Lecture 3

Universal apx continued

Ann.

- hw1 tomorrow
- typed notes ...
- tablet tweaks
- hybrid

Plan for next lectures:

1-4	shallow constructive apx
5-8	initialization / overparameterization
9-11	deeper topics / RNN transformer / distribution modeling
12-	opt
20 -	??

Last week: folklore constructive apx over $\mathbb{R}^1$, $\mathbb{R}^d$

Theorem ("universal approximation", Hornik-Stinchcombe-White '89, Leshno '93).

Let any $\sigma:\mathbb{R}\to\mathbb{R}$ continuous & not a polynomial. For any cont. $g:\mathbb{R}^d\to\mathbb{R}$, any $\varepsilon>0$ $\exists$ 2-layer σ-network $f:\mathbb{R}^d$ s.t. $|f(x)-g(x)|\le\varepsilon$ $\forall x\in[0,1]^d$.

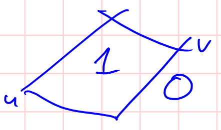
↑ choice of m

Remarks.

- e-Lip, 3-layer, avg dist.
- "celebrated", even though true for all ML models
- Bakes in exponential dependence on dimension
- false if m fixed or σ a polynomial

Lemma. Same statement, except $\sigma(r)=\exp(r)$.

Remark on proof. Recall our folklore proof need to apx $\prod_{i=1}^{d}\mathbb{1}[x_i\ge u_i]\,\mathbb{1}[x_i<v_i]$ ←



⇒ products are helpful

⇒ proof technique is a reduction to polynomials.

Proof. Define $\mathcal{F}:=\{x\mapsto a^T\exp(Vx):\overset{\text{varies}}{m\ge0},\ a\in\mathbb{R}^m,\ V\in\mathbb{R}^{m\times d}\}$.

Proof is complete if we can show $\mathcal{F}$ satisfies conditions of the Stone-Weierstrass (aka $\mathcal{F}$ is polynomial-like).

① $\mathcal{F}\ni f$ is continuous ✓

② $\forall x\in[0,1]^d$, $\exists f\in\mathcal{F}$, $f(x)\ne0$ (easy: $x\mapsto\exp(0^Tx)=1$). ✓

③ $\forall x\ne x'$, $\exists f$, $f(x)\ne f(x')$ (easy: $z\overset{f}{\mapsto}\underbrace{\exp(\langle -x',x-x'\rangle)}_{a}\exp(\langle z,\overbrace{x-x'}^{v}\rangle)$

$=\exp(\langle z-x',x-x'\rangle)$) ✓

note $f(x')=\exp(0)=1\ne\exp(\|x-x'\|^2)$.

④ $\mathcal{F}$ "closed under VS & poly operations": given $a^T\exp(Vx)$ & $u^T\exp(Wx)$, $b\in\mathbb{R}$, $c\in\mathbb{R}$

$$\underset{b\sum_{i=1}^m a_i\exp(v_i^Tx)}{b\,a^T\exp(Vx)}+\underset{c\sum_{j=1}^n u_j\exp(w_j^Tx)}{c\,u^T\exp(Wx)}=\underbrace{\begin{bmatrix}\mathbb{R}^m\{\ ba\\ \hline \mathbb{R}^n\{\ cu\end{bmatrix}}_{\mathbb{R}^{m+n};\ m+n}\exp\left(\begin{bmatrix}V\\ \hline W\end{bmatrix}x\right);$$

for products

$$\left[\sum_{i=1}^m a_i\exp(v_i^Tx)\right]\left[\sum_{j=1}^n u_j\exp(w_j^Tx)\right]=\sum_{ij}\underbrace{a_iu_j}_{\substack{\text{new}\\\text{outer}\\\text{weights}}}\exp(\langle v_i+w_j,x\rangle).$$ ✓ //

Remarks.

- ④ fails for polynomials of fixed degree.
- Going through proof in detail reveals exponential dependence on dim

[note to self: adaptive open problem empirical part & math part.]

To handle other activations:

① Given $\varepsilon>0$, g, pick $(a,V)\in\mathbb{R}^m\times\mathbb{R}^{m\times d}$ s.t. $\forall x\in[0,1]^d$ $|a^T\exp(Vx)-g(x)|\le\varepsilon/2$.

② Write $\underbrace{\exp(r)=\sum_{j=1}^k u_j\,\sigma(w_jr+b_j)}_{\text{univariate approximation; easier, maybe hard problem.}}$, define $f(x)=\sum_{i=1}^m a_i\sum_{j=1}^k u_j\,\sigma(w_jv_i^Tx+b_j)$. //

Apx. with infinite width

- popular in recent years
- can always convert to finite width (e.g., via sampling)
- can often write target with equality
- maybe helps with adaptivity

Recall our univariate approximation proof

$$\sum_{j=1}^m a_j\,\mathbb{1}[x\ge b_j]$$

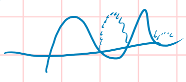
$$m=\left\lceil\frac{e}{\varepsilon}\right\rceil.$$

Proposition. Let $g:\mathbb{R}\to\mathbb{R}$ be continuously differentiable $g(0)=0$. Then, $\forall x\in[0,1)$

$$g(x)=\underbrace{\int_0^1 g'(x)\,\mathbb{1}[x\ge b]\,dx.}_{\text{"infinite width network"}}$$

Proof. By FTC

$$g(x)-g(0)=\int_0^x g'(b)\,db=\int_0^1 g'(b)\,\mathbb{1}[x\ge b]\,db.$$ //



Remarks.

- Sampling. Define $Z:=\int_0^1|g'(b)|$, and note $\frac{|g'(b)|}{Z}$ is a probability dist over $[0,1]$.
  - sample $b_j\sim(\cdot)$, define $a_j:=Z\,\mathrm{sgn}(g'(b_j))$;

  note via proposition that

$$\mathbb{E}\,a_j\,\mathbb{1}[x\ge b_j]=\int_0^1\underbrace{[Z\,\mathrm{sgn}(g'(b))]}_{a_j}\,\mathbb{1}[x\ge b]\underbrace{\frac{|g'(b)|}{Z}}_{\text{probability dist}}db$$

$$=\int_0^1\left(\mathrm{sgn}(g'(b))\cdot|g'(b)|\right)\mathbb{1}[x\ge b]\,\frac{Z}{Z}\,db$$

$$=\int_0^1 g'(b)\,\mathbb{1}[x\ge b]\,db=g(x),$$

  so this single node is an unbiased estimator of $f(x)$ $\forall x\in[0,1]$

- Sample m such nodes, define $f(x):=\frac{1}{m}\sum_j a_j\,\mathbb{1}[x\ge b_j]$.

  Sampling theorems (in notes) say $\mathbb{E}(f(x)-g(x))^2\le\frac{1}{m}\int$

  Define $f(x):=\frac{1}{m}\sum_{i=1}^m a_i\,\mathbb{1}[x\ge b_i]$, then $\mathbb{E}(f(x)-g(x))^2\le\frac{Z^2}{b}$;

  width does not pay for large flat regions (somewhat adaptivity).

Next time: we'll write multivariate continuous functions using Fourier transforms, # nodes will scale some function of the fourier ("Barron norm").