MA 751
M. Kon

Problem Set 4
Due Thurs. 2/24/22

Note that the coming week will have no Tuesday class, and the Monday discussion section will be held on Tuesday because of the changed schedule.

Lectures 7, 8

Study of neural networks for high dimensional approximation predated machine learning, and has now been incorporated into the area. The mathematical models that they provide are a natural extension of the classes of approximators we have considered. They are currently enjoying increased interest in the context of deeper multilayer networks.

Reading: 11.1-11.8, class material

Problems:

1. Newer activation functions: Consider a neural network of the type described in class, with activations x_i for the k neurons in the first layer, y_j for the m neurons in the second layer, and q for the single neuron in the third layer. Assume that k = 3, m = 3. Assume that the activation function has the form $H(x) = \frac{1}{\pi}\tan^{-1}x + 1/2$.

(a) Let $q = \widehat{f}(x)$ (with $\mathbf{x} = (x_1,x_2,x_3)$) be the function which gives the activation of the output neuron q in terms of the input $\mathbf{x}$. Give the general form of $\widehat{f}(\mathbf{x})$ in terms of the function H and any appropriate constants (i.e., $V^{(j)}$, θ_j, w_j) determined by the network.

(b) Fix values of the above constants to any values you like, and for the values $x_2 = 0$ and $x_3 = 1$, sketch the output q as a function of x_1.

(c) Show that for k = 1 and m fixed, if H(x) = cos x, then for appropriate choices of the constants the function $\widehat{f}(x)$ can approximate any desired input-output function f(x) in $L^2[0,\pi]$ to within any accuracy $\epsilon > 0$, (i.e., $\| f(\mathbf{x}) - \widehat{f}(\mathbf{x}) \|_2 < \epsilon$) if m (which can depend on ϵ) is sufficiently large. What familiar problem does this reduce to in this case?

2. Changing error measures: Let $K \subseteq \mathbb{R}^k$ be a compact subset. Suppose that a neural network is able to compute a certain class of continuous functions $\mathcal{B}$ on $\mathbb{R}^k$ with the property that given any function $f(x) \in C(K)$ (i.e., a continuous function on K) together with an $\epsilon > 0$, there exists a $g \in \mathcal{B}$ such that

$$\|f - g\|_\infty < \epsilon, \tag{1}$$

where for any function h,

$$\|h\|_\infty \equiv \sup_{\boldsymbol{x}\in K} |h(\boldsymbol{x})|.$$

Now let μ be a Borel measure on K. Assuming that (1) holds as stated above, show that (1) above must still then hold if we replace the $\|\cdot\|_\infty$ norm with the norm $\|\cdot\|_p$ for any $1 \le p < \infty$, where by definition

$$\|h\|_p = \left(\int_K |h(\boldsymbol{x})|^p d\mu(\boldsymbol{x})\right)^{1/p}.$$

For notions involving measures you can refer to the introductory probability lecture (see course web page). Note also that if $f(\boldsymbol{x})$ is a real-valued continuous function on a set $K \subset \mathbb{R}^k$ with finite measure $\mu(K)$ then

$$\int_K f(\boldsymbol{x}) d\mu(\boldsymbol{x}) \leq ||f||_\infty \mu(K). \tag{1a}$$

Try proving (1a) either for a general measure μ, or if you like just for the case of standard Lebesgue measure on $K = [0, 1]$ (i.e. in 1 dimension).

3. Neural networks with more than one output neuron:

Consider a neural network as developed in class, with k neurons with activations x_i in the first layer, n neurons with activations y_i in the second layer, and m neurons with activations q_i in the third layer.

In class we have considered the case m = 1, and Funahashi's theorem stated that it is possible to approximate any function $f(\boldsymbol{x}) = f(x_1,x_2,\ldots,x_k)$: $\mathbb{R}^k \rightarrow \mathbb{R}$ (which represents the desired output of the single output neuron) with the neural net input-output (i-o) function

$$\widehat{f}(\boldsymbol{x}) = \sum_{j=1}^{n} w_j \, H(V^j \cdot \boldsymbol{x} - \theta_j), \tag{2}$$

where H is a non-constant nondecreasing function, if the constants w_j, θ_j, and a collection of vectors $V^1, V^2, \ldots$ are chosen properly.

To review this, the vector $\boldsymbol{x} = (x_1,x_2,\ldots,x_k)$ represented the activation levels of neurons in the first layer, and $q = f(x_1,x_2,\ldots,x_k)$ represented the activation of a *single* neuron in the third layer (i.e., we set m = 1 there). We assumed that w_j represent connection strengths from each neuron in the second layer to the single neuron in the third layer, and V^j is the vector whose i^{th} entry is the connection strenth from neuron x_i in the first layer to neuron y_j in the second layer. We showed that the neural net which we constructed would, given an input $\boldsymbol{x}$, yield an output q (in the output neuron) given by the right side of (2), which is supposed to be a good approximation of the desired output $f(\boldsymbol{x})$ on the left side.

Show that this result also allows us to generalize to the situation with m neurons $q_1, \ldots, q_m$ in the third layer, where m > 1. That is, given a function $\boldsymbol{f} : \mathbb{R}^k \rightarrow \mathbb{R}^m$ show the new network (now with m output neurons) can compute a function $\widehat{\boldsymbol{f}}(\boldsymbol{x})$ such that $||\widehat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})||_\infty < \epsilon$, for any required $\epsilon > 0$. As usual, the l component $\widehat{f}_l(\boldsymbol{x})$ of $\widehat{\boldsymbol{f}}(\boldsymbol{x})$ will be computed by the network as the activation q_l of the l^{th} output neuron. Here for any function $\boldsymbol{f} : \mathbb{R}^k \rightarrow \mathbb{R}^m$ on a set $K \subset \mathbb{R}^k$, we will define

$$||\boldsymbol{f}||_\infty \equiv \max_l \sup_{\boldsymbol{x} \in K} |f_l(\boldsymbol{x})|,$$

where $f_l(\boldsymbol{x})$ is the l^{th} component of $\boldsymbol{f}(\boldsymbol{x})$.

4. Recall that Funahashi proved that any continuous function on a compact set $K \subset \mathbb{R}^k$ can be uniformly approximated by a neural network of the form

$$\widehat{f}(\boldsymbol{x}) = \sum_k w_j H(V^j \cdot \boldsymbol{x} - \theta_j), \tag{3}$$

if H is monotone increasing. Prove the Corollary to Funahashi's theorem, namely, that functions of the form (3) are then dense in $L^p(K)$ for $1 \leq p < \infty$. Note that given a set C of functions (e.g. continuous functions) and a subset C' of these functions (e.g. the set of possible neural network functions), the density of the smaller set C' in the larger one C has been defined in the notes. How does our approximability within any ϵ of any $f \in C$ by some $f' \in C'$ prove that C' is dense in C.

5. Problem **11.3** in Hastie, Tibshirani