Assignment 5 (Sol.)

Introduction to Machine Learning
Prof. B. Ravindran

1. You are given the following neural networks which take two binary valued inputs $x_1, x_2 \in \{0, 1\}$ and the activation function is the threshold function($h(x) = 1$ if $x > 0$; 0 otherwise). Which of the following logical functions does it compute?

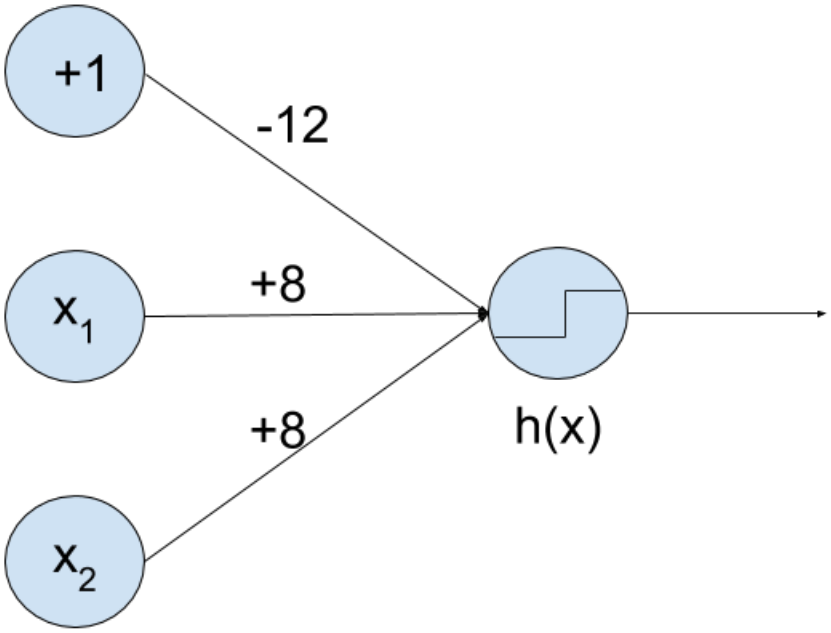

Figure 1: Q1

(a) OR
(b) AND
(c) NAND
(d) None of the above.

Solution: B
You can construct the truth table and see the values and decide which gate the network mimics.

0	0	0
0	1	0
1	0	0
1	1	1

2. We have a function which takes a two-dimensional input $x = (x_1, x_2)$ and has two parameters $w = (w_1, w_2)$ given by $f(x, w) = \sigma(\sigma(x_1 w_1)w_2 + x_2)$ where $\sigma(x) = \frac{1}{1+e^{-x}}$. We use backpropagation to estimate the right parameter values. We start by setting both the parameters to

0. Assume that we are given a training point $x_1 = 1, x_2 = 0, y = 5$. Given this information answer the next two questions. What is the value of $\frac{\partial f}{\partial w_2}$?

(a) 0.5

(b) -0.25

(c) 0.125

(d) -0.5

Solution: C
Write $\sigma(x_1 w_1)w_2 + x_2$ as o_2 and $x_1 w_1$ as o_1

$$\frac{\partial f}{\partial w_2} = \frac{\partial f}{\partial o_2}\frac{\partial o_2}{\partial w_2}$$

$$\frac{\partial f}{\partial w_2} = \sigma(o_2)(1 - \sigma(o_2)) \times \sigma(o_1)$$

$$\frac{\partial f}{\partial w_2} = 0.5 * 0.5 * 0.5$$

3. If the learning rate is 0.5, what will be the value of w_2 after one update using backpropagation algorithm?

(a) 0.0625

(b) -0.0625

(c) 0.5625

(d) - 0.5625

Solution: C
The update equation would be

$$w_2 = w_2 - \lambda \frac{\partial L}{\partial w_2}$$

where L is the loss function, here $L = (y - f)^2$

$$w_2 = w_2 - \lambda \times 2(y - f) \times (-1) \times \frac{\partial f}{\partial w_2}$$

Now putting in the given values we get the right answer.

4. Given N samples $x_1, x_2, \ldots, x_N$ drawn independently from a Gaussian distribution with variance σ_2 and unknown mean μ, find the MLE of the mean.

(a) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{\sigma^2}$

(b) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{2\sigma^2 N}$

(c) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$

(d) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N-1}$

Solution C
We will write the log likelihood as the following,

$$L = \sum_i \log(\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x_i-\mu)^2}{2\sigma^2}})$$

$$L = K + \sum_i \frac{(x_i-\mu)^2}{2\sigma^2}$$

Now we need to maximize this L, which we do by setting $\frac{\partial L}{\partial \mu}$ to 0, which gives us option C as the solution.

