Approximation by ridge functions with weights in a specified set

David Stewart & Palle Jorgensen

University of Iowa
Mathematics

September 11, 2018

Extreme Learning Machines

Extreme Learning Machines are neural networks with one hidden layer where the training is only carried out on the weights in the *output*. The weights before the activation/sigmoidal functions are generated randomly are left unchanged.

Approximation by ridge functions

We consider approximating functions $f(\boldsymbol{x})$ for $\boldsymbol{x} \in [-1, +1]^m = J^m$ by functions in

$$V_{\mathcal{W}} := \text{span}\left\{ \boldsymbol{x} \mapsto \varphi(\boldsymbol{w}^T \boldsymbol{x}) \mid \varphi \in C(\mathbb{R}),\ \boldsymbol{w} \in \mathcal{W} \right\}$$

for $\mathcal{W}$ a specific subset in $\mathbb{R}^m$.

Questions:

- For what sets $\mathcal{W}$ is $V_{\mathcal{W}}$ dense in $C(J^m)$?
- If $V_{\mathcal{W}}$ is not dense in $C(J^m)$, how well can we approximate (nice) functions by functions in $V_{\mathcal{W}}$?

General framework

We work in a Banach space X; the approximating functions form a subspace $V \subset X$.

If $\overline{V} = X$ then every object in X can be approximated (arbitrarily well) by elements of V.

But if $\overline{V} \neq X$, then for every $\epsilon > 0$ there are functions $0 \neq f \in X$ where $\inf_{g \in V} \|f - g\|_X \geq (1 - \epsilon) \|f\|_X$. So

$$\sup_{f : \|f\|_X = 1} \inf_{g \in V} \|f - g\|_X \qquad \text{is either 0 or 1.}$$

Choose a Banach subspace Z compactly embedded in X and we look determine

$$m(V; Z, X) = \sup_{f:\|f\|_Z=1} \inf_{g\in V} \|f - g\|_X .$$

In our case we use:

- $X = C(J^m)$
- $V = V_{\mathcal{W}}$
- $Z = \mathrm{Lip}(J^m)$, the space of Lipschitz functions with semi-norm
$$|f|_Z = \sup_{\boldsymbol{x},\boldsymbol{y}\in J^m} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|_2}$$
(We can quotient out constant functions since $V_{\mathcal{W}}$ always contains these.)

We say $f \in X$ is *unapproximable* by V if

$$\|f\|_X \leq \|f - g\|_X \qquad \text{for all } g \in V.$$

For any $f \in X$ if $h \in V$ is the closest point in V to f then $f - h$ is unapproximable by V.

The existence of a closest point is assured if X is a reflexive Banach space, but generally false otherwise.

Separating Hyperplane Theorem

Theorem: If $C \subset X$ is closed and convex and $y \notin C$, then there is a $\mu \in X'$ and $b \in \mathbb{R}$ where

$$\begin{aligned} \langle y, \mu \rangle + b &> 0 \\ \langle z, \mu \rangle + b &\leq 0 \qquad \text{for all } z \in C \end{aligned}$$

Specifically, if C is a closed subspace of X, then ν satisfies

$$\begin{aligned} \langle y, \mu \rangle &> 0 \\ \langle z, \mu \rangle &= 0 \qquad \text{for all } z \in C. \end{aligned}$$

Cybenko's universal approximation result

George Cybenko's paper from 1989 shows that if (for example) $\sigma(u) = \tanh(u)$ then

$$\overline{\operatorname{span}\left\{\, \boldsymbol{x} \mapsto \sigma(\boldsymbol{w}^T\boldsymbol{x} + b) \mid \boldsymbol{w} \in \mathbb{R}^m,\ b \in \mathbb{R} \,\right\}} = C(J^m).$$

The proof uses the Separating Hyperplane Theorem

Note: $C(J^m)' = \mathcal{M}(J^m)$, the space of signed Borel measures on J^m with bounded variation and $\langle g, \mu \rangle = \int g(\boldsymbol{x})\, d\mu(\boldsymbol{x})$.

