NNSP-1

Homework #3

Instructor: Prof., Shayan Garani Srinivasa *TA: Prayag Gowgi*

Solution for 5.1

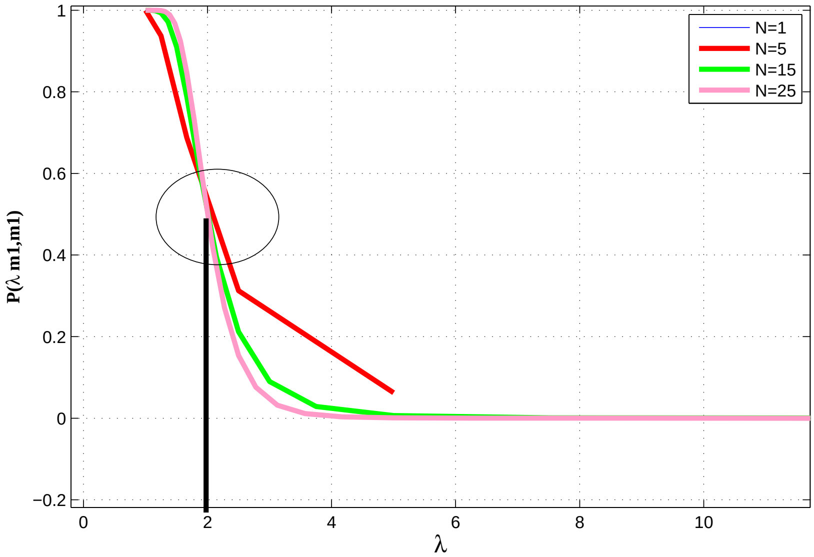

FIGURE 1 – Plot of $P(\lambda m_1, m_1)$ versus λ for $N = 1, 5, 15, 25$

In Figure 1, we see that P value is $\frac{1}{2}$ for each value of m_1.

Solution for 5.8

Given :

$$y(i) = \sum_{j=1}^{K} w_j(n) \exp\left(\frac{1}{2\sigma(n)^2}||x(i) - \mu_j(n)||^2\right)$$

$$E = \frac{1}{2}\sum_{i=1}^{n} e^2(i)$$

$$e(i) = d(i) - y(i)$$

1. The partial derivative of E with respect to $w_j(n)$, $\mu_j(n)$ and σ is given by

   a $\frac{\partial E}{\partial w_j(n)} = -e(j)\exp\left(-\frac{1}{2\sigma^2(n)}||x(j) - \mu_j(n)||^2\right)$

   b $\frac{\partial E}{\partial \mu_j(n)} = -\frac{1}{2\sigma^2(n)}e(j)w_j \exp\left(-\frac{1}{2\sigma^2(n)}||x(j) - \mu_j(n)||^2\right)(x(i) - \mu_j(n))$

   c $\frac{\partial E}{\partial \sigma(n)} = -\frac{1}{\sigma^3(n)}\sum_{j=1}^{N} e(j)w_j \exp\left(-\frac{1}{2\sigma^2(n)}||x(j) - \mu_j(n)||^2\right)||x(j) - \mu_j(n)||^2$

2. The update formulas for all the network parameters are as follows :

   a $w_j(n+1) = w_j(n) - \eta_w \frac{\partial E}{\partial w_j}$

   b $\mu_j(n+1) = \mu_j(n) - \eta_\mu \frac{\partial E}{\partial \mu_j}$

   c $\sigma(n+1) = \sigma(n) - \eta_\sigma \frac{\partial E}{\partial \sigma}$

3. In clustering the potential function is sum of the squared distances between the cluster center and the data point. The gradient $\frac{\partial E}{\partial \mu_j(n)}$ is trying to minimize the distance between μ_j, cluster center and the data point $x(i)$.

Solution for 7.4

Given :

$$E = \sum_{i=1}^{N} \left(d_i - \sum_{j=1}^{m_1} w_j G\left(||x_i - t_j||\right) \right)^2 + \lambda ||DF^*||^2 \tag{1}$$

where

$$\begin{aligned} F^* &= \sum_{i=1}^{m_1} w_i G\left(||x - t_i||\right) \\ ||DF^*||^2 &= w^{\mathrm{T}} G_0 w \end{aligned} \tag{2}$$

We would like to minimize E with respect to w. Differentiating E with respect to w_l we get,

$$\frac{\partial E}{\partial w_l} = -2 \sum_{i=1}^{N} \left(d_i G\left(||x_i - t_l||\right) - \left(\sum_{j=1}^{m_1} w_j G\left(||x_i - t_i||\right) \right) G\left(||x_i - t_l||\right) \right) + 2\lambda G_0 w \tag{3}$$

