Approximation by superposition of sigmoidal functions

1.1 Discriminatory functions are dense

Setting. For our show, we shall live in the compact metric space $X = I_n = [0,1]^n$ with its usual metric. We shall look at approximations to functions in the Banach space $(C(I_n), \|\cdot\|_u)$. We shall let the signed Radon measures on I_n be denoted by $M(I_n)$. For a fix $\sigma : \mathbb{R} \to \mathbb{R}$, we shall be interested in testing whether the set

$$S_\sigma = \left\{ \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j) \, : \, y_j \in \mathbb{R}^n, \, \theta_j \in \mathbb{R}, \, N \in \mathbb{Z}^+ \right\}$$

is dense in $C(I_n)$.

Remark. The measures in $M(I_n)$ are automatically finite, since I_n is compact.

Definition 1.1. A function $\sigma : \mathbb{R} \to \mathbb{R}$ is said to be **discriminatory** if the only measure $\mu \in M(I_n)$ such that

$$\int_{I_n} \sigma(y^T x + \theta) d\mu(x) = 0 \qquad \forall y \in \mathbb{R}^n, \, \theta \in \mathbb{R}$$

is $\mu = 0$.

Example 1.2. The zero function is not discriminatory. A function which is almost-everywhere zero, with respect to the Lebesgue measure is not discriminatory.

Example 1.3. It is not true that, given a measure $\mu \in M(I_n)$ that any function f that is zero almost-everywhere will be automatically discriminatory. Let $n = 1$, put $\mu = \lambda_{\mathbb{R}^+}$ and let $f = \chi_{\mathbb{R}^-}$. Observe that the hypothesis for the definition has not been met by this measure. We have to observe that such hypothesis is met by measures which have a form of translation invariance property.

Definition 1.4. A function $\sigma : \mathbb{R} \to \mathbb{R}$ is **sigmoidal** if

$$\lim_{x \to \infty} \sigma(x) = 1 \qquad \lim_{x \to -\infty} \sigma(x) = 0$$

Example 1.5. The function $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoidal function of excellence to computer scientists and statisticians.

We park this definition to obtain our first nice result.

Theorem 1.6. *Let σ be any continuous discriminatory function. Then $\overline{S_\sigma} = C(I_n)$.*

Proof. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous discriminatory function. Arguing by contradiction, suppose $\overline{S_\sigma} \subset C(I_n)$. Observe that $\overline{S_\sigma}$ is a closed proper vector subspace of $C(I_n)$.

By the Hahn-Banach theorem, there exists a bounded linear function $L : C(I_n) \to \mathbb{R}$ such that $L|_{\overline{S_\sigma}} = 0$ but $L \neq 0$. By the Riesz Representation theorem for bounded linear functionals, there exists a unique $\mu \in M(I_n)$ such that

$$L(h) = \int_{I_n} h(x) d\mu(x)$$

for all $h \in C(I_n)$. Notice that $\sigma_{y,\theta}(x) = \sigma(y^T x + \theta) \in \overline{S_\sigma}$ for any choice of $y \in \mathbb{R}^n, \theta \in \mathbb{R}$. Thus,

$$L(\sigma_{y,\theta}) = \int_{I_n} \sigma(y^T x + \theta) d\mu(x) = 0 \qquad \forall y, \theta$$

But since σ was assumed to be discriminatory, we must have that $\mu = 0$, and thus $L = 0$. This contradicts the HB theorem, so that $\overline{S_\sigma} = C(I_n)$. ■

We specialise this result to one particular class of functions.

Theorem 1.7. *Any bounded measurable sigmoidal function, σ, is discriminatory. A fortiori, any continuous discriminatory function is discriminatory.*

Proof. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a bounded measurable sigmoidal function. We first observe that:

$$\sigma(\lambda(y^T x + \theta) + \varphi) \begin{cases} \to 1 & y^T x + \theta > 0 \text{ as } \lambda \to \infty \\ \to 0 & y^T x + \theta < 0 \text{ as } \lambda \to \infty \\ = \sigma(\varphi) & y^T x + \theta = 0 \quad \forall \lambda \end{cases}$$

Thus, for any sequence $(\lambda_k)_{k=1}^{\infty} \subset \mathbb{R}$ with $\lambda_k \to +\infty$ (in the extended sense) we have that $\sigma_{\lambda_k}(x) = \sigma(\lambda_k(y^T x + \theta) + \varphi)$ converges pointwise and boundedly to

$$\gamma(x) = \begin{cases} 1 & y^T x + \theta > 0 \\ 0 & y^T x + \theta < 0 \\ \sigma(\varphi) & y^T x + \theta = 0 \end{cases}$$