For the μ in the Separating Hyperplane Theorem

$$\int \sigma(a\boldsymbol{w}^T\boldsymbol{x} + b)\, d\mu(\boldsymbol{x}) = 0 \qquad \text{for all } a, b \in \mathbb{R}$$

so we can show that

$$0 = \mu^{\boldsymbol{w}}(F) := \mu\left(\left\{\, \boldsymbol{x} \mid \boldsymbol{w}^T\boldsymbol{x} \in F \,\right\}\right) \qquad \text{for all Borel } F \subset \mathbb{R}.$$

Fourier Transforms

A Borel measure μ with bounded variation has a Fourier Transform

$$\widehat{\mu}(\boldsymbol{\xi}) = \int_{\mathbb{R}^m} e^{-i\boldsymbol{\xi}^T\boldsymbol{x}}\, d\mu(\boldsymbol{x}).$$

Note that for the μ from the SHT

$$\begin{aligned}\widehat{\mu}(s\boldsymbol{w}) &= \int_{\mathbb{R}^m} e^{-is\boldsymbol{w}^T\boldsymbol{x}}\, d\mu(\boldsymbol{x}) \\ &= \int_{\mathbb{R}} e^{-ist}\, d\mu^{\boldsymbol{w}}(t) = \widehat{\mu^{\boldsymbol{w}}}(s) = 0.\end{aligned}$$

For Cybenko's result, this is true for all $\boldsymbol{w} \in \mathbb{R}^m$ so $\widehat{\mu}(\boldsymbol{\xi}) = 0$ for all $\boldsymbol{\xi}$, and so $\mu = 0$ contradicting the SHT.

Thus there is no f in $C(J^m)$ that is not in

$$\overline{\operatorname{span}\left\{\, \boldsymbol{x} \mapsto \sigma(\boldsymbol{w}^T\boldsymbol{x} + b) \mid \boldsymbol{w} \in \mathbb{R}^m,\ b \in \mathbb{R} \,\right\}}$$

What about specific (finite) $\mathcal{W}$?

What about spans of ridge functions

$$V_{\mathcal{W}} := \operatorname{span}\left\{ \boldsymbol{x} \mapsto \varphi(\boldsymbol{w}^T \boldsymbol{x}) \mid \varphi \in C(\mathbb{R}),\ \boldsymbol{w} \in \mathcal{W} \right\} ?$$

We can get *lower bounds* on how badly a Lipschitz function f can be approximated by $V_{\mathcal{W}}$ as follows: Pick a measure μ with support in J^m where $\mu^{\boldsymbol{w}} = 0$ for every $\boldsymbol{w} \in \mathcal{W}$. Then look for a function f where $\langle f, \mu \rangle = \|f\|_\infty \|\mu\|_{\mathcal{M}} \neq 0$.

The measure μ has to satisfy $\widehat{\mu}(t\boldsymbol{w}) = 0$ for all $t \in \mathbb{R}$ and $\boldsymbol{w} \in \mathcal{W}$.

Example: $\mathcal{W} = \{\boldsymbol{e}_1, \boldsymbol{e}_2\}$.

The Fourier transform $\widehat{\mu}(t\boldsymbol{e}_1) = \widehat{\mu}(t\boldsymbol{e}_2) = 0$ for all $t \in \mathbb{R}$.

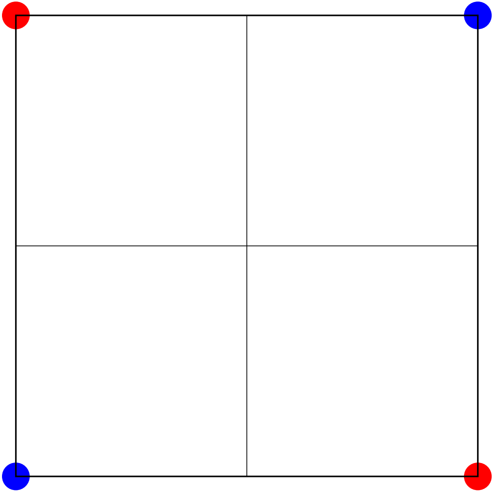
Since the Fourier transform of $\delta_{\boldsymbol{v}} = \delta(\cdot - \boldsymbol{v})$ is $\widehat{\delta_{\boldsymbol{v}}}(\boldsymbol{\xi}) = \exp(-i\boldsymbol{\xi}^T\boldsymbol{v})$, so we look for $\widehat{\mu}(\boldsymbol{\xi})$ that involves complex exponentials $\exp(-i\boldsymbol{\xi}^T\boldsymbol{v})$ for $\boldsymbol{v} \in J^m$.

