Neural Network Theory

Lecturer: Prof. Helmut Elbrächter
Arranged by Tongyu Lu

Chapter 1: approximating functions using sigmoid basis

Introduction: neuron as basis

Consider a single-layer neural network $G(x)$ (which has a d-dimensional vector input $x \in \mathbb{R}^d$) with sigmoid-like activation function $\sigma: \mathbb{R} \to [0,1]$:

$$G(x) = \sum_{j=1}^{N} a_j \sigma\left(y_j^T x + \theta_j\right)$$

where $y_j \in \mathbb{R}^d, \theta_j \in \mathbb{R}$
It is OK to treat each term $\sigma\left(y_j^T x + \theta_j\right)$ as a basis defined by $\left(y_j, \theta_j\right)$, which could be written as

$$K_{y_j,\theta_j}(x) = \sigma\left(y_j^T x + \theta_j\right)$$

Then, we find ourself actually treating a single-layer neural network as a linear combination of basis:

$$G(x) = \sum_{y,\theta} a_{y,\theta} K_{y,\theta}(x)$$

This is somehow like the Fourier transform:

$$x(t) = \int_f \hat{x}_f e^{2\pi\sqrt{-1}ft} df$$

where each basis $\phi_f(t) = e^{2\pi\sqrt{-1}ft}$, and there is uncountable-many of them. We know that $\left\{\phi_f(t): f \in \mathbb{R}\right\}$ is a complete set of bases for functions (vectors) which are subject to Dirichlet conditions.
In the neural network context, we also have uncountable-many bases $K_{y,\theta}(x) = \sigma(y^T x + \theta)$, but the magic is that: we only select finite-many of those bases and we can the approximate a given function in $\mathbb{R}^d$. (That is why we use j to denote y and θ).
We want to see whether $\left\{K_{y,\theta}(x): y \in \mathbb{R}^d, \theta \in \mathbb{R}\right\}$ is a complete set of bases for continuous functions (vectors) whose domain is $\mathbb{R}^d$. And further, we wish to show that we could only select finite-many of those bases in order to approximate a given continuous function.
Our following analysis is first restricted within such a Hilbert space $(X, \|\cdot\|_\infty)$, where $X = \{f: I_d \to \mathbb{R}, f \in C(I_d)\}$, $I_d = [0,1]^d$. The norm is defined as $\|f\|_\infty = \sup_{x \in I_d} |f(x)|$. (In the following contexts, we use $\|\cdot\|$ instead of $\|\cdot\|_\infty$) The inner product between f and g is defined as $\langle f, g \rangle = \int_{I_d} f(x) g(x) dx$.

Question: why restricting x within I_d? Why not use $\mathbb{R}^d$?
Question: the normed space has infinity norm, which defines the inner product as $\langle x, y \rangle = (\|x + y\|^2 - \|x - y\|^2)/4$. But the inner product is defined as an integral. Why shall we define integral as the inner-product?

Is a single-layer neural network capable to represent arbitrary function?

The answer for this question is as follows:

Theorem (single-layer approximation): let $\sigma: \mathbb{R} \to \mathbb{R}$ be a continuous discriminatory (?) function, and let space $U(\sigma) = \text{span}\{K_{y,\theta}(x): y \in \mathbb{R}^d, \theta \in \mathbb{R}\}$. Then, $U(\sigma)$ is dense (?) in $C(I_d)$.

This statement may look intimidating at first glance. This might because there are two unfamiliar (maybe not) concepts: discriminatory property and density property.

The density property is easier to explain: subset M is dense in X, if $\bar{M} = X$.

Recall that the concept of closure: $\bar{M} = \{x: B_\epsilon(x) \cap M \neq \emptyset, \forall \epsilon > 0\}$ ($B_\epsilon(x)$ means a ball of radius ϵ centered by x).

In other words, if M is dense in X, then each point in X is neighbored by points from M.

Formally, the statement "M is dense in X" means that "$\forall x \in X, \forall \epsilon > 0, \exists y \in M \; s.t. \; \|x - y\| < \epsilon$".

Therefore, the statement "$U(\sigma)$ is dense in $C(I_d)$" means that "$\forall f \in C(I_d), \forall \epsilon > 0, \exists G \in U(\sigma) \; s.t. \; \|G - f\| < \epsilon$".

This is the explanation for why this theorem is about approximation.

Next, what is the discriminatory property? It is to be introduced in the next section. And after having an idea of the following concepts, we are ready to prove the theorem.

Preparation knowledge: functional-analysis-related concepts

(It is recommended that readers have known a few concepts about linear functional. But anyway, let us go on.)

- **Discriminatory property and completeness**

  Definition (discriminatory property): σ is discriminatory if statement "given $\forall g \in C(I_d)$, then $\langle g, K_{y,\theta} \rangle = 0, \forall y \in \mathbb{R}^d, \theta \in \mathbb{R}$" implies $g = 0$.

