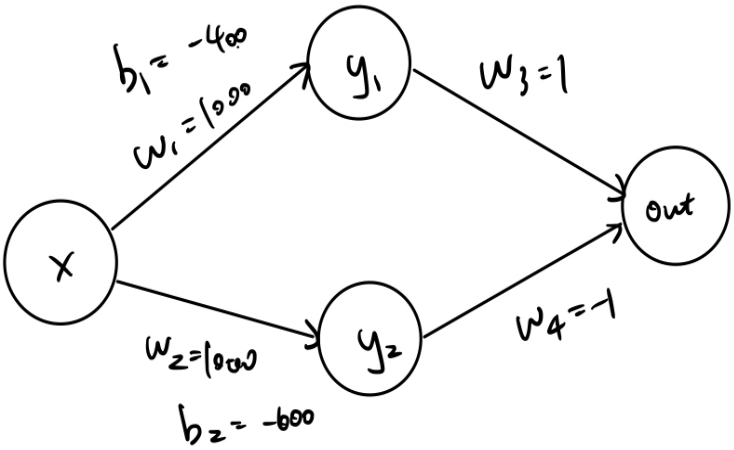
Homework 6

Xixiang Chen

1 Neural Networks and Universal Approximation Theorem

1.1

(a)



$y_1 = \sigma (w_1 X_1 + b_1))$

$y_2 = \sigma (w_2 X_2 + b_2))$

$out = w_3 y_1 + w_4 y_2$

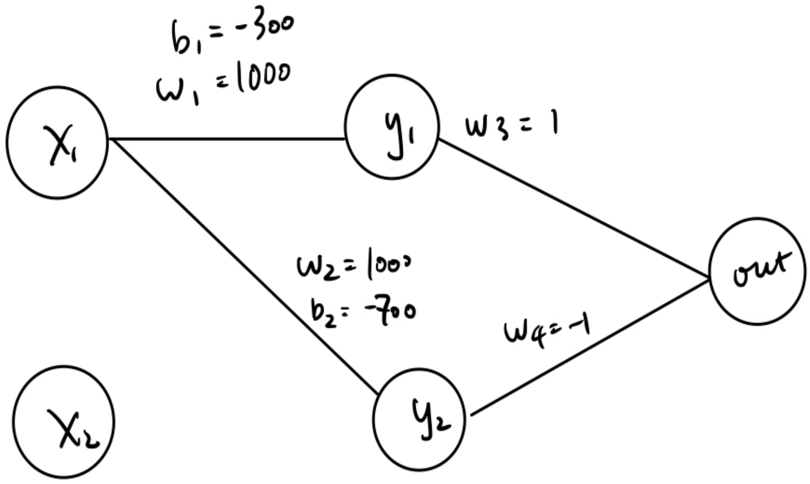
The minimum number of hidden neurons needed is 2.

(b)

(1) $w1, w2$ determine the steepness of the step-up and step-down part of the bump respectively.
(2) $-\frac{w1}{b1}, -\frac{w2}{b2}$ determine the step-up and step-down locations respectively.
(3) $w3, w4$ determine the height of the bump.

1.2

(a)



