CS 886
Applied Machine Learning
Introduction Part 1 - Overview, Regression

Dan Lizotte

University of Waterloo

10 Sept 2014

Welcome to CS 886 (Fall 2014)

Instructor **Dan Lizotte** Office: DC3617 but use these first:

- Piazza: `piazza.com/class#fall2014/cs886`
- e-mail: dlizotte@uwaterloo.ca, 886 in subject line
  Use your UW e-mail.

Wiki Main resource for materials, requirements, etc.
`www.cs.uwaterloo.ca/~dlizotte/teaching/cs886`
READ THE WHOLE THING.

Lectures: Wednesdays and Fridays, 10:30am–11:50am, DC2568
Based on material courtesy of Prof. Doina Precup
`www.cs.mcgill.ca/~dprecup`
and **Pattern Recognition and Machine Learning**
by Chris Bishop
`research.microsoft.com/en-us/um/people/cmbishop/prml/`

Required Text: The Elements of Statistical Learning
`www-stat.stanford.edu/~tibs/ElemStatLearn/`

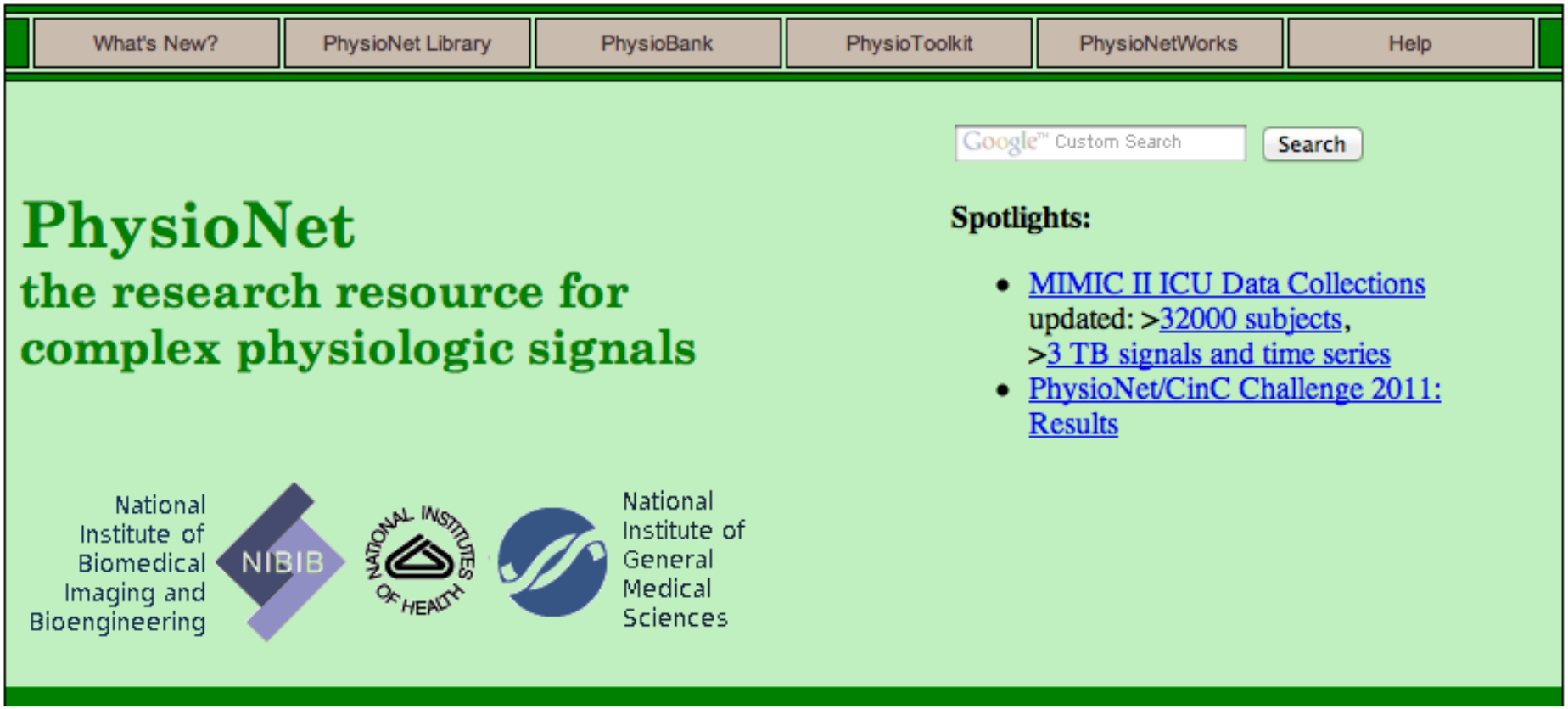
Objective

- Introduce students to machine learning techniques,
  with a focus on application to substantive (i.e. non-ML) problems.
- Gain experience in identifying
  1. which problems can be tackled by machine learning methods
  2. which specific ML methods are applicable to the problem at hand
- Students will gain an in-depth understanding of a particular (substantive problem, ML solution) pair, and present their findings.
- Evaluation: Quizzes, Project Proposal, Brainstorming Presentation, Draft, Report, Reviews

Topics

- Machine Learning:
  - Supervised learning
  - Unsupervised learning
  - Sequential decision making
- Substantive areas:
  - Astronomy
  - Cardiology
  - Criminology
  - Conservation
  - Education
  - Energy Consumption
  - History
  - Kinesiology
  - Marketing
  - Music
  - Neurology
  - ...