Note: $\widehat{\mu}(\boldsymbol{\xi})$ is complex analytic everywhere (entire) so we can look for Taylor series. So...

$$\widehat{\mu}(\boldsymbol{\xi}) = c\,\xi_1\,\xi_2 + \cdots$$

We can put $\widehat{\mu}(\boldsymbol{\xi}) = \left(e^{-i\xi_1} - e^{+i\xi_1}\right)\left(e^{-i\xi_2} - e^{+i\xi_2}\right)$, $(c = (-2i)^2)$.

Here μ is a sum of δ-functions at (x_1, x_2) with each $x_i = \pm 1$, and the weights at each of these points is $x_1 x_2$.



$= \operatorname{supp} \mu_-$, $= \operatorname{supp} \mu_+$

Now we need to find a function f with, say, $\|f\|_\infty = 1$ and $\langle f, \mu \rangle = \|f\|_\infty \|\mu\|_{\mathcal{M}}$.
Since μ is a sum of (scaled) δ-functions, we can choose $f(\boldsymbol{v}) = \pm 1$ at each of these points, choosing the sign of $f(\boldsymbol{v})$ to match the sign of the scaling of the associated δ-function.

We can put $f(x_1, x_2) = x_1 x_2$

Note that this f is Lipschitz with Lipschitz constant $\sqrt{2}$.

In general: if $A = \operatorname{supp} \mu_+$ and $B = \operatorname{supp} \mu_-$ we have a Lipschitz function f where $f(\boldsymbol{x}) = +1$ for $\boldsymbol{x} \in \operatorname{supp} \mu_+$ and $f(\boldsymbol{x}) = -1$ for $\boldsymbol{x} \in \operatorname{supp} \mu_-$:

$$f(\boldsymbol{x}) = \frac{d(\boldsymbol{x}, B) - d(\boldsymbol{x}, A)}{d(\boldsymbol{x}, B) + d(\boldsymbol{x}, A)}$$

$$\operatorname{Lip} f = \frac{2}{\min_{\boldsymbol{a} \in A, \boldsymbol{b} \in B} \|\boldsymbol{a} - \boldsymbol{b}\|}$$

Non-trivial lower bounds

What if there are many vectors in $\mathcal{W}$? How many do we need to get a reasonable approximation?

Choose

$$\boldsymbol{z}_{k+1} \perp \left\{\boldsymbol{z}_1, \dots, \boldsymbol{z}_k, \boldsymbol{w}_{s(k)}, \dots, \boldsymbol{w}_{s(k+1)-1}\right\}$$

where $s(k+1) = s(k) + m - 1 - k$ for $k = 1, 2, \dots, m-2$. Put

$$\widehat{\mu}(\boldsymbol{\xi}) = c \prod_{k=1}^{m-1} \left(\exp(-i\boldsymbol{z}_k^T \boldsymbol{\xi}) - \exp(+i\boldsymbol{z}_k^T \boldsymbol{\xi})\right) \qquad \text{so}$$

$$\mu = c \sum_{\boldsymbol{u} \in \{\pm 1\}^{m-1}} \left(\prod_{k=1}^{m-1} u_k\right) \delta_{\sum_{k=1}^{m-1} u_k \boldsymbol{z}_k}.$$

Choose $\|\boldsymbol{z}_k\|_2 = 1/\sqrt{m-1}$ so that $\left\|\sum_k u_k \boldsymbol{z}_k\right\|_\infty \leq 1$ for all $\boldsymbol{u} \in \{\pm 1\}^{m-1}$.

$$\operatorname{supp} \mu_+ = \left\{ \sum_k u_k \boldsymbol{z}_k \mid \boldsymbol{u} \in \{\pm 1\}^{m-1} \ \& \ \# \{k \mid u_k > 0\} \text{ is even} \right\}$$

$$\operatorname{supp} \mu_- = \left\{ \sum_k u_k \boldsymbol{z}_k \mid \boldsymbol{u} \in \{\pm 1\}^{m-1} \ \& \ \# \{k \mid u_k > 0\} \text{ is odd} \right\}$$

Thus given $\mathcal{W}$ with $|\mathcal{W}| \leq \frac{1}{2} m(m-1)$ there is a function f of Lipschitz constant $\sqrt{m-1}$ with $\|f\|_\infty = 1$ that is unapproximable by span $\left\{\boldsymbol{x} \mapsto \varphi(\boldsymbol{w}^T \boldsymbol{x}) \mid \varphi \in C(\mathbb{R}),\ \boldsymbol{w} \in \mathcal{W}\right\}$.