5. Continuing with the above question, assume that the prior distribution of the mean is also a Gaussian distribution, but with parameters mean μ_p and variance σ_p^2. Find the MAP estimate of the mean.

   (a) $\mu_{MAP} = \frac{\sigma^2\mu_p + \sigma_p^2 \sum_{i=1}^N x_i}{\sigma^2 + N\sigma_p^2}$

   (b) $\mu_{MAP} = \frac{\sigma^2 + \sigma_p^2 \sum_{i=1}^N x_i}{\sigma^2 + \sigma_p^2}$

   (c) $\mu_{MAP} = \frac{\sigma^2 + \sigma_p^2 \sum_{i=1}^N x_i}{\sigma^2 + N\sigma_p^2}$

   (d) $\mu_{MAP} = \frac{\sigma^2\mu_p + \sigma_p^2 \sum_{i=1}^N x_i}{N(\sigma^2 + \sigma_p^2)}$

   Solution C
   For a MAP estimate, we try to maximize $f(\mu)f(X|\mu)$

$$f(\mu)f(X|\mu) = \frac{1}{\sigma_p\sqrt{2\pi}} e^{\frac{(\mu-\mu_p)^2}{2\sigma_p^2}} \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x_i-\mu)^2}{2\sigma^2}}$$

   We will maximize this with respect to μ after taking a logarithm. This will yield the following equation,

$$\frac{\sum_i x_i}{\sigma} + \frac{\mu_p}{\sigma_p} - \mu(\frac{N}{\sigma} + \frac{1}{\sigma_p}) = 0$$

   Thus solution will be C

6. Which among the following statements is (are) true?

   (a) MAP estimates suffer more from overfitting than maximum likelihood estimates.

   (b) MAP estimates are equivalent to the ML estimates when the prior used in the MAP is a uniform prior over the parameter space.

   (c) One drawback of maximum likelihood estimation is that in some scenarios (hint: multinomial distribution), it may return probability estimates of zero.

   (d) The parameters which minimize the expected Bayesian L1 Loss is the median of the posterior distribution.

   Solution - B, C, D

7. Using the notations used in class and the tutorial document, evaluate the value of the neural network with a 3-3-1 architecture (2-dimensional input with 1 node for the bias term in both the layers). The parameters are as follows

$$\alpha = \begin{bmatrix} 1 & 0.2 & 0.4 \\ -1 & 0.3 & 0.5 \end{bmatrix}$$

$$\beta = \begin{bmatrix} 0.3 & 0.4 & 0.5 \end{bmatrix}$$

Using sigmoid function as the activation functions at both the layers, the output of the network for an input of (0.8, 0.7) will be

(a) 0.6710

(b) 0.6617

(c) 0.6948

(d) 0.3369

Solution C
This is a straight forward computation task. First pad x with 1 and make it the X vector,

$$X = \begin{bmatrix} 1 \\ 0.8 \\ 0.7 \end{bmatrix}$$

The output of the first layer can be written as

$$o_1 = \alpha X$$

Next apply the sigmoid function and compute

$$a_1(i) = \frac{1}{1 + e^{-o_1(i)}}$$

Then pad the a_1 vector also with 1 for bias, then compute the output of the second layer.

$$o_2 = \beta a_1$$

$$a_2 = \frac{1}{1 + e^{-o_2}}$$

$$a_2 = 0.6948$$

8. Which of the following statements is/are true about Neural Networks?

(a) Neural Networks can model arbitrarily complex decision boundaries.

(b) Neural Networks can be used to emulate a Gaussian kernel SVM

(c) Training of a neural network is very sensitive to the initial weights.

(d) Ideal initialization for weights would be setting all of them to zeros

Solution A, B, C
A - Neural networks are also called as universal approximators, because of their ability to learn complex functions by varying the number of layers and nodes.
B - The decision from any SVM is given by $\hat{y} = (\sum_{i=0}^{h} \alpha K(x_i, x) + b)$ where x_i represent the Support Vectors and K is the gaussian kernel. This can be implemented using a RBF-Neural Network. The first layer would be the input layer. Second layer would be the radial basis nodes, with as many nodes as support vectors in the SVM. And a single node in the final layer. The centers of the gaussian basis functions would be the support vectors of the SVM. The would be same as that of the kernel. The weights connected the hidden layer to the last layer would be given by i and a bias b. The activation function for the last layer would be the sgn function. C This is true because bad initializations might hinder the learning of the neural network, for example if you use all zeros the network might not be able to learn anything because of zero gradients.