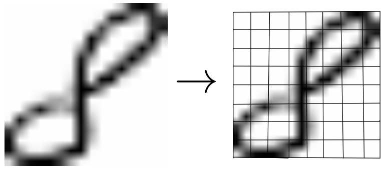


Handwritten digits



http://alex.seewald.at/digits/

A 64-dimensional space



Turn each cell into a number (somehow, see notes)
Unravel into a column vector, a **feature vector**
⇒ represented digit as point in $64D$

http://alex.seewald.at/digits/

Euclidean distance

Distance between $2D$ vectors: (x, y) and (x', y')

$$r_2 = \sqrt{(x - x')^2 + (y - y')^2}$$

Distance between D-dimensional vectors: $\mathbf{x}$ and $\mathbf{x}'$

$$r_2(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{d=1}^{D} (x_d - x'_d)^2}$$

Measures similarities between feature vectors
i.e., similarities between digits, movies, sounds, galaxies, . . .

Question

Have high-resolution scans of digits.

How many pixels should be sample?

What are pros and cons of:

2×2, 4×4, 16×16, or 100×100?