Introduction to Learning and Data

What is (machine) learning?

Administrative stuff

Welcome to Inf2b!

Inf2b: Algorithms, Data Structures (ADS), and Learning

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

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

Constituents:

- 30 lectures (including review) - Tuesdays, Thursdays, Fridays
- Tutorials (starting in week 2)
- Drop-in labs for Learning (Fridays 14:10 - 15:00)
- 2 assessed assignments (with drop-in labs)

CW2 (Learning): 13/Mar. – 04/Apr.

Equal split into two threads:

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

Building models that describe a data set and can make predictions about new data

Drop-in labs for Learning

Fridays 14:10-15:00 in AT-5.05 (West Lab)

Starting in Week 1 (on Friday 19th January)

Worksheets available on the course webpage

- Practice on machine learning using Matlab
- Work on toy problems for the topics taught in the course
- Demonstrator: Andreas Kapourain

Face detection

How would you detect a face?

How does album software tag your friends?

Viola–Jones Face detection (2001)

Face detector consists of linear combination of ‘weak’ classifiers that utilise five types of primitive features.

- The detector is trained on a training data set of a large number of positive and negative samples.
- Scan the input image with a sub-window (24 x 24 pixels) to detect a face.

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

A nice demo: http://vimeo.com/12774628

Within informatics:

- Vision: as we’ve seen. [another e.g.]
- Graphics: increasingly data driven
- AI & Natural Language Processing (NLP): text search/summarisation, speech recognition/synthesis, e.g. IBM Watson
- 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 Tesco’s http://www.tescos.co.uk/discovery/big-data-tesco.html
- Speech recognition and synthesis (e.g. Siri, Echo), Machine Translation, … with self-driving cars

Fit numbers in a program to data (i.e. train machines on 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

Hiding from the machines (cameras)

The Viola-Jones face detector is fast, but has some drawbacks.

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

Useful webpage to check your maths:
http://www.mathsisfun.com/algebra

- Laws of exponents (Exponent rules)
  \[ x^a x^b = x^{a+b}, \quad (x^a)^b = x^{ab} \]
- Log and exponential
  \[ \log(x y^m) = \log x + m \log y, \quad e^{\ln x} = x \]
- Quadratic equations and their solutions
  \[ ax^2 + bx + c = 0, \quad x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \]
- Vectors \[ v = (v_1, v_2, \ldots, v_n)^T \]
  - Notation: column/row vectors, transpose
  - Addition and subtraction \[ v + w \]
  - Dot product (inner product) \[ u \cdot v = u^T v \]
  - Equation of a straight line, linear equations

Matrices \[ A = (a_{ij}), \quad A^{-1} = a_{ij} \]
- Addition, subtraction \[ A + B, \quad A - B \]
- Multiplication \[ (AB)_{ij} = \sum_k a_{ik} b_{kj} \]
- Transpose \[ (ABC)^T = C^T B^T A^T \]
- Inverse \[ A^{-1} A = A A^{-1} = I \]
- Eigen values and eigenvectors
- Vector spaces, subspaces, linear independence, basis and dimension, rank and nullity
- Linear transformations \[ y = Ax \]

NB: See Section 4 of Learning Note No. 1 for the notation we use.

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
- Learn MATLAB try the lab sheets for the 1st lab this week.
- 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

Classroom

- Have a look at the lecture note and slides in advance to the lecture.
- Have questions prepared to ask.
- Laptops, tablets, phones are not allowed to use during lectures unless permitted.

NB: Lectures will be recorded, and videos will be published in a few days after the lecture.

Classification of oranges and lemons

A two-dimensional space

Represent each sample as a point \((w, h)\) in a 2D space

- Oranges:
- Lemons:

Photo image – pixels

Represent each sample as a point \((w, h)\) in a 2D space

- Oranges:
- Lemons:

Laptops, tablets, phones are not allowed to use during lectures unless permitted.

NB: Lectures will be recorded, and videos will be published in a few days after the lecture.

Classification of oranges and lemons

A two-dimensional space

Represent each sample as a point \((w, h)\) in a 2D space

- Oranges:
- Lemons:

Photo image – pixels

Represent each sample as a point \((w, h)\) in a 2D space

- Oranges:
- Lemons:

Laptops, tablets, phones are not allowed to use during lectures unless permitted.

NB: Lectures will be recorded, and videos will be published in a few days after the lecture.
Pixel image to a feature vector

Turn each cell (pixel) into a number (somehow, see notes)
Unravel into a column vector, a feature vector
\[ x = (x_1, x_2, \ldots, x_{64})^T \]
represented digit as point in 64D
\[ x \in [0, 127] \text{ or } x \in [0, 1] \]
http://alex.seewald.at/digits/

Image data as a point in a vector space

Distance between 2D vectors: \[ u = (u_1, u_2)^T \text{ and } v = (v_1, v_2)^T \]
\[ d_2(u, v) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2} \]
Distance between \( d \)-dimensional vectors: \[ u = (u_1, \ldots, u_d)^T \text{ and } v = (v_1, \ldots, v_d)^T \]
\[ d_2(u, v) = \sqrt{\sum_{k=1}^{d} (u_k - v_k)^2} \]
Measures similarities between feature vectors
i.e., similarities between digits, movies, sounds, galaxies, ...

Euclidean distance

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?

Example of image resolutions

Exercises in the lecture note 1

Try the exercises in the lecture note 1.
No solutions will be published.
In case you’re not sure if your answers are correct.
- Discuss them with your classmates
- Use the Inf2b-Learning discussion board on Piazza

Summary

- Self-study both ADS and Learning everyday.
- Drop-in for Learning starts on Friday, 19th Jan at 14:10.
  Try the worksheets before the lab.
- Tutorial starts in Week 2.
- Discussion forum for “Learning” in Piazza
- Office hours: Wednesdays at 14:00-15:00 (TBC) in IF-3.04

AccountabilityAcc