Homework 5

Machine Learning: Foundations and Applications — Paul J. Atzberger
MATH CS 120 — http://atzberger.org/

1. (Neural Networks) Consider a basic Multi-Layer Perceptron (MLP) with two inputs x_1, x_2, single output y, and a hidden layer with n units h_i. Corresponding to this MLP is the hypothesis space $\mathcal{H} = \{q : \mathbb{R}^2 \to \mathbb{R} | q(x_1, x_2; \mathcal{W}) = \sum_{i=1}^n w_i^{(2)} h_i, \text{where } h_i = \sigma(w_{i1}^{(1)} x_1 + w_{i1}^{(1)} x_2)\}$. The output is $y = q(x_1, x_2; \mathcal{W})$ where the $\mathcal{W}$ denotes the collection of weights.

   Consider the case where we set $x_2 = 1$ and $x_1 \in [0,1]$ with activation the Rectified Linear Unit (ReLU) $\sigma = \max(0, z)$. Show that with at most $n = k+2$ hidden units we can exactly represent any function $f(x_1)$ that is continuous piece-wise linear on $[0,1]$ with k interior transition points and $f(x) = 0$ for $x \notin (0,1)$. For instance, show that $f(x) = x$ for $x \leq 1/2$ and $f(x) = 1 - x$ for $x > 1/2$ can be exactly represented on $[0,1]$ by a MLP with $n = 3$ units.

2. (Backpropagation and Training) Consider approximating a general function $f(x)$ on $[0,1]$ by using a gradient descent $\dot{\mathbf{w}} = -\nabla_{\mathbf{w}} L$ to minimize the loss $L(q) = \frac{1}{m} \sum_{i=1}^m (f(z_i) - q(z_i; \mathbf{w}))^2$ for m data points $z_i \in [0,1]$, where we take in the MLP $x_1 = z_i$ and $x_2 = 1$. State for the MLP the back-propagation method for computing the gradient in $\mathbf{w}$. Draw the computational graph in the case when $n = 1$ and $m = 1$ for both the "forward pass" and the "backward pass." Explain techniques for how you might mitigate getting stuck in local minima or overfitting the data?

3. (Neural Networks as Universal Approximators) The Cybenko Theorem states that if a continuous activation function $g(z)$ is discriminatory on the unit cube $I_n \subset \mathbb{R}^n$ then the linear space $\mathcal{V} = \{q | q(\mathbf{x}) = \sum_{j=1}^n \alpha_j g(\mathbf{w}_j^T \mathbf{x} + b_j), n \in \mathbb{N}\}$ is dense in the space of continuous functions $\mathcal{C}(I_n)$. In other words, for any continuous function $f \in \mathcal{C}(I_n)$ and $\epsilon > 0$, there exists a $q \in \mathcal{V}$ such that $|f(\mathbf{x}) - q(\mathbf{x})| < \epsilon$ for all $\mathbf{x} \in I_n$. An activation function $g(z)$ is said to be discriminatory if for a Borel measure $\mu \in \mathcal{M}$ we have for all weights $\mathbf{w}, b$ that $\int g(\mathbf{w}^T \mathbf{x} + b) d\mu(\mathbf{x}) = 0$ then the measure must be zero $\mu \equiv 0$.

   (a) Show that the sigmoid activation function $g(z) = 1/1 + e^{-z}$ is discriminatory on $I_1 = [0,1]$. Hint: Use that $\int g(\mathbf{w}^T \mathbf{x} + b) d\mu(\mathbf{x}) = 0$ for all $\mathbf{w}, b$ iff $\int q(\mathbf{x}) d\mu(\mathbf{x}) = 0$ for all $q \in \mathcal{V}$.

   (b) Show that the ReLU activation function $g(z) = \max(z, 0)$ is discriminatory on I_1. Hint: Use that $\int g(\mathbf{w}^T \mathbf{x} + b) d\mu(\mathbf{x}) = 0$ for all $\mathbf{w}, b$ iff $\int q(\mathbf{x}) d\mu(\mathbf{x}) = 0$ for all $q \in \mathcal{V}$.

   (c) Show that the linear activation function $g(z) = z$ is not discriminatory on I_1. Hint: Construct a counter-example using a measure of the form $\mu(x) = a_1 \delta(x - x_1) + a_2 \delta(x - x_2) + a_3 \delta(x - x_3)$, where $\delta(\cdot)$ denotes here the Dirac δ-function (measure).