Machine learning lecture slides

COMS 4771 Fall 2020

Optimization II: Neural networks

Outline

- Architecture of (layered) feedforward neural networks
- Universal approximation
- Backpropagation
- Practical issues

Parametric featurizations

- So far: data features (x or $\varphi(x)$) are fixed during training
  - Consider a (small) collection of feature transformations φ
  - Select φ via cross-validation – outside of normal training
- "Deep learning" approach:
  - Use φ with many tunable parameters
  - Optimize parameters of φ during normal training process
- *Neural network*: parameterization for function $f: \mathbb{R}^d \to \mathbb{R}$
  - $f(x) = \varphi(x)^\top w$
  - Parameters include both w and parameters of φ
  - Varying parameters of φ allows f to be essentially any function!
  - Major challenge: optimization

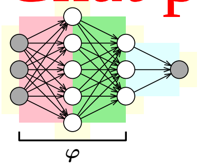

Figure 1: Neural networks as feature maps

Feedforward neural network

- *Architecture* of a *feedforward neural network*
  - Directed acyclic graph $G = (V, E)$
  - One *source* node (vertex) per input to the function $(x_1, \ldots, x_d)$
  - One *sink* node per output of the function
  - Internal nodes are called *hidden units*
  - Each edge $(u, v) \in E$ has a weight parameter $w_{u,v} \in \mathbb{R}$
  - *Value* h_v of node v given values of parents $\pi_G(v) = \{u \in V : (u, v) \in E\}$ is

$$h_v := \sigma_v(z_v), \quad z_v := \sum_{u \in \pi_G(v)} w_{u,v} \cdot h_u.$$

  - $\sigma_v : \mathbb{R} \to \mathbb{R}$ is the *activation function* (a.k.a. *link function*)
  - E.g., sigmoid function $\sigma_v(z) = 1/(1 + e^{-z})$
- Inspired by neurons in the brain

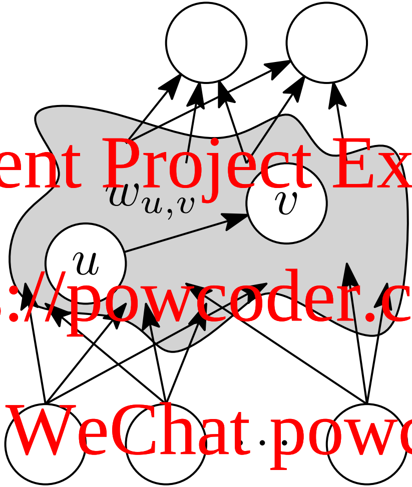

Figure 2: Computation DAG of a feedforward neural network

Standard layered architectures

- Standard feedforward architecture arranges nodes into *layers*
  - Initial layer (layer zero): source nodes (input)
  - Final layer (layer L): sink nodes (output)
  - (Layer counting is confusing; usually don't count input)
- Edges only go from one layer to the next
  - (Non-standard feedforward architectures (e.g., ResNets) break this rule)
- Can write function using matrices of weight parameters

$$f(x) = \sigma_L(W_L \sigma_{L-1}(\cdots \sigma_1(W_1 x) \cdots))$$

  - Layer ℓ has d_ℓ nodes
  - $W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ are the weight parameters for layer ℓ
  - Scalar-valued activation function $\sigma_\ell \colon \mathbb{R} \to \mathbb{R}$ (e.g., sigmoid) is applied coordinate-wise to input
- Often also include "bias" parameters $b_\ell \in \mathbb{R}^{d_\ell}$

$$f(x) = \sigma_L(b_L + W_L \sigma_{L-1}(\cdots \sigma_1(b_1 + W_1 x) \cdots))$$

- The tunable parameters: $\theta = (W_1, b_1, \ldots, W_L, b_L)$

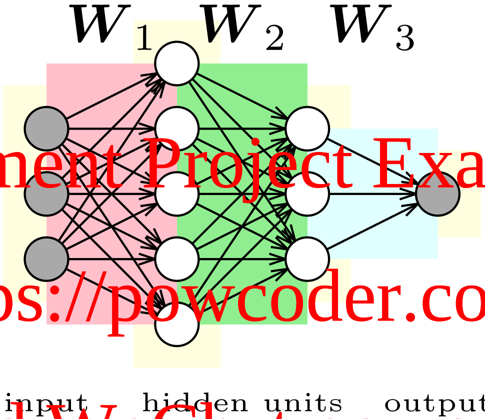

Figure 3: Layered feedforward neural network

Well-known activation functions

- *Heaviside*: $\sigma(z) = \mathbf{1}_{\{z \geq 0\}}$
  - Popular in the 1940s; also called *step function*
- *Sigmoid* (from logistic regression): $\sigma(z) = 1/(1 + e^{-z})$
  - Popular since 1970s
- *Hyperbolic tangent*: $\sigma(z) = \tanh(z)$
  - Similar to sigmoid, but range is $(-1, 1)$ rather than $(0, 1)$
- *Rectified Linear Unit* (*ReLU*): $\sigma(z) = \max\{0, z\}$
  - Popular since 2012
- *Identity*: $\sigma(z) = z$
  - Popular with luddites

