ECE 599/692 – Deep Learning

Lecture 2 - Background

Hairong Qi, Gonzalez Family Professor
Electrical Engineering and Computer Science
University of Tennessee, Knoxville
http://www.eecs.utk.edu/faculty/qi
Email: hqi@utk.edu

Outline

- Instructor and TA
  - Dr. Hairong Qi (hqi@utk.edu)
  - Chengcheng Li (cli42@vols.utk.edu)
- What's the difference between different courses and terminologies?
- Why deep learning?
  - Seminar works
  - Engineered features vs. Automatic features
- What do we cover?
- What's the expectation?
  - ECE599
  - ECE692
- Programming environment
  - Tensorflow and Google Cloud Platform (GCP)
- Preliminaries
  - Linear algebra, probability and statistics, numerical computation, machine learning basics

2

Different Courses

- Machine Learning (ML) (CS425/528)
- Pattern Recognition (PR) (ECE471/571)
  - Reinforcement Learning (RL) (ECE517)
  - Biologically-Inspired Computation (CS527)
- Deep Learning (DL) (ECE599/692)
- Artificial Intelligence (AI) (CS529 – Autonomous Mobile Robots)

???! Sept. 2017: https://www.alibabacloud.com/blog/deep-learning-vs-machine-learning-vs-pattern-recognition_207110

???! Mar. 2015, Tombone's Computer Vision Blog: http://www.computervisionblog.com/2015/03/deep-learning-vs-machine-learning-vs.html

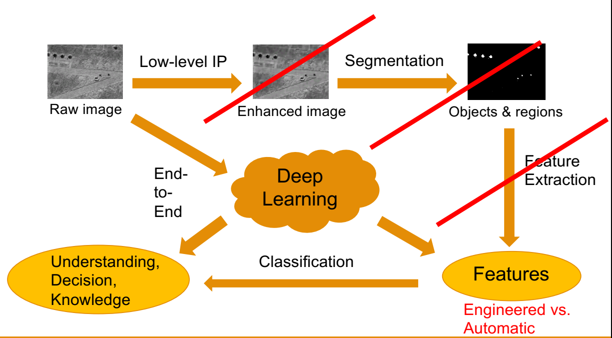
3

Different Terminologies

- Pattern Recognition vs. Pattern Classification
- Machine Learning vs. Artificial Intelligence
- Machine Learning vs. Pattern Recognition
- Engineered Features vs. Automatic Features

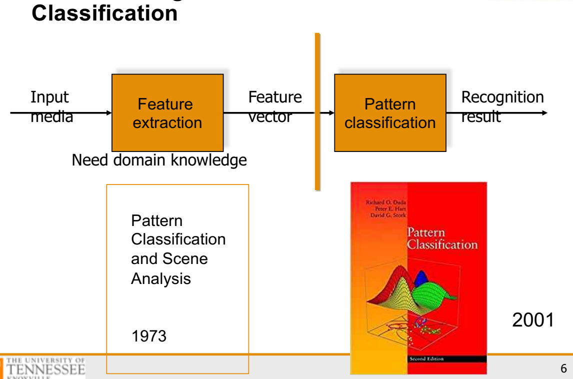
4

The New Deep Learning Paradigm



5

Pattern Recognition vs. Pattern Classification



6

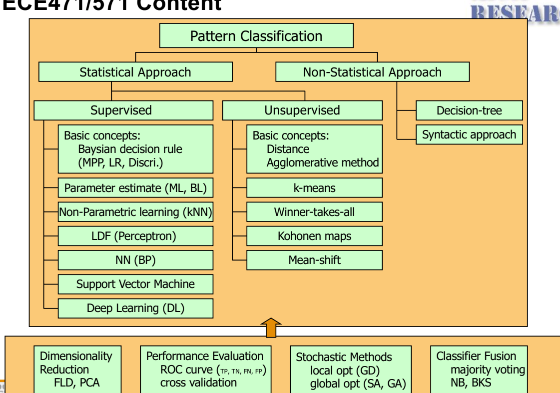
AI vs. ML or PR

PR + Reasoning (RNN) → AI
PR + Planning & RL → AI

CS425/528 Content

- Introduction (ch. 1)
- Supervised Learning (ch. 2)
- Bayesian Decision Theory (ch. 3)
- Parametric Methods (chs. 4–5)
- Dimensionality Reduction (ch. 6)
- Clustering (ch. 7)
- Non-Parametric Methods (ch. 8)
- Decision Trees (ch. 9)
- Neural Networks (chs. 10–11)
- Local Models (ch. 12)
- Kernel Machines (ch. 13)
- Reinforcement Learning (ch. 18)
- Machine Learning Experiments (ch. 19)

ECE471/571 Content



What Do We Cover?

AICIP RESEARCH

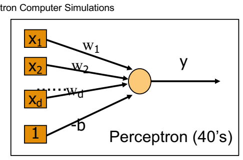
- Neural networks
  - Multi-layer Perceptron
  - Backpropagation Neural Network (Project 1, Due 09/07)
- Feedforward networks
  - Supervised learning - CNN (Project 2, Due 09/21)
  - Unsupervised learning – AE (Project 3, Due 10/12)
- Generative networks
  - GAN (Project 4, Due 10/26)
- Feedback networks
  - RNN (Project 5, Due 11/09)
- Final project (Due TBD)

THE UNIVERSITY OF TENNESSEE KNOXVILLE 10

A Bit History

AICIP RESEARCH

- 1943 (McCulloch and Pitts):
- 1957 - 1962 (Rosenblatt):
  - From Mark I Perceptron to the Tobermory Perceptron to Perceptron Computer Simulations
  - Multilayer perceptron with fixed threshold
- 1969 (Minsky and Papert):
- The dark age: 70's ~25 years
- 1986 (Rumelhart, Hinton, McClelland): BP
- 1989 (LeCun et al.): CNN (LeNet)
- Another ~25 years
- 2006 (Hinton et al.): DL
- 2012 (Krizhevsky, Sutskever, Hinton): AlexNet
- 2014 (Goodfellow, Benjo, et al.): GAN



- W.S. McCulloch, W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, 5(4):115-133, December **1943**.
- F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Spartan Books, **1962**.
- Minsky, S. Papert, *Perceptrons: An Introduction to Computational Geometry*, **1969**.
- D.E. Rumelhart, G.E. Hinton, R.J. Williams, "Learning representations by back-propagating errors," Nature, 323(9):533-536, October **1986**. (BP)
- Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, "Backpropagation applied to handwritten zip code recognition," Neural Computation, 1(4):541-551, **1989**. (LeNet).
- G.E. Hinton, S. Osindero, Y. Teh, "A fast learning algorithm for deep belief nets," Neural Computation, 18:1527-1554, **2006**. (DL)
- G.E. Hinton, R.R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, 313(5786):504-507, **2006** (DL)
- A. Krizhevsky, I. Sutskever, G.E. Hinton, "ImageNet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, pages 1097-1105, **2012**. (AlexNet)
- I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative adversarial networks," NIPS, **2014**.

THE UNIVERSITY OF TENNESSEE KNOXVILLE 11

A Bit History - Revisited

AICIP RESEARCH

- 1956-1976
  - 1956, The Dartmouth Summer Research Project on Artificial Intelligence, organized by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon

    > We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College ... The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

  - The rise of symbolic methods, systems focused on limited domains, deductive vs. inductive systems
  - 1973, the Lighthill report by James Lighthill, "Artificial Intelligence: A General Survey" - automata, robotics, neural network
  - 1976, the AI Winter
- 1976-2006
  - 1986, BP algorithm
  - ~1995, The Fifth Generation Computer
- 2006-???
  - 2006, Hinton (U. of Toronto), Bingio (U. of Montreal, LeCun (NYU)
  - 2012, ImageNet by Fei-Fei Li (2010-2017) and AlexNet

THE UNIVERSITY OF TENNESSEE KNOXVILLE 12

https://en.wikipedia.org/wiki/Dartmouth_workshop
https://en.wikipedia.org/wiki/Lighthill_report

AICIP RESEARCH

Why Deep Learning?

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

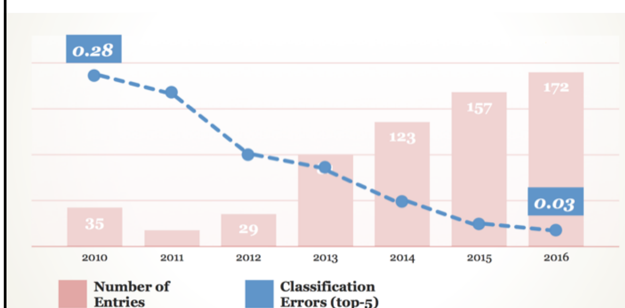
Year	Top-5 Error	Model
2010 winner	28.2%	Fast descriptor coding
2011 winner	25.7%	Compressed Fisher vectors
2012 winner	15.3%	AlexNet (8, 60M)
2013 winner	14.8%	ZFNet
2014 winner 2014 runner-up	6.67%	GoogLeNet (22, 4M) VGGNet (16, 140M)
2015 winner	3.57%	ResNet (152)

Human expert: 5.1%

http://image-net.org/challenges/talks_2017/imagenet_ilsvrc2017_v1.0.pdf

THE UNIVERSITY OF TENNESSEE 13

AICIP RESEARCH



http://image-net.org/challenges/talks_2017/imagenet_ilsvrc2017_v1.0.pdf

THE UNIVERSITY OF TENNESSEE 14

AICIP RESEARCH

Preliminaries

- Math and Statistics
  - Linear algebra
  - Probability and Statistics
  - Numerical computation
- Machine learning basics
  - Neural networks and backpropagation
- Programming environment
  - Tensorflow
  - GCP

THE UNIVERSITY OF TENNESSEE 15

Linear Algebra

- Scalars, vectors, matrices, tensors
- Linear dependence and span
- Norms
  - l_p norms, l_0 norm
  - Frobenius norm - l_2 norm for matrices
- Matrix decomposition
  - Eigendecomposition (for square matrices)
  - Singular value decomposition (SVD) (for any matrices)
  - [Snyder&Qi:2017]

THE UNIVERSITY OF TENNESSEE KNOXVILLE 16

Probability

- Frequentist probability vs. Baysian probability
- Probability distribution
  - Discrete variable and probability mass function (PMF)
  - Continuous variable and probability distribution function (PDF)
- Marginal probability
- Conditional probability (e.g., Baye's rule)

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)}$$

THE UNIVERSITY OF TENNESSEE KNOXVILLE 17

Information Theory

- Measuring information
  - Self-information of an event x=x, $I(x)$ = -log$P(x)$
    - Base e: unit (nats) information gained by observing an event of probability 1/e
    - Base 2: unit (bits or shannons)
  - Shannon entropy: $H(x) = E_{x\sim P}[I(x)] = -E_{x\sim P}[\log P(x)]$
- Kullback-Leibler (KL) divergence
  - $D_{KL}(P||Q) = E_{x\sim P}[\log P(x)/Q(x)] = E_{x\sim P}[\log P(x) - \log Q(x)]$
- Cross-entropy
  - $H(P,Q) = H(P) + D_{KL}(P||Q)$

THE UNIVERSITY OF TENNESSEE KNOXVILLE 18

Numerical Computation

- Global vs. local optimization
- Gradient descent
- Constrained optimization
  - Langrange optimization
  - Karush-Kuhn-Tucker (KKT) approach

19

Pattern Classification Approaches

- Supervised vs. unsupervised
- Parametric vs. non-parametric
- Classification vs. regression vs. generation
- Training set vs. test set vs. validation set
- Cross-validation

20

Pattern Classification Approaches

- Supervised
  - Maximum a-posteriori probability $P(\omega_j \mid x) = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)}$
  - kNN $P(\omega_m \mid x) = \frac{p(x \mid \omega_m) P(\omega_m)}{p(x)} = \frac{\frac{k_m}{n_m V} \frac{n_m}{n}}{\frac{k}{nV}} = \frac{k_m}{k}$
  - NN, when n -> infty, $g_k(\mathbf{x};\mathbf{w})$ -> $P(w_k|x)$

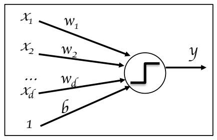
21

Neural Networks

- Perceptrons

$$y = \begin{cases} 0 & \mathbf{w}^T\mathbf{x} + b \le 0 \\ 1 & \mathbf{w}^T\mathbf{x} + b > 0 \end{cases}$$

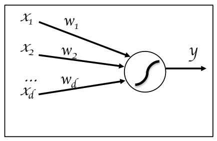
where b = -threshold



- Sigmoid neurons

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

$$y = \frac{1}{1 + \exp(-(\mathbf{w}^T\mathbf{x} + b))}$$



Network Example – MNIST Recognition

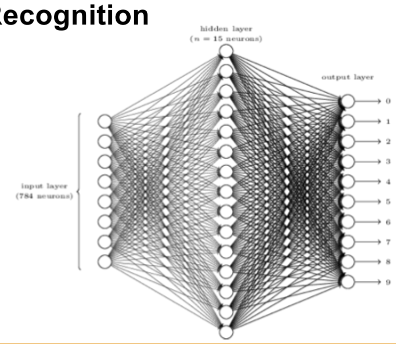
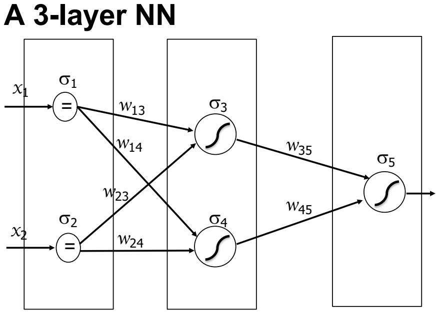


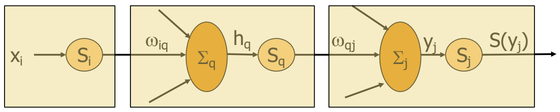
Image from: [Nielson]

A 3-layer NN



BP – 3-layer Network

AICIP RESEARCH



$$E = \frac{1}{2}\sum_j \left(T_j - S\left(y_j\right)\right)^2$$

Choose a set of initial ω_{st}

$$\omega_{st}^{\ k+1} = \omega_{st}^{\ k} - c^k \frac{\partial E^k}{\partial \omega_{st}^{\ k}}$$

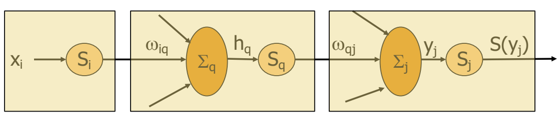
ω_{st} is the weight connecting input *s* at neuron *t*

The problem is essentially "how to choose weight ω to minimize the error between the expected output and the actual output"

The basic idea behind BP is gradient descent

The Derivative – Chain Rule

AICIP RESEARCH



$$\Delta\omega_{qj} = -\frac{\partial E}{\partial \omega_{qj}} = -\frac{\partial E}{\partial S_j}\frac{\partial S_j}{\partial y_j}\frac{\partial y_j}{\partial \omega_{qj}}$$

$$= -\left(T_j - S_j\right)\left(S_j'\right)\left(S_q\left(h_q\right)\right)$$

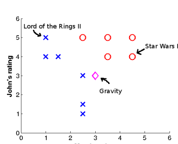
$$\Delta\omega_{iq} = -\frac{\partial E}{\partial \omega_{iq}} = \left[\sum_j \frac{\partial E}{\partial S_j}\frac{\partial S_j}{\partial y_j}\frac{\partial y_j}{\partial S_q}\right]\frac{\partial S_q}{\partial h_q}\frac{\partial h_q}{\partial \omega_{iq}}$$

$$= \left[\sum_j \left(T_j - S_j\right)\left(S_j'\right)\left(\omega_{qj}\right)\right]\left(S_q'\right)\left(x_i\right)$$

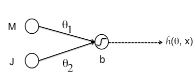
Why Deeper?

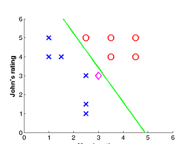
AICIP RESEARCH

Movie name	Mary's rating	John's rating	I like?
Lord of the Rings II	1	5	No
...	...	...	...
Star Wars I	4.5	4	Yes
Gravity	3	3	?



$$h(x;\theta,b) = \theta_1 x_1 + \theta_2 x_2 + b,$$



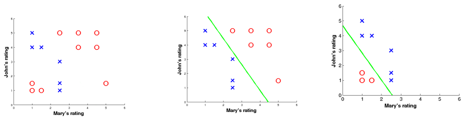


$$J(\theta,b) = \left(h(x^{(1)};\theta,b) - y^{(1)}\right)^2 + \left(h(x^{(2)};\theta,b) - y^{(2)}\right)^2 + \ldots + \left(h(x^{(m)};\theta,b) - y^{(m)}\right)^2$$

$$= \sum_{i=1}^{m} \left(h(x^{(i)};\theta,b) - y^{(i)}\right)^2$$

http://ai.stanford.edu/~quocle/tutorial2.pdf

Why Deeper? - Another Example



Movie name	Output by decision function h_1	Output by decision function h_2	Susan likes?
Lord of the Rings II	$h_1(x^{(1)})$	$h_2(x^{(2)})$	No
...	...	...	...
Star Wars I	$h_1(x^{(n)})$	$h_2(x^{(n)})$	Yes
Gravity	$h_1(x^{(n+1)})$	$h_2(x^{(n+1)})$	?