Data


What's New?
PhysioNet Library
PhysioBank
PhysioToolkit
PhysioNetWorks
Help
Google™ Custom Search
Search
PhysioNet
the research resource for
complex physiologic signals
Spotlights:
MIMIC II ICU Data Collections
updated: >32000 subjects,
>3 TB signals and time series
PhysioNet/CinC Challenge 2011:
Results
National
Institute of
Biomedical
Imaging and
Bioengineering
NIBIB
NATIONAL INSTITUTES OF HEALTH
National
Institute of
General
Medical
Sciences

Data

MIMIC II Database

Quick Links

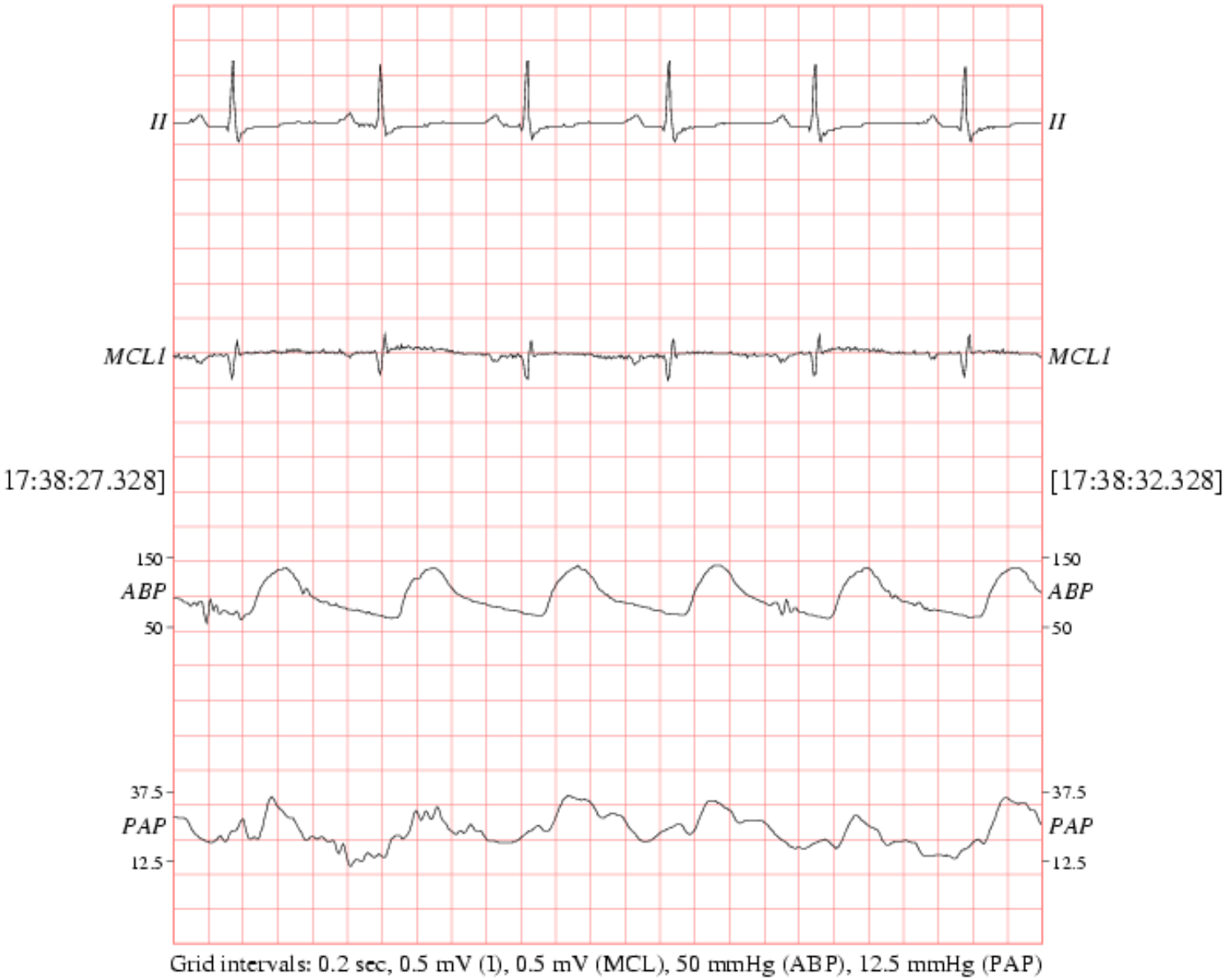
Getting Access
Waveform Database Overview
Clinical Database Overview
Waveform Data
Clinical Data

MIMIC II Statistics

MIMIC II FAQs

The MIMIC II (**M**ultiparameter **I**ntelligent **M**onitoring in **I**ntensive **C**are) Database contains comprehensive clinical data from tens of thousands of Intensive Care Unit (ICU) patients*. Data were collected between 2001 and 2008 from a variety of ICUs (medical, surgical, coronary care, and neonatal) in a single tertiary teaching hospital. The database contains clinical data from bedside workstations as well as hospital archives. The database also includes thousands of records of continuous high-resolution physiologic waveforms and minute-by-minute numeric time series (trends) of physiologic measurements. Many, but not all, of the waveform records are matched to corresponding clinical data (for more information, see Record Matching). The database is thoroughly de-identified (all PHI has been removed and all dates have been changed).

Data


II
II
MCL1
MCL1
17:38:27.328]
[17:38:32.328]
150
ABP
50
150
ABP
50
37.5
PAP
12.5
37.5
PAP
12.5
Grid intervals: 0.2 sec, 0.5 mV (I), 0.5 mV (MCL), 50 mmHg (ABP), 12.5 mmHg (PAP)

Data

"Recorded waveforms and numerics vary depending on choices made by the ICU staff. Waveforms almost always include one or more ECG signals, and often include continuous arterial blood pressure (ABP) waveforms, fingertip photoplethysmogram (PPG) signals, and respiration, with additional waveforms (up to 8 simultaneously) as available. Numerics typically include heart and respiration rates, SpO2, and systolic, mean, and diastolic blood pressure, together with others as available. Recording lengths also vary; most are a few days in duration, but some are shorter and others are several weeks long."

Data

ICU Intensive Care Unit

ECG Electrocardiogram - "...electrical activity of the heart over a period of time." MCL1 and II in the graph are ECG readings from different electrodes.

ABP Arterial Blood Pressure - (Near-)continuous measurement of pressure in the artery. PAP is same for pulmonary artery.

PPG Photoplethysmogram - "As you can see here in the photophym... in the uh, photoplethmohrp...

in the **cardiac pulse waveform**..."

What now? Find the problems people care about.

Crit Care Med. 2011 May; 39(5): 952-960. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L. Lehman, G.B. Moody, T. Heldt, T.H. Kyaw, B.E. Moody, R.G. Mark.

Crit Care Med. 2001 Feb;29(2):427-35. Artificial intelligence applications in the intensive care unit. Hanson CW 3rd, Marshall BE.

PLoS Comput Biol. 2007 Nov;3(11):e204. Epub 2007 Sep 6. From inverse problems in mathematical physiology to quantitative differential diagnoses. Zenker S, Rubin J, Clermont G.

... ...

What now?

- Back to the data to see if what you have can address the problems.
- Back to the methods to see if you can apply them to your data.
- Back to the problems to see if your output addresses them.
- ...

Seeking students who:

- Like to read - have a desire to understand substantive problems
- Like to think - make connections between methods and problems
- Like to hack - be willing to munge data into usability
- Like to speak - teach us about what you found!

ML methods knowledge an asset, but not required.

Project - Big Picture

- The project will require quite a bit of independent study of methods. Use the book, and other online resources.
- **The data must be interesting. No irises allowed.**
- My guess: Most projects will be supervised, prediction-oriented
- *A high quality project must thoroughly describe the problem and the data,* **justify and explain** *the methods used, and give a* **sound empirical evaluation** *of the results.*

Project - Big Picture

- I have a secret... ...your project might not work.
- That is okay. Prove to me and to your classmates that:
  - You thoroughly understand the substantive area and problem
  - You thoroughly understand the data
  - You know what methods are reasonable to try and why
  - You tried several and evaluated them rigorously, but your predictions are just not that good.
  - You can't get blood from a turnip. (But prove it.)

A **turnip** cannot be coaxed, squeezed, or cajoled into producing blood. All efforts at obtaining blood from this vegetable will be futile.

Project - Big Picture

- Downside to "real data": Might not work. (Probably won't work?)
- Upside is, given effort, you will gain much more relevant experience.
- Project components:
  - Proposal: Two-page document detailing the plan for the project
  - Draft: A draft of the final report will be due approximately midway through the term
  - Brainstorming Presentation: 30 minutes, after the halfway point
  - Report: ICML conference format, submitted to EasyChair
  - Reviews: Each student reads a few papers, writes reviews
- The wiki is the gold standard for project requirements.
- **Expectations:** The quality of writing in the report should be comparable to a paper in ICML, IAAI, ICMLA or another good conference. Therefore you need to **read** a few of these to get an idea of what's expected.

Quizzes

- **Daily** starting on Friday.
- **Beginning of class** so don't be late
- **Multiple choice/short answer**, just a few questions, **based on last day's material, plus readings in HTF**
- **Closed book, no cheating.** I will not hesitate to send you to academic integrity.
- **Don't panic.** They're 5% and I will drop each person's worst one.