- What happens if every activation function is linear/affine?
  - Overall function is affine
  - An unusual way to parameterize an affine function
- Therefore, use non-linear/non-affine activation functions

Necessity of multiple layers (1)

- Suppose only have input and output layers, so function f is

$$f(x) = \sigma(b + w^\mathsf{T} x)$$

where $b \in \mathbb{R}$ and $w \in \mathbb{R}^d$ (so $w^\mathsf{T} \in \mathbb{R}^{1 \times d}$)
- If σ is monotone (e.g., Heaviside, sigmoid, hyperbolic tangent, ReLU, identity), then f has same limitations as a linear/affine classifier

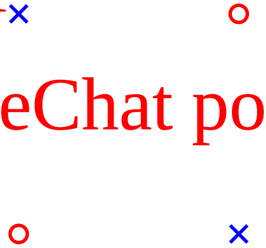

Figure 4: XOR data set

- XOR problem
  - Let $x^{(1)} = (+1, +1)$, $x^{(2)} = (+1, -1)$, $x^{(3)} = (-1, +1)$, $x^{(4)} = (-1, -1)$.
  - $y^{(i)} = +1$ iff coordinates of $x^{(i)}$ are the same. (XNOR)
  - Suppose $(w, b) \in \mathbb{R}^2 \times \mathbb{R}$ satisfies

$$\begin{aligned} -b - w^\top x^{(1)} &< 0, \\ b + w^\top x^{(2)} &< 0, \\ b + w^\top x^{(3)} &< 0. \end{aligned}$$

  - Add up the equations:

$$b + w^\top (x^{(2)} + x^{(3)} - x^{(1)}) < 0.$$

  - But $x^{(2)} + x^{(3)} - x^{(1)} = x^{(4)}$, so

$$b + w^\top x^{(4)} < 0.$$

    In other words, cannot correctly label $x^{(4)}$.

- **Theorem** (Cybenko, 1989; Hornik, Stinchcombe, & White, 1989): Let σ_1 be any continuous non-linear activation function from above. For any continuous function $f : \mathbb{R}^d \to \mathbb{R}$ and any $\varepsilon > 0$, there is a two-layer neural network (with parameters $\theta = (W_1, b_1, w_2)$) s.t.

$$\max_{x \in [0,1]^d} |f(x) - w_2^\top \sigma_1(b_1 + W_1 x)| \leq \varepsilon.$$

- This property of such families of neural networks is called *universal approximation*.
- Many caveats
  - "Width" (number of hidden units) may need to be very large
  - Does not tell us how to find the network
  - Does not justify deeper networks

Theorem (Weierstrass, 1885): For any continuous function $f\colon [a,b] \to \mathbb{R}$, and any $\epsilon > 0$, there exists a polynomial p such that

$$\sup_{x \in [a,b]} |f(x) - p(x)| < \epsilon.$$

Theorem (Stone, 1937): Let $K \subset \mathbb{R}^d$ be any bounded set. Let A be a set of continuous functions on K such that the following hold.

(1) A is an *algebra* (i.e., A is closed under addition, multiplication, and scalar multiplication).

(2) A *does not vanish on K* (i.e., for all $x \in K$, there exists $h \in A$ such that $h(x) \neq 0$).

(3) A *separates points in K* (i.e., for all distinct $x, y \in K$, there exists $h \in A$ such that $h(x) \neq h(y)$).

For any continuous function $f : K \to \mathbb{R}$ and any $\epsilon > 0$, there exists $h \in A$ such that

$$\sup_{x \in K} |f(x) - h(x)| < \epsilon.$$

Let $K = [0,1]^d$, and let

$$A = \Big\{ x \mapsto \sum_{k=1}^{m} a_k \cos(x^\top w_k + b_k) :$$
$$m \in \mathbb{N}, a_k, b_k \in \mathbb{R}, w_k \in \mathbb{R}^d \text{ for } k = 1, \dots, m \Big\}.$$

(Check that A satisfies properties of Stone-Weierstrass theorem.)

Let $K = [0,1]^d$, and let

$$A = \left\{ x \mapsto \sum_{k=1}^{m} a_k \exp(x^\top w_k + b_k) : \right.$$

$$\left. m \in \mathbb{N}, a_k, b_k \in \mathbb{R}, w_k \in \mathbb{R}^d \text{ for } k = 1, \dots, m \right\}.$$

(Check that A satisfies properties of Stone-Weierstrass theorem.)

- Training data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$
- Fix architecture: $G = (V, E)$ and activation functions
- ERM: find parameters θ of neural network f_θ to minimize empirical risk (possibly with a surrogate loss)
  - Regression $y_i \in \mathbb{R}$

$$\widehat{\mathcal{R}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} (f_\theta(x_i) - y_i)^2$$

  - Binary classification $y_i \in \{-1, +1\}$

$$\widehat{\mathcal{R}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln(1 + \exp(-y_i f_\theta(x_i)))$$

    (Could use other surrogate loss functions . . .)
  - Can also add regularization terms, but also common to use *algorithmic regularization*
- Typically objective is not convex in parameters θ
- Nevertheless, local search (e.g., gradient descent, SGD) often works well!

Backpropagation

- *Backpropagation (backprop)*: Algorithm for computing partial derivatives wrt weights in a feedforward neural network
  - Clever organization of partial derivative computations with *chain rule*
  - Use in combination with gradient descent, SGD, etc.
- Consider loss on a single example (x, y), written as

$$J := \ell(y, f_\theta(x))$$

- Goal: compute $\frac{\partial J}{\partial w_{u,v}}$ for every edge $(u, v) \in E$
- Initial step of backprop: *forward propagation*
  - Compute z_v's and h_v's for every node $v \in V$
  - Running time: linear in size of network
- We'll see that rest of backprop also just requires time linear in size of network

Derivative of loss with respect to weights

- Let $\hat{y}_1, \hat{y}_2, \ldots$ denote the values at the output nodes.
- Then by chain rule,

$$\frac{\partial J}{w_{u,v}} = \sum_i \frac{\partial J}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{w_{u,v}}.$$

- $\frac{\partial J}{\partial \hat{y}}$ is just determined by the loss function (e.g., squared loss)
- So just have to focus on $\frac{\partial \hat{y}_i}{w_{u,v}}$
- Assume for simplicity there is just a single output, $\hat{y}$

- Chain rule, again:

$$\frac{\partial \hat{y}}{\partial w_{u,v}} = \frac{\partial \hat{y}}{\partial h_v} \cdot \frac{\partial h_v}{\partial w_{u,v}}$$

- First term: trickier; we'll handle later
- Second term:

  - $h_v = \sigma_v(z_v)$
  - $z_v = w_{u,v} \cdot h_u +$ (terms not involving $w_{u,v}$)
  - Therefore

$$\frac{\partial h_v}{\partial w_{u,v}} = \frac{\partial h_v}{\partial z_v} \cdot \frac{\partial z_v}{\partial w_{u,v}} = \sigma'(z_v) \cdot h_u.$$

  - z_v and h_u were computed during forward propagation

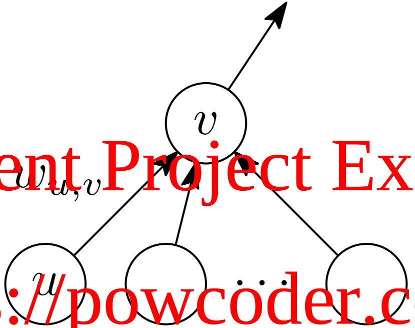

Figure 5: Derivative of a node's output with respect to an incoming weight

Derivative of output with respect to hidden units

- Key trick: compute $\frac{\partial \hat{y}}{\partial h_v}$ for all vertices in decreasing order of layer number
- If v is not the output node, then by chain rule (yet again)

$$\frac{\partial \hat{y}}{\partial h_v} = \sum_{v' : (v,v') \in E} \frac{\partial \hat{y}}{\partial h_{v'}} \cdot \frac{\partial h_{v'}}{\partial h_v}$$

  - $\frac{\partial \hat{y}}{\partial h_{v'}}$ was already computed since v' is in higher layer than v
  - $h_{v'} = \sigma_{v'}(z_{v'})$
  - $z_{v'} = w_{v,v'} \cdot h_v +$ (terms not involving h_v)
  - Therefore

$$\begin{aligned} \frac{\partial h_{v'}}{\partial h_v} &= \frac{\partial h_{v'}}{\partial z_{v'}} \cdot \frac{\partial z_{v'}}{\partial h_v} \\ &= \sigma'(z_{v'}) \cdot w_{v,v'} . \end{aligned}$$