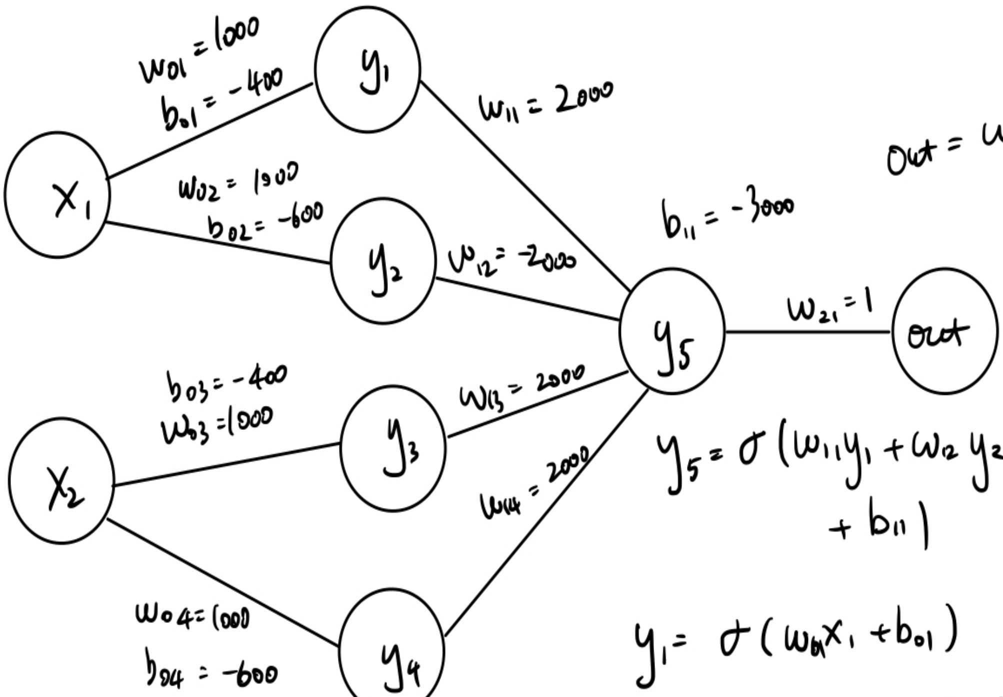
$$y_1 = \sigma(w_1 x_1 + b_1)$$

$$y_2 = \sigma(w_2 x_1 + b_2)$$

$$out = w_3 y_1 + w_4 y_2$$

THe minimum number of hidden neurons needed is 2.

(b)



$$out = w_{21} y_5$$

$$y_5 = \sigma(w_{11} y_1 + w_{12} y_2 + w_{13} y_3 + w_{14} y_4 + b_{11})$$

$$y_1 = \sigma(w_{01} x_1 + b_{01})$$

$$y_2 = \sigma(w_{02} x_1 + b_{02})$$

$$y_3 = \sigma(w_{03} x_2 + b_{03})$$

$$y_4 = \sigma(w_{04} x_2 + b_{04})$$

The minimum number of hidden neurons in the 1st-layer is 4.

(c)

2 EM

(a)

x_i the recorded number of failed cycles during visit i, $i \in 1, ..., m$.
z_i the machine used during visit i.
$w_i k$ the probability using machine k during visit i, $k \in 1, 2$.
X_i the random variable for number of failed cycles during visit i.
Since each failure cycle is independent, we have $P(X_i = x_i | z_i = k, \theta_k) = \binom{n}{x_i} \theta_k^x{}_i (1 - \theta_k)^{(n - x_i)}$.
E-step:

$$
\begin{aligned}
P(z_i = k | x_i, w_t, \theta_t) &= \frac{P(X_i = x_i | z_i = k, w_t, \theta_t) P(z_i = k | w_t, \theta_t)}{P(X_i = x_i | w_t, \theta_t)} \\
&= \frac{\binom{n}{x_i} \theta_k^{t\, x_i} (1 - \theta_k^t)^{(n - x_i)} P(z_i = k | w_t, \theta_t)}{\sum_k (\binom{n}{x_i} \theta_k^{t\, x_i} (1 - \theta_k^t)^{(n - x_i)} P(z_i = k | w_t, \theta_t))} \\
&= \frac{\binom{n}{x_i} \theta_k^{t\, x_i} (1 - \theta_k^t)^{(n - x_i)} w_{ik}^t}{\sum_k (\binom{n}{x_i} \theta_k^{t\, x_i} (1 - \theta_k^t)^{(n - x_i)} w_{ik}^t)} =: \gamma_{ik}^{t+1}
\end{aligned}
$$

M-step:

$$
\begin{aligned}
A(w, \theta, w_t, \theta_t) &= \sum_i \sum_k \gamma_{ik}^{t+1} