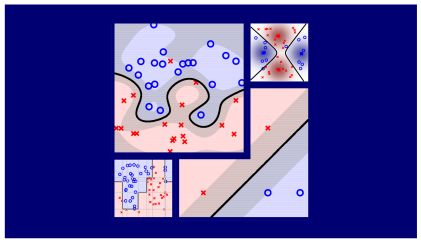
Machine Learning Soundings
(機器學習深測)



Lecture 4: Regularization in Deep Learning

Hsuan-Tien Lin (林軒田)

htlin@csie.ntu.edu.tw

Department of Computer Science
& Information Engineering

National Taiwan University
(國立台灣大學資訊工程系)

Roadmap

1. Deep Learning Foundations

Lecture 1: Neural Network

automatic **pattern feature extraction** from **layers of neurons** with **backprop** for GD/SGD

Lecture 4: Regularization in Deep Learning

2. Deep Learning Models

Co-Adapatation Issue

* difficulty in DL generalization
  - co-adaptation :

(consistent mistakes from some neurons) corrected by fitting other neurons

⇓ ⇓

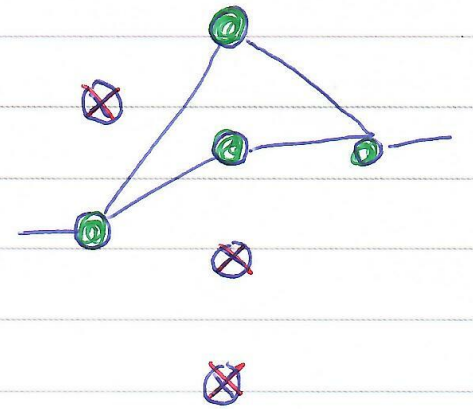
like noise like overfit

and vice versa

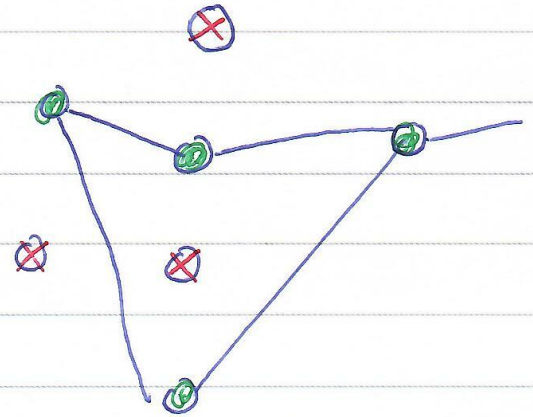
(dependence)

The Dropout Technique

* break the dependence : shut down neurons (randomly)



mini-batch-1



mini-batch-2

dropout : drop p
keep (1-p) ⇒ many thin networks (combined)

⇓ slows down convergence but faster per-iteration

↘ noise as regularization like DAE

Dropout During Testing

* dropout during testing
  - full-net prediction w/o changing

$$\underbrace{\mathbb{E}(s_i^{(l)})}_{\text{in training}} = \underbrace{(1-p)\, s_i^{(l)}}_{\text{in testing}}$$

  - test-time "pseudo-" dropout

$$x_i^{(l)} = \theta\left(\underbrace{(1-p)}_{\text{need to record } p,\ \text{less flexibility for changing } p \text{ per neuron or dynamically}}\, s_i^{(l)}\right)$$

  - inverted dropout

training : dropout & $x_i^{(l)} = \theta\left(s_i^{(l)} / (1-p)\right)$

testing : $x_i^{(l)} = \theta\left((1-p)\, s_i^{(l)} / (1-p)\right)$

$= \theta(s_i^{(l)})$ unchanged (usually preferred)