  - $z_{v'}$'s were computed during forward propagation
  - $w_{v,v'}$'s are the values of the weight parameters at which we want to compute the gradient

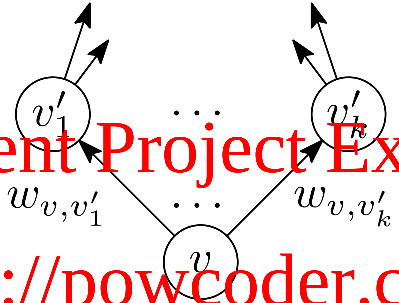

Figure 6: Derivative of the network output with respect to hidden unit values

Example: chain graph (1)

- Function $f_\theta \colon \mathbb{R} \to \mathbb{R}$
- Architecture
  - DAG: $0 \longrightarrow 1 \longrightarrow 2 \longrightarrow \cdots \longrightarrow L$
  - Same activation σ in every layer
- Parameters $\theta = (w_{0,1}, w_{1,2}, \ldots, w_{L-1,L})$
  - Input is at vertex 0, and output is at vertex L
- Fix input value $x \in \mathbb{R}$; what is $\frac{\partial h_L}{\partial w_{i-1,i}}$ for $i = 1, \ldots, L$?
- Forward propagation:
  - $h_0 := x$
  - For $i = 1, 2, \ldots, L$:

$$z_i := w_{i-1,i} h_{i-1}$$
$$h_i := \sigma(z_i)$$

- Backprop:
  - For $i = L, L-1, \ldots, 1$:

$$\frac{\partial h_L}{\partial h_i} := \begin{cases} 1 & \text{if } i = L \\ \frac{\partial h_L}{\partial h_{i+1}} \cdot \sigma'(z_{i+1}) w_{i,i+1} & \text{if } i < L \end{cases}$$

$$\frac{\partial h_L}{\partial w_{i-1,i}} := \frac{\partial h_L}{\partial h_i} \cdot \sigma'(z_i) h_{i-1}$$

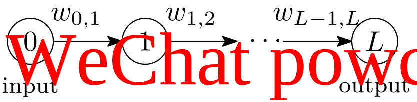

Figure 7: Neural network with a chain computation graph

Practical issues I: Initialization

- Ensure inputs are *standardized*: every feature has zero mean and unit variance (wrt training data)
  - Even better: different features have zero covariance (again, on training data)
  - But this can be expensive
- Initialize weights randomly for gradient descent / SGD
  - Standard normal random variables (or similar)
  - What should variance be?
  - Heuristic: ensure h_v have similar statistics as inputs
  - E.g., using $\tanh$-activation, if v has in-degree k, use variance $1/k$ for all weights $w_{u,v}$
  - Many initialization schemes for other activations (e.g., ReLU), dealing with bias parameters, . . .

Practical issues II: Architecture choice

- Architecture can be regarded as a "hyperparameter"
  - Could use cross-validation to select, but ...
  - Many "good" architectures are known for popular problems (e.g., image classification)
  - Unclear what to do for completely new problems
- Optimization-inspired architecture choice
  - With wide enough network, can get zero training error
  - Use the smallest network that lets you get zero training error
  - Then add regularization term to objective (e.g., sum of squares of weights), and optimize the regularized ERM objective
- Entire research communities are trying to figure out good architectures for their problems

Multi-class

Vector-valued activation: $\sigma \colon \mathbb{R}^{d_\ell} \to \mathbb{R}^{d_\ell}$

- *Softmax* activation: $\sigma(v)_i = \exp(v_i) / \sum_j \exp(v_j)$
- Common to use this in final layer

- Neural networks with *convolutional layers*
  - Useful when inputs have locality structure
  - Sequential structure (e.g., speech waveform)
  - 2D grid structure (e.g., image)
  - ...
- Weight matrix W_ℓ is highly-structured
  - Determined by a small *filter*
  - Time to compute $W_\ell h_{\ell-1}$ is typically $\ll d_\ell \times d_{\ell-1}$ (e.g., closer to $\max\{d_\ell, d_{\ell-1}\}$)

Convolutions I

- Convolution of two continuous functions: $h := f * g$

$$h(x) = \int_{-\infty}^{+\infty} f(y)g(x-y)\,\mathrm{d}y$$

- If $f(x) = 0$ for $x \notin [-w, +w]$, then

$$h(x) = \int_{-w}^{+w} f(y)g(x-y)\,\mathrm{d}y$$

  - Replaces $g(x)$ with weighted combination of g at nearby points

- For functions on discrete domain, replace integral with sum

$$h(i) = \sum_{j=-\infty}^{\infty} f(j)\,g(i-j)$$

E.g., suppose only $f(0), f(1), f(2)$ are non-zero. Then:

$$\begin{bmatrix} h(1) \\ h(2) \\ h(3) \\ h(4) \\ h(5) \\ h(6) \\ h(7) \end{bmatrix} = \begin{bmatrix} f(0) & 0 & 0 & 0 & 0 \\ f(1) & f(0) & 0 & 0 & 0 \\ f(2) & f(1) & f(0) & 0 & 0 \\ 0 & f(2) & f(1) & f(0) & 0 \\ 0 & 0 & f(2) & f(1) & f(0) \\ 0 & 0 & 0 & f(2) & f(1) \\ 0 & 0 & 0 & 0 & f(2) \end{bmatrix} \begin{bmatrix} g(1) \\ g(2) \\ g(3) \\ g(4) \\ g(5) \end{bmatrix}$$

(Here, we pretend $g(i) = 0$ for $i < 1$ and $i > 5$.)

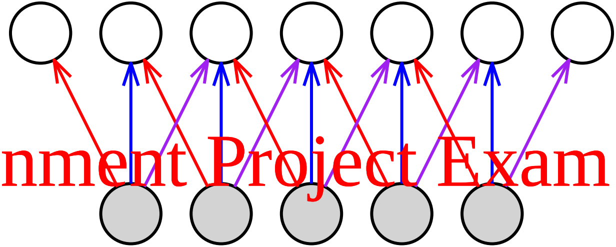

Figure 8: Convolutional layer

Convolutions III

- Similar for 2D inputs (e.g., images), except now sum over two indices
  - g is the input image
  - f is the filter
  - Lots of variations (e.g., padding, strides, multiple "channels")
- Use additional layers/activations to down-sample after convolution
  - E.g., max-pooling

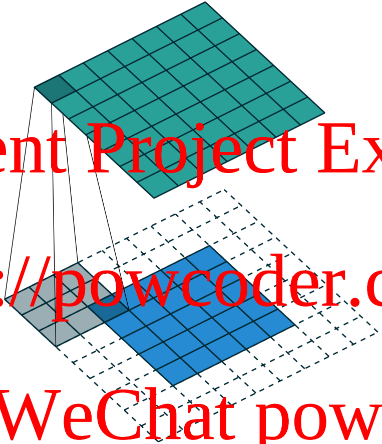

Figure 9: 2D convolution

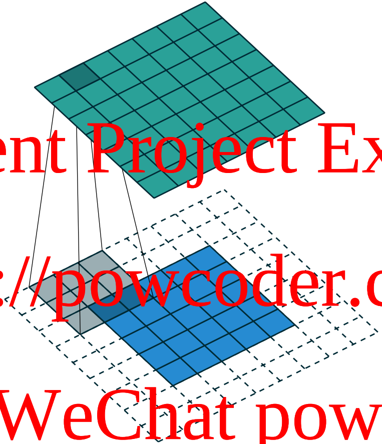

Figure 10: 2D convolution

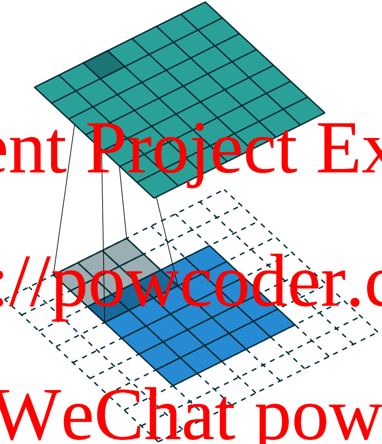

Figure 11: 2D convolution

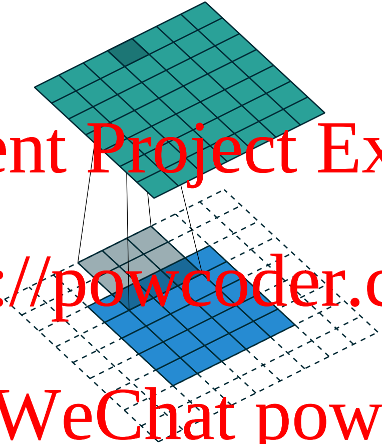

Figure 12: 2D convolution