If $\hat{w}$ is an optimum w in equation (3), then we get

$$\begin{aligned} G^{\mathrm{T}} d &= \left(GG^{\mathrm{T}} + \lambda G_0\right) \hat{w} \\ \hat{w} &= \left(GG^{\mathrm{T}} + \lambda G_0\right)^{-1} G^{\mathrm{T}} d \end{aligned} \tag{4}$$

Solution for 7.5

Given :

$$\int_{\Re^{m_0}} ||DF(x)||^2 dx = \sum_{k=0}^{\infty} \int_{\Re^{m_0}} ||D^k F(x)||^2 dx \tag{5}$$

$$\begin{aligned} a_k &= \frac{\sigma^{2k}}{k! 2^k} \\ D^{2k} &= \left(\nabla^2\right)^k \\ D^{2k+1} &= \nabla \left(\nabla^2\right)^k \end{aligned} \tag{6}$$

where ∇ and ∇^2 is the usual gradient and Laplacian operator respectively.
From equation (5) LHS can be written as

$$\langle DF, DF \rangle_{\mathcal{H}} = \langle F, \tilde{D} DF \rangle_{\mathcal{H}} \tag{7}$$

We know that $L = D\tilde{D} = \sum_{k=0}^{\infty} (-1)^k \nabla^{2k}$ with

$$D = \sum_{k}^{\infty} \alpha_k^{\frac{1}{2}} \left(\frac{\partial}{\partial x_1} + \cdots + \frac{\partial}{\partial x_{m_0}} \right)^k \tag{8}$$

$$\tilde{D} = \sum_{k}^{\infty} (-1)^k \alpha_k^{\frac{1}{2}} \left(\frac{\partial}{\partial x_1} + \cdots + \frac{\partial}{\partial x_{m_0}} \right)^k \tag{9}$$

Therefore from equations (7) and (8) we get

$$DF(x) = \sum_{k=0}^{\infty} \frac{\sigma^k}{k! 2^k} \nabla^k F(x) \tag{10}$$

Solution for 2

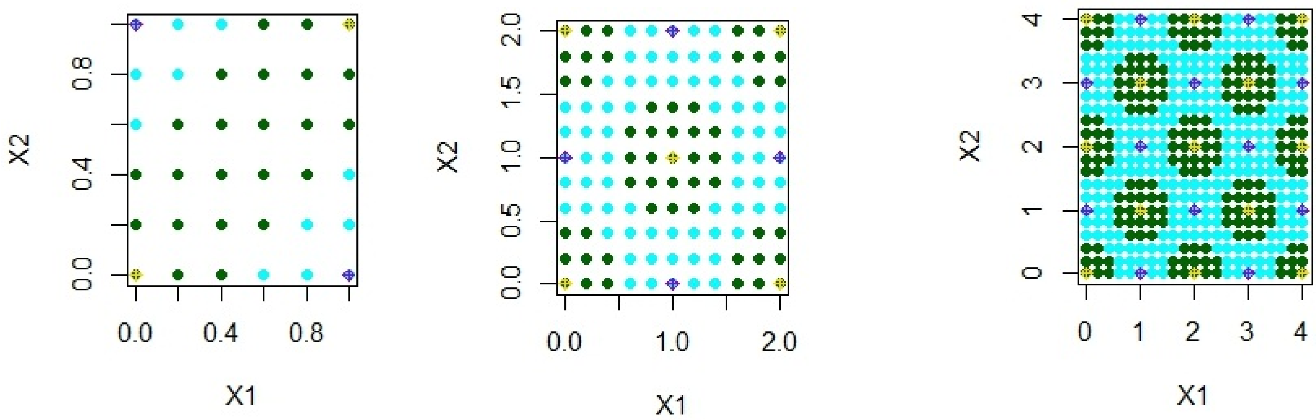

FIGURE 2 – Classification of **a)** Simple XOR **b)** Tiled XOR of size 3×3 and **c)** Tiled XOR of size 4×4.

Observations : We have used MASS package in R to perform this experiment. The complexity of hidden nodes, keeping the kernel width same, increased with increase in size of tiled XOR. Set of decision boundaries shown in Figure 2 shows that as the size of tiled XOR increases, the decision boundary has to bend itself to accommodate the data points of similar class. The other observation is regarding the width (σ) of the kernel. The width $\sigma = 0.6$ was used for simple XOR and $\sigma = 0.4$ for the other two cases. If the spread of tiled XOR is more, one would require larger σ. However, if there is any addition of tiled XOR to the data set, one would have to introduce additional nodes in the system.