Now, to show that σ is discriminatory, we shall let $\mu \in M(I_n)$ be a measure such that $\int_{I_n} \sigma(y^T x + \theta) d\mu(x) = 0$ for all $y \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$. For notational convenience, given y, θ, define the hyperplane $\Pi_{y,\theta} = \{x \in I_n : y^T x + \theta = 0\}$ and the open half-space $H_{y,\theta} = \{x \in I_n : y^T x + \theta > 0\}$. We may then compute:

$$\begin{aligned} 0 &= \lim_{k\to\infty} \int_{I_n} \sigma_{\lambda_k}(x) d\mu(x) \\ &= \int_{I_n} \lim_{k\to\infty} \sigma_{\lambda_k}(x) d\mu(x) && \text{(LDCT)} \\ &= \int_{I_n} \gamma(x) d\mu(x) \\ &= \sigma(\varphi)\mu(\Pi_{y,\theta}) + \mu(H_{y,\theta}) \qquad (\dagger) \end{aligned}$$

for any choice of $y \in \mathbb{R}^n, \theta, \varphi \in \mathbb{R}$.

Now, fix $y \in \mathbb{R}^n$ and for any bounded measurable function h put $F_y : L^\infty(\mathbb{R}) \to \mathbb{R}$, with $F_y(h) = \int_{I_n} h(y^T x) d\mu(x)$. Since μ is finite, F_y is a bounded linear functional. Put $h = \chi_{[\theta,\infty)}$ and compute

$$F_y(h) = \int_{I_n} \chi_{[\theta,\infty)}(y^T x) d\mu(x) = \mu(\Pi_{y,-\theta}) + \mu(H_{y,-\theta}) = 0$$

by ($\dagger$). Likewise, we may put $h = \chi_{(\theta,\infty)}$ to get $F_y(h) = 0$. Using linearity, we get that for any interval I, we have $F_y(\chi_I) = 0$. Thus, for any linear combination of indicators of intervals (any step function), say s, we have that $F_y(s) = 0$. Since step functions approximate simple functions, and simple functions are dense in L^∞, we have that $F_y = 0$. In particular, for the functions $s(x) = \sin(x)$ and $c(x) = \cos(x)$ we have that

$$0 = F_y(c + is) = \int_{I_n} \cos(y^T x) + i \sin(y^T x) d\mu(x) = \int_{I_n} \exp(iy^T x) d\mu(x) = \hat{\mu}$$

for any y. That is, the Fourier transform of μ is zero, and thus μ itself is zero. Hence, σ is discriminatory.

1.2 Applications to learning theory

In the setting of deep learning, we may be interested in learning parameters to approximate any continuous function via sigmoidal functions. A useful corollary of the above results is:

Theorem 1.8. *Let σ be any continuous sigmoidal function. Then $\overline{S_\sigma} = C(I_n)$.*

We can say more. Some problems in learning theory are not about regression, but also about classification. Let $(I_n, \mathcal{B}(I_n), \lambda_n)$ be the Lebesgue measure space on I_n. Let $P_1, \dots, P_k$ be a finite Borel partition of I_n. Define the decision function f by

$$f(x) = j \iff x \in P_j$$

The question posed in learning theory is whether we can approximate this decision function with a single-layer network. The answer is below:

Theorem 1.9. *Let σ be a continuous sigmoidal function. Let f be a decision function for a finite Borel partition of I_n. For any $\epsilon > 0$, there exists a $G(x) \in S_\sigma$ and a compact set $K \subseteq I_n$ such that $\mu(I_n \setminus D) < \epsilon$ and $|G(x) - f(x)| < \epsilon$ for $x \in K$.*

Proof. Let $\epsilon > 0$. Observe that σ is measurable in a finite measure space. By Lusin's theorem, for the given ϵ, there exists a compact set $K \subset I_n$ such that $h = f|_K$ is continuous and $\lambda(I_n \setminus K) < \epsilon$. Since $h \in C(K)$, we may find $G(x) \in S_\sigma$ such that $|G(x) - h(x)| = |G(x) - f(x)| < \epsilon$ for all $x \in K$. ■

Moral. The total measure of incorrectly classified points can be made arbitrarily small.

1.3 Extensions to other activation functions

Theorem 1.10. *Let μ be a Radon measure. The set $S_\sigma / \sim_\mu$ is dense in $L^1(I_n, \mathcal{B}(I_n), \mu)$.*

Proof. Follows since $C_c(X) / \sim_\mu$ is dense in $L_p(\mu)$. ■