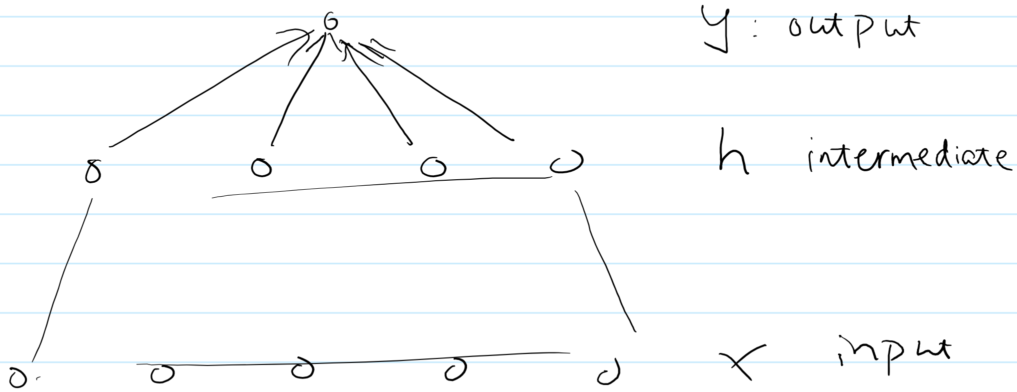
Lecture 20 Representer Theorems

Tuesday, November 8, 2016 4:32 PM

- idea: two-layer neural net ≈ combination of nonlinear functions

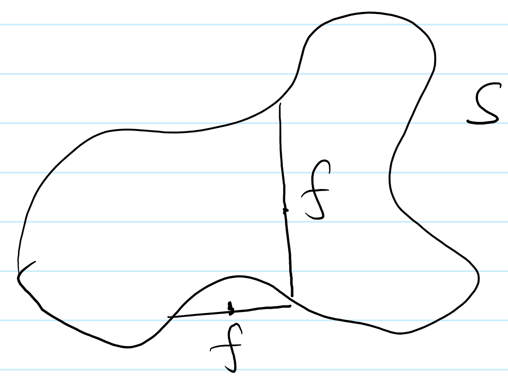


$$h_i = \sigma(w_i^T x + b_i)$$
$$y = c^T h$$

- Q: What functions can be approximated by a 2-layer net?
- Barron's representer theorem
  - for any function $f(x) \in L_1(\mathbb{R}^d)$, there is a function $g(x)$ representable by a two-layer neural net with k hidden units such that $\|f(x) - g(x)\|^2 \le O\left(\frac{1}{k}\right)$.
- $L_1(\mathbb{R}^d)$: set of functions whose Fourier coefficients have l_1-norm $= 1$.
  in particular: low frequency, low deg polynomials.
- abstraction of representer theorems
  - Given a set S of functions of norm at most 1.
    (think: S = set of functions representable by 1 hidden unit)
  - Given a function $f \in conv(S)$
  - There exists $c_1, c_2, \cdots, c_k$, $f_1, \cdots, f_k \in S$ s.t.

if we set $g(x) = \sum_{i=1}^{k} c_i f_i(x)$ then

$\|f - g\|^2 \le \frac{1}{k}$



- in general in d-dimensional space, if a point is in convex hull, can be represented by d+1 points, but functions have infinite dimensions.

- Proof: $f \in \mathrm{conv}(S)$ means there is a distribution D on S such that

$$\mathbb{E}_{h \sim D}[h(x)] = f(x)$$

sample $f_i(x) \sim D$,

we know: $\mathbb{E}[f_i(x)] = f(x)$

$$\mathbb{E}[\|f_i - f\|^2] \le 1$$

so if we let $g(x) = \frac{1}{k} \sum_{i=1}^{k} f_i(x)$

we know $\mathbb{E}[g(x)] = f(x)$

$$\mathbb{E}[\|g - f\|^2] = \frac{1}{k} \qquad \square$$

- problem: representer theorem is not constructive.

Know <u>existence</u> of a representation, but not how to find a representation.

- Greedy approach (Frank-Wolfe algorithm)

Suppose we currently have

$$g_k = \sum_{i=1}^{k} c_i f_i(x)$$

$$\| g_k - f \|^2 \le \frac{1}{k}$$

how to find g_{k+1} such that $\|g_{k+1} - f\|^2 \le \frac{1}{k+1}$?

Plan: let $g_{k+1} = \lambda g_k + (1-\lambda) f_{k+1}(x)$

λ, f_{k+1} : unknown

solve the optimization problem

$$\min_{\lambda, f_{k+1}} \|g_{k+1} - f\|^2$$

s.t. $\lambda \in [0,1]$

$f_{k+1} \in S$.

Claim: $\|g_{k+1} - f\|^2 \le \frac{1}{k+1}$

Proof: let $g_{k+1} = \lambda g_k + (1-\lambda) f_{k+1}(x)$

where $f_{k+1} \sim D$ (recall $\mathbb{E}_{h\sim D}[h] = f$)

$$\mathbb{E}[\| g_{k+1} - f\|^2] = \mathbb{E}[\| \lambda(g_k - f) + (1-\lambda)(f_{k+1} - f) \|^2]$$

$$= \lambda^2 \|g_k - f\|^2 + (1-\lambda)^2 \mathbb{E}[\|f_{k+1} - f\|^2]$$

no cross term because $\mathbb{E}[f_{k+1}] = f$

$$\le \frac{\lambda^2}{k} + (1-\lambda)^2$$

$$= \frac{(k+1)\lambda^2 - 2\lambda + 1}{k}$$

let $\lambda = \frac{1}{k+1}$ we have

$$\mathbb{E}[\|g_{k+1} - f\|^2] \le \frac{1}{k+1}.$$

- Problem: even the simpler problem of optimizing a single unit (f_{K+1}) is not easy.

This is still open for neural nets, but in special cases it can be solved / approximated.

Example: if $\sigma(x) = x^2$ (nonlinear), then finding optimal f_{K+1} can be solved by SVD.