Assignment 1

CS7.601
Deep Learning: Theory and Practices
Spring 2020

Submission Deadline: 5.00 PM, 18/02/2020
Submission Venue: F-20, Machine Learning Lab, KRB, IIIT-H

Instructions

- All questions are compulsory to solve.
- Total marks are 35.
- **Only handwritten submissions are allowed.**

Problems

Problem 1. [8 Marks] Anti-symmetric sigmoidal activation function can be described as follows.

$$\begin{aligned} f(net) &= a\ \tanh(b\ net) = a\ \frac{[1 - e^{b\ net}]}{[1 + e^{b\ net}]} \\ &= \frac{2a}{1 + e^{b\ net}} - a \end{aligned} \tag{1}$$

Show that the Anti-symmetric sigmoidal activation function acts to transmit the maximum information if its inputs are distributed normally. Recall that the entropy (a measure of information) is defined as $H = -\int p(y) \log p(y) dy$.

1. Consider a continuous input variable x drawn from the density $p(x) \sim \mathcal{N}(0, c^2)$ (normal distribution with mean 0 and variance c^2). What is entropy for this distribution? [2 Marks]

2. Suppose samples x are passed through an anti-symmetric sigmoidal function to give $y = f(x)$, where the zero crossing of the sigmoid occurs at the peak of the Gaussian input, and the effective width of the linear region of sigmoidal equals to the range $-c < x < c$. What are the values of a and b in Eq.1 insures this? [2 Marks]

3. Calculate the entropy of the output distribution p(y). [2 Marks]

4. Suppose instead that the transfer function were a Dirac delta function $\delta(x-\theta)$. What is the entropy of the resulting output distribution $p(y)$? [2 Marks]

Problem 2. [3 Marks] Consider the sigmoidal transfer function described in Eq.1.

1. Show that its derivative $f'(net)$ can be written simply in terms of $f(net)$ itself. [1 Mark]

2. What are $f(net)$, $f'(net)$ and $f''(net)$ at $net = -\infty,\ 0,\ \infty$? [1 Mark]

3. For which value of net is the second derivative $f''(net)$ extremal? [1 Mark]

Problem 3. [3 Marks] Consider a standard three-layer back-propagation net with d input units, n_H hidden units, c output units, and bias. Let the activation function used be anti-symmetric sigmoid (Eq.1).

1. How many weights are in the net?[1 Mark]

2. Consider the symmetry in the value of the weights. In particular, show that if the sign is flipped on every weight, the network function is unaltered. [2 Marks]

Problem 4. [8 Marks] Assume that the criterion function $J(\mathbf{w})$ is well described to second order by a Hessian matrix H.

1. Show that convergence of learning is assured if the learning rate obeys $\eta < \frac{2}{\lambda_{max}}$, where λ_{max} is the largest eigenvalue of H. [3 Marks]

2. Show that the learning time is thus dependent upon the ratio of the largest to the smallest non-negligible eigenvalue of H. [3 Marks]

3. Explain why "standardizing" the training data can therefore reduce learning time. [2 Mark]

Problem 5. [6 Mark)] Consider a quadratic error function of the form

$$E = E_0 + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T H (\mathbf{w} - \mathbf{w}^*)$$

where $\mathbf{w}^*$ represents the minimum, and the Hessian matrix H is positive definite and constant. Suppose the initial weight vector $\mathbf{w}^{(0)}$ is chosen to be at the origin and is updated using simple gradient descent

$$\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau-1)} - \rho \nabla E$$

where τ denotes the iteration number, and ρ is the learning rate (which is assumed to be small).

1. Show that, after τ steps, the components of the weight vector parallel to the eigenvectors of H can be written

$$w_j^{(\tau)} = \{1 - (1 - \rho\eta_j)^\tau\} w_j^*$$

where $w_j = \mathbf{w}^T \mathbf{u}_j$, and $\mathbf{u}_j$ and η_j are the eigenvectors and eigenvalues, respectively, of H so that $H\mathbf{u}_j = \eta_j \mathbf{u}_j$. [2 Marks]

2. Show that as $\tau \to \infty$, this gives $\mathbf{w}^{(\tau)} \to \mathbf{w}^*$ as expected, provided $|1 - \rho\eta_j| < 1$. [2 Marks]

3. Now suppose that training is halted after a finite number τ of steps. Show that the components of the weight vector parallel to the eigenvectors of the Hessian satisfy (a) $w_j^{(\tau)} \simeq w_j^*$ when $\eta >> (\rho\tau)^{-1}$, (b) $|w_j^{(\tau)}| << |w_j^*|$ when $\eta << (\rho\tau)^{-1}$. [2 Marks]

Problem 6. [2 Marks] Show that if the transfer function of the hidden units is linear, a three-layer network is equivalent to a two-layer one. Explain why, therefore, that a three-layer network with linear hidden units cannot solve a non-linearly separable problem such as XOR or n-bit parity.

Problem 7 . [5 Marks]

1. Let $\mathcal{F}$ be a finite function class. Then VC-dimension of $\mathcal{F}$ is less than or equal to $\log |\mathcal{F}|$. [1 Mark]

2. Let M_n denotes the hypothesis space of monomial concepts defined on $\{0, 1\}^n$. Find the upper bound on the VC dimension of M_n. Also, find the lower bound on the VC dimension of M_n. [2 Marks]

3. Let $\mathcal{F}$ represents all rectangle shaped classifiers in 2 dimension. Show that the VC-dimension of $\mathcal{F}$ is 4. [2 Mark]