  This means that: the bases defined using σ("affine") consist of a complete set of bases in $C(I_d)$. $U(\sigma) = \{K_{y,\theta}: y \in \mathbb{R}^d, \theta \in \mathbb{R}\}$ is complete because there is no other vector orthogonal to $U(\sigma)$ except zero vector $g = 0$.

  In the measure theory language, the definition is stated as such:

  Definition (discriminatory property): σ is discriminatory if statement "given $\forall \mu \in M(I_d)$, then $\int_{I_d} K_{y,\theta}(x) d\mu(x) = 0, \forall y \in \mathbb{R}^d, \theta \in \mathbb{R}$" implies $\mu = 0$.

  Actually, we could get the equivalent bridge by setting $d\mu(x) = g(x)dx$

- **Hahn-Banach Theorem**

  This theorem says that:

  Let $(M, \|\cdot\|)$ be a subspace of Banach space $(X, \|\cdot\|)$. Let l be a bounded linear functional (b.l.f.) on $(M, \|\cdot\|)$ (and it shall be bounded by a sublinear functional actually). Then, $\exists$ b.l.f. $L: X \to \mathbb{R}$ on $(X, \|\cdot\|)$ that is an extension of l and satisfies $\|L\| = \|l\|$.

  Please refer to any functional analysis book for the proof of this theorem. Here in our discussion, this theorem is applied to proving the following theorem:

  Theorem (a condition of density): Let $(M, \|\cdot\|)$ be a subspace of Banach space $(X, \|\cdot\|)$. Then, statement "$\forall$b.l.f. $F: X \to \mathbb{R}$, such that if $F(x) = 0, x \in M$, then $F(x) = 0, x \in X$" implies that "M is dense in X", or that $\bar{M} = X$.

  Proof: to prove this, we suppose that $\bar{M} \subset X$. Then, $\exists x_0 \in X \backslash \bar{M}$. Then, $\exists$ b.l.f. $F: X \to \mathbb{R}$ such that $F(M) = \{0\}$ but $F(x_0) \neq 0$. (This is a corollary of H.B. theorem, which shall be proved.) However, this contradicts with the assumption that $F(x_0)$ should be zero.

- **Riesz Representation Theorem**

  This theorem says that: given Hilbert space X, $\forall$ b.l.f. $F: X \to \mathbb{R}$, $\exists$unique $g \in X$ such that $F(x) = \langle x, g \rangle, \forall x \in X$.
  In our context, it gives us that $\forall$ b.l.f. $F: C(I_d) \to \mathbb{R}$, $\exists$unique $g \in C(I_d)$ such that $F(f) = \langle f, g \rangle, \forall f \in C(I_d)$.

  In language of measure theory: $\forall$ b.l.f. $F: C(I_d) \to \mathbb{R}$, $\exists$unique $\mu \in M(I_d)$ such that $F(f) = \int_{I_d} f d\mu, \forall f \in C(I_d)$.

  Please refer to any functional analysis book for the proof of this theorem.

Proof for the single-layer approximation theorem

Restate the single-layer approximation theorem: let $\sigma: \mathbb{R} \to \mathbb{R}$ be a continuous discriminatory function, and let space $U(\sigma) = \text{span}\{K_{y,\theta}(x): y \in \mathbb{R}^d, \theta \in \mathbb{R}\}$. Then, $U(\sigma)$ is dense in $C(I_d)$.

Hopefully, we have already understood this statement. Now comes the proof:
By definition, $U(\sigma)$ is a linear subspace of $C(I_d)$. This is because each $K_{y,\theta}(x)$ is orthogonal to each other and is member of $C(I_d)$.
Let $L: C(I_d) \to \mathbb{R}$ be a b.l.f. and $L\big(U(\sigma)\big) = \{0\}$. Then, by Riesz representation, $\exists$unique $g \in C(I_d)$ such that $L(f) = \langle f, g \rangle, \forall f \in C(I_d)$. Since $L\big(U(\sigma)\big) = \{0\}$, we have $L\big(K_{y,\theta}\big) = \langle K_{y,\theta}, g \rangle = 0, \forall y \in \mathbb{R}^d, \theta \in \mathbb{R}$. According to the assumption that $\sigma: \mathbb{R} \to \mathbb{R}$ is discriminatory, we have that $g = 0$, which implies that $L\big(C(I_d)\big) = \{0\}$.
In other words, we reached the statement that "$L\big(U(\sigma)\big) = \{0\}$ implies $L\big(C(I_d)\big) = \{0\}$". Therefore, according to the discussion in the previous section, $U(\sigma)$ is dense in $C(I_d)$, which finishes the proof.

Now, we may wonder: what kind of function σ have the discriminatory property?
And why are we able to pick only finite-many bases (neurons) to approximate the target function?