Logistics

- First homework: Sit down and *carefully* read the wiki,
  pick brainstorming slot,
  sign up for Piazza **with your UW e-mail.**
- Data available online; if you find more, add it to the wiki
- Note: You are responsible if the data require an "agreement" for use, or if there is an application required, etc.
- You *may* use proprietary data; if so post it in the table
  (no link of course)

Outline for Unit 1

- What is machine learning?
- Types of machine learning
- Supervised learning
- Linear and polynomial regression
- Performance evaluation
- Overfitting
- Cross-validation

What is learning?

- Herbert A. Simon: Any process by which a system improves its performance
- Marvin Minsky: Learning is making useful changes in our minds
- Ryszard S. Michalski: Learning is constructing or modifying representations of what is being experienced
- Leslie Valiant: Learning is the process of knowledge acquisition in the absence of explicit programming

Any system that accomplishes its task using a combination of prior knowledge and data.

Why study machine learning?

- Easier to build a learning system than to hand-code a working program! E.g.:
  - Robot that learns a map of the environment by exploring
  - Programs that learn to play games by playing against themselves
- Discover knowledge and patterns in highly dimensional, complex data
  - *Sky surveys*
  - Sequence analysis in bioinformatics
  - Social network analysis
  - Ecosystem analysis
  - *Forest fire prediction*
  - *Power consumption prediction*
  - *Predicting hospital stay length*
  - *Characterizing muscle pathologies*
  - ...

Why study machine learning?

- Solving tasks that require a system to be adaptive, e.g.
  - Speech and handwriting recognition
  - "Intelligent" user interfaces
- Understanding animal and human learning
  - How do we learn language?
  - How do we recognize faces?
- Creating real AI!

"If an expert system–brilliantly designed, engineered and implemented–cannot learn not to repeat its mistakes, it is not as intelligent as a worm or a sea anemone or a kitten."
— Oliver Selfridge

Very brief history

- Studied ever since computers were invented (e.g. Arthur Samuel's checkers player in 1956!!)
- Very active in 1960s (neural networks)
- Died down in the 1970s
- Revival in early 1980s (decision trees, backpropagation, temporal-difference learning) - coined as "machine learning"
- Exploded starting in the 1990s
- Now: very active research field, several yearly conferences (e.g., ICML, ECML, NIPS), major journals (e.g., Machine Learning, Journal of Machine Learning Research)
- The time is right to study in the field!
  - Lots of recent progress in algorithms and theory
  - **Flood of data to be analyzed**
  - Computational power is available
  - Growing demand for industrial applications

Related disciplines

- Artificial intelligence
- Probability theory and statistics
- Computational complexity theory
- Control theory
- Information theory
- Philosophy
- Psychology and neurobiology

What are good machine learning tasks?

- There is no human expert
  E.g., predicting hospital stay length
- Humans can perform the task but cannot explain how
  E.g., character recognition
- Desired function changes frequently
  E.g., predicting stock prices based on recent trading data
- Each user needs a customized function
  E.g., news filtering

Kinds of learning

Based on the information available:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Based on the role of the learner

- Passive learning
- Active learning

Supervised learning (HTF Ch. 2)

- Training experience: a set of *labeled examples* of the form

$$\langle x_1, x_2, \ldots x_p, y \rangle,$$

  where x_j are *feature values* and y is the *output*
- Task: **Given a new** $x_1, x_2, \ldots x_p$**, predict** y
- What to learn: A *function* $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_p \to \mathcal{Y}$, which maps the features into the output domain
- Goal: minimize the error (loss function) on the *future predictions*
  Plan: minimize the error (loss function) on the training examples

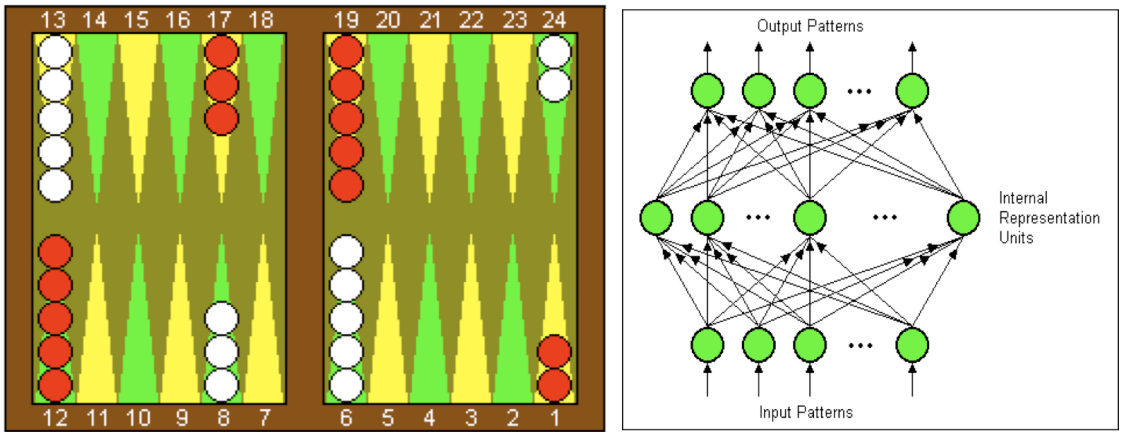
Example: Face detection and recognition

- $x_1, x_2, \ldots x_p$ are **features** that describe an image
- y could be...
  - ...$\in \{0, 1\}$ (face present/no face present)
  - ...$\in \{0, 1, 2, ...\}$ how many faces?
  - ...$\in \{\text{rectangles}\}$ where are the faces?

Reinforcement learning

- Training experience: interaction with an environment; the agent receives a numerical reward signal
- E.g., a trading agent in a market; the reward signal is the profit
- What to learn: a way of choosing actions that is very rewarding in the long run
- Goal: estimate and maximize the long-term cumulative reward

Example: TD-Gammon (Tesauro)

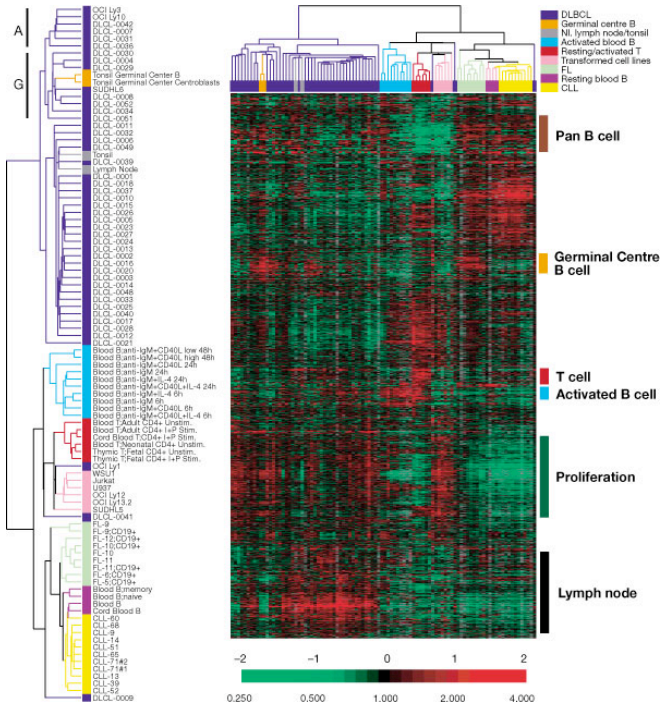


- Learning from self-play, using TD-learning
- Became the best player in the world
- Discovered new ways of opening not used by people before

Unsupervised learning

- Training experience: unlabelled data – no targets!
- What to learn: interesting associations and patterns in the data
- E.g., image segmentation, clustering
- Often there is no single correct answer. Evaluation can be troublesome.
- Can potentially be used as a pre-processing step for a supervised problem.

Example: Oncology (Alizadeh et al.)

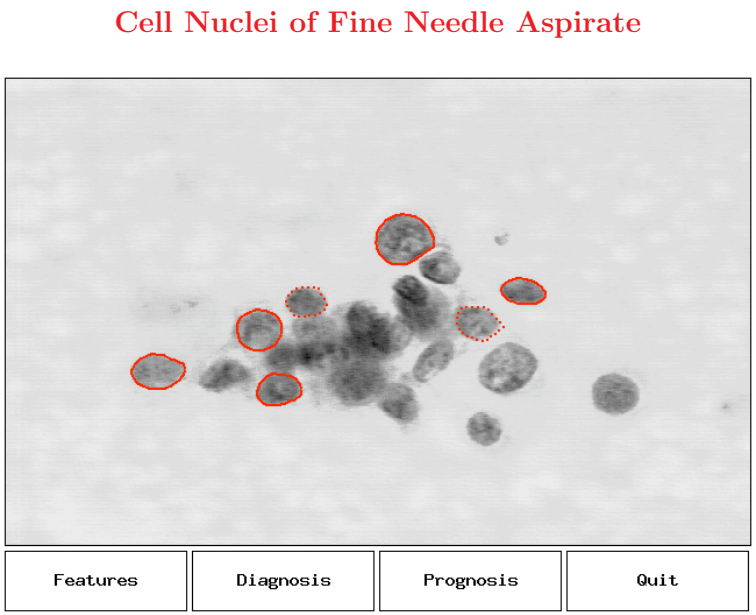


- Activity levels of all ($\approx$ 25,000) genes were measured in lymphoma patients
- Cluster analysis determined three different subtypes (where only two were known before), having different clinical outcomes

Passive and active learning

- Traditionally, learning algorithms have been *passive learners*, which take a given batch of data and process it to produce a hypothesis or model
  Data $\rightarrow$ Learner $\rightarrow$ Predictive Model
- *Active learners* are instead allowed to query the environment
  - Ask questions
  - Perform experiments
- Open issues: how to query the environment optimally? how to account for the cost of queries?

Today: Introduction to Supervised Learning



- Cell samples were taken from tumors in breast cancer patients before surgery, and imaged
- Tumors were excised
- Patients were followed to determine whether or not the cancer recurred, and how long until recurrence or disease free

Wisconsin data (continued)

- Thirty real-valued features per tumor.
- Two variables that can be predicted:
  - Outcome (R=recurrence, N=non-recurrence)
  - Time (until recurrence, for R, time healthy, for N).

tumor size	texture	perimeter	...	outcome	time
18.02	27.6	117.5		N	31
17.99	10.38	122.8		N	61
20.29	14.34	135.1		R	27
...					

Terminology

tumor size	texture	perimeter	...	outcome	time
18.02	27.6	117.5		N	31
17.99	10.38	122.8		N	61
20.29	14.34	135.1		R	27
...					

- Columns are called *input variables* or **features** or *attributes*
- The outcome and time (which we are trying to predict) are called *output variables* or *targets*
- A row in the table is called *training example* or *instance*
- The whole table is called *(training) data set*.

Prediction problems

tumor size	texture	perimeter	...	outcome	time
18.02	27.6	117.5		N	31
17.99	10.38	122.8		N	61
20.29	14.34	135.1		R	27
...					

- The problem of predicting the recurrence is called *(binary) classification*
- The problem of predicting the time is called *regression*

More formally

tumor size	texture	perimeter	...	outcome	time
18.02	27.6	117.5		N	31
17.99	10.38	122.8		N	61
20.29	14.34	135.1		R	27
...					

- A training example i has the form: $\langle x_{i,1}, \ldots x_{i,p}, y_i \rangle$ where p is the number of features (30 in our case).
- We will use the notation $\mathbf{x}_i$ to denote the row vector with elements $x_{i,1}, \ldots x_{i,p}$.
- The training set D consists of n training examples
- We denote the $n \times p$ matrix of features by X and the size-n column vector of outputs from the data set by $\mathbf{y}$.
- In statistics, X is called the data matrix or the *design matrix.*
- Let $\mathcal{X}$ denote the space of input values
- Let $\mathcal{Y}$ denote the space of output values

Supervised learning problem

- Given a data set $D \subset (\mathcal{X} \times \mathcal{Y})^n$, find a function:

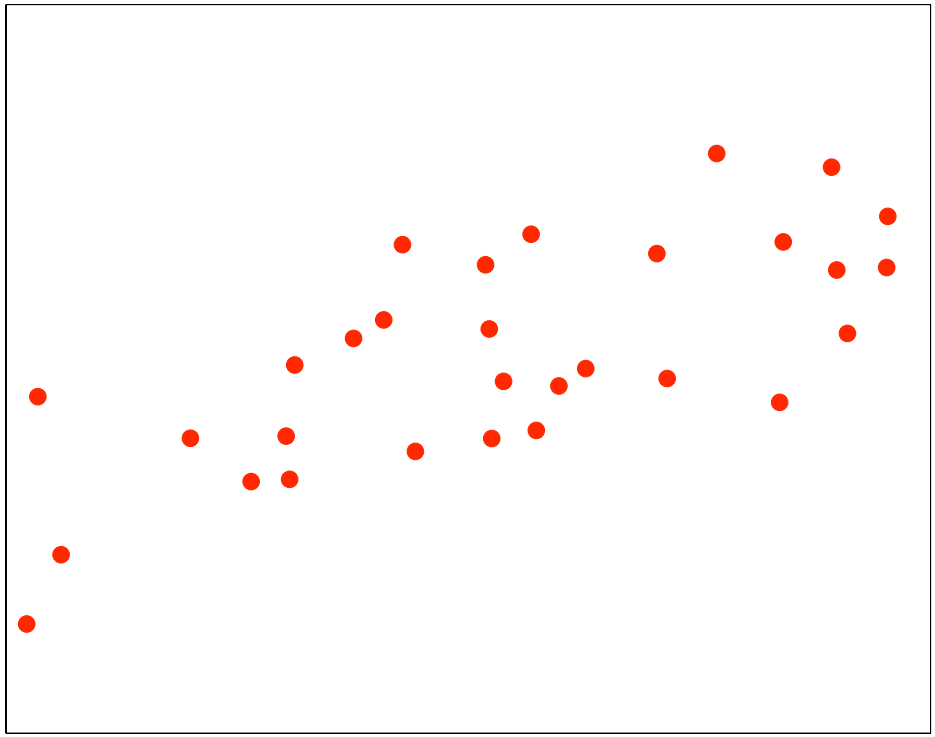
$$h : \mathcal{X} \to \mathcal{Y}$$

  such that $h(\mathbf{x})$ is a *"good predictor"* for the value of y.
- h is called a *hypothesis*
- Problems are categorized by the type of output domain
  - If $\mathcal{Y} = \mathbb{R}$, this problem is called *regression*
  - If $\mathcal{Y}$ is a finite discrete set, the problem is called *classification*
  - If $\mathcal{Y}$ has 2 elements, the problem is called *binary classification* or *concept learning*

Steps to solving a supervised learning problem

1. Decide what the input-output pairs are.
2. Decide how to encode inputs and outputs.
   This defines the input space $\mathcal{X}$, and the output space $\mathcal{Y}$.
   (We will discuss this in detail later)
3. Choose a class of hypotheses/representations $\mathcal{H}$.
4. ...

Example: What hypothesis class should we pick?



x	y
0.86	2.49
0.09	0.83
-0.85	-0.25
0.87	3.10
-0.44	0.87
-0.43	0.02
-1.10	-0.12
0.40	1.81
-0.96	-0.83
0.17	0.43

Linear hypothesis (HTF Ch. 3)

- Suppose y was a linear function of $\mathbf{x}$:

$$h_{\mathbf{w}}(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots$$

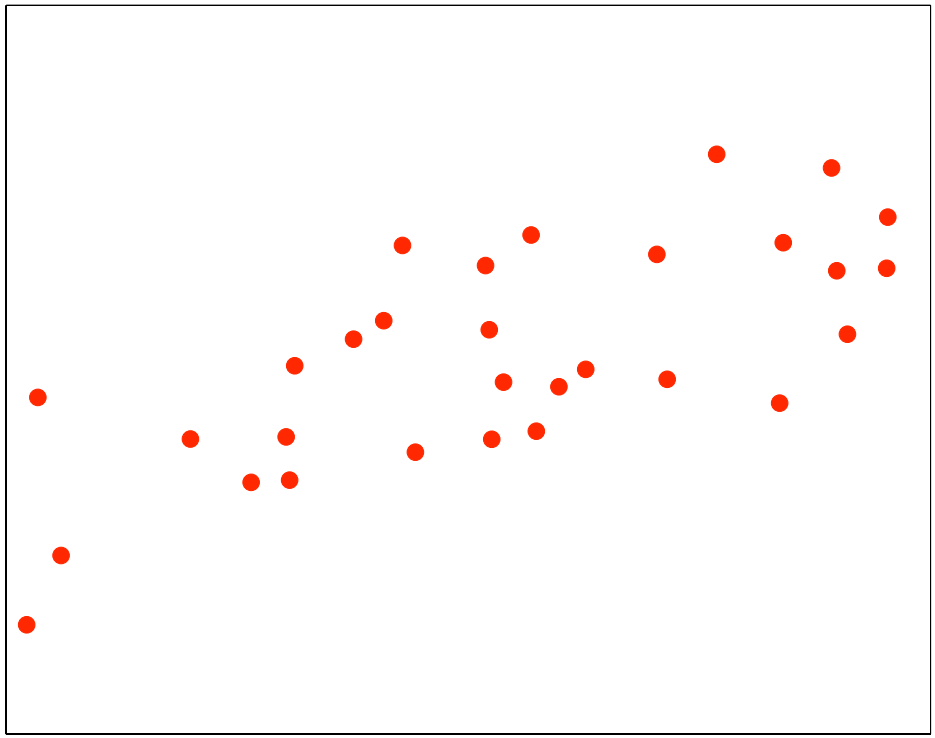
- w_i are called *parameters* or *weights*[1]
- We typically include an attribute $x_0 = 1$ (also called *bias term* or *intercept term*) so that the number of weights is $p + 1$. We then write:

$$h_{\mathbf{w}}(\mathbf{x}) = \sum_{i=0}^{p} w_i x_i = \mathbf{x}\mathbf{w}$$

  where $\mathbf{w}$ and $\mathbf{x}$ are vectors of size $p + 1$.
- The design matrix X is now n by $p + 1$.

[1] In statistics, β is commonly used. Also, in engineering, the word "parameter" sometimes means "feature".

Example: Design matrix with bias term



x_0	x_1	y
1	0.86	2.49
1	0.09	0.83
1	-0.85	-0.25
1	0.87	3.10
1	-0.44	0.87
1	-0.43	0.02
1	-1.10	-0.12
1	0.40	1.81
1	-0.96	-0.83
1	0.17	0.43

Hypotheses will be of the form

$$h_{\mathbf{w}}(\mathbf{x}) = x_0 w_0 + x_1 w_1 \quad (1)$$

$$= w_0 + x_1 w_1 \quad (2)$$

How should we pick $\mathbf{w}$*?*

Error minimization!

- Intuitively, $\mathbf{w}$ should make the predictions of $h_{\mathbf{w}}$ close to the true values y_i on on the training data
- Hence, we will define an *error function* or *cost function* to measure how much our prediction differs from the "true" answer on on the training data
- We will pick $\mathbf{w}$ such that the error function is minimized
- Hopefully, new examples are somehow "similar" to the training examples, and will also have small error.

How should we choose the error function?

Least mean squares (LMS)

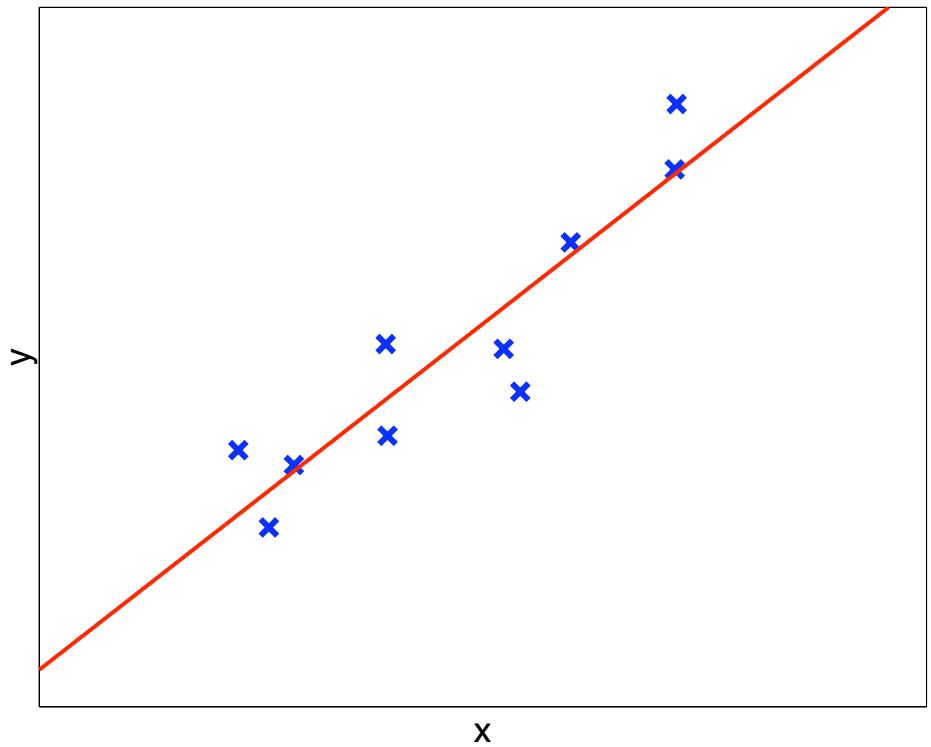
- Main idea: try to make $h_{\mathbf{w}}(\mathbf{x})$ close to y on the examples in the training set
- We define a *sum-of-squares* error function

$$J(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(h_{\mathbf{w}}(\mathbf{x}_i) - y_i)^2$$

  (the 1/2 is just for convenience)
- We will choose $\mathbf{w}$ such as to minimize $J(\mathbf{w})$
- One way to do it: compute $\mathbf{w}$ such that:

$$\frac{\partial}{\partial w_j} J(\mathbf{w}) = 0, \ \forall j = 0 \ldots p$$

Data and line $y = 1.05 + 1.60x$

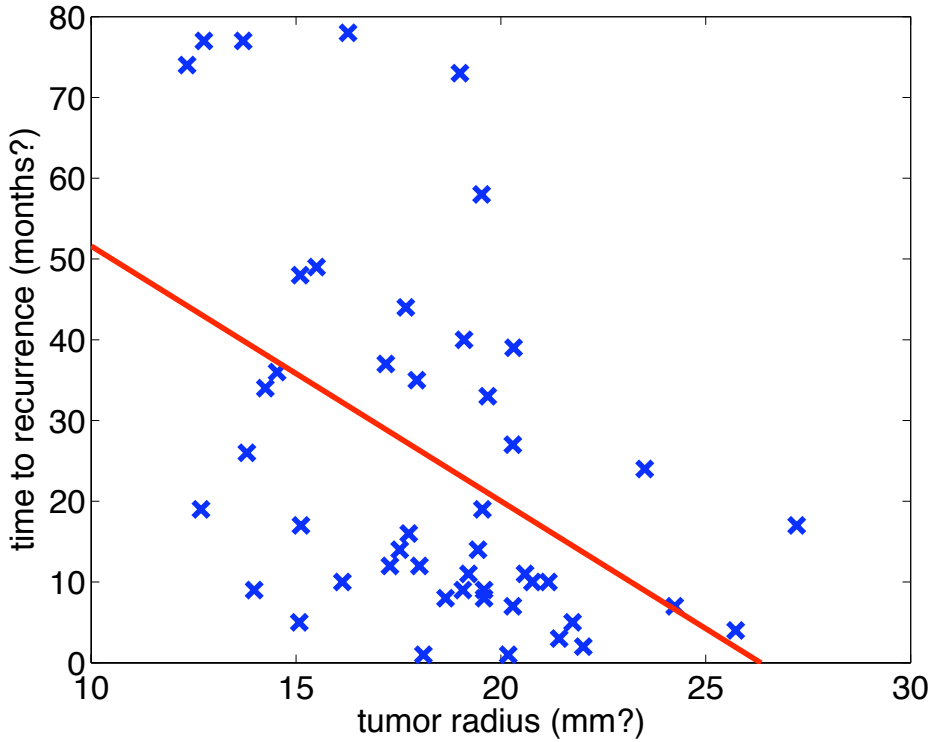


Here, $\mathbf{w} = (1.05, 1.60)^{\mathsf{T}}$

Steps to solving a supervised learning problem

1. Decide what the input-output pairs are.
2. Decide how to encode inputs and outputs.
   This defines the input space $\mathcal{X}$, and the output space $\mathcal{Y}$.
3. Choose a class of hypotheses/representations $\mathcal{H}$.
4. Choose an error function (cost function) to define the best hypothesis
5. Choose an algorithm for searching efficiently through the space of hypotheses.

Predicting recurrence time based on tumor size



Next time

- Solution to linear regression
- Non-linear regression
- Performance evaluation
- Overfitting
- Model selection