Learning theory for neural computation models

Daniel Hsu

COMS 6998-7 Spring 2025

What is a neural net?

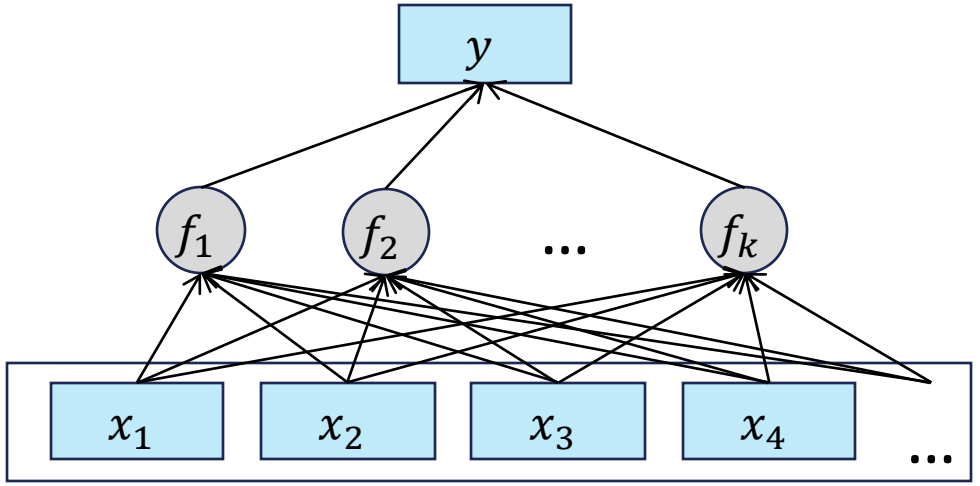
- Straight-line program (a.k.a. circuit)
  - Each line of program assigns a value to a new variable
    - Value can be a constant
    - Or, value can be result of an operation applied to other variables
      - Other variables are either previously-defined or "free variables" (inputs to program)
    - Some variables designated as outputs
- Neural net: operations are various differentiable functions (in pytorch)
- Arithmetic circuit: operations are + and ×
  - Computes polynomials
- "$\mathcal{A}_{k,n}$ w/o jumps" from Goldberg & Jerrum: operations are +,−,×,÷
  - Computes rational functions

What is a two-layer neural networks?

- Two-layer neural net is a linear combination of hidden units

$$y = \sum_{i=1}^{k} w_i \, f_i(x)$$

  - Each "hidden unit" is some function of the input $x = (x_1, \ldots, x_N)$



Statistical theory for learning

- Training data, test data: iid samples from same distribution
- Learning: pick hypothesis from hypotheses class using training data
- Successful generalization: good performance on test data
- Why is generalization possible? Many plausible reasons
  - Learning theory 101 (COMS 4252, COMS 4773) reason: uniform convergence
  - But many other plausible reasons are known!
    - E.g., different learning criteria, distribution-specific learning, leveraging dependence structures, alternative models of supervision/teaching
- Many tools for understanding uniform convergence
  - VC dimension
  - Metric entropy (i.e., log of covering number)
  - …

Sparsity and norm bounds

- Linear classifiers $f_{w,b}: \{0,1\}^n \to \{-1,1\}$

$$f_{w,b}(x) = \text{sign}(w \cdot x - b)$$

  - Monotone disjunction over variables x_i for $i \in S$:
    Represent as linear classifier with $b = 1/2$, $w_i = 1$ for $i \in S$, $w_i = 0$ for $i \notin S$
    - L^1 norm of w is $\|w\|_1 = |S|$
  - Winnow mistake bound:

$$O(b\|w\|_1 \log n)$$

  - Sample complexity for learning:

$$O\left(\frac{b\|w\|_1 \log n}{\epsilon}\right)$$

  - Can also derive uniform convergence bounds of this form using Rademacher complexity, covering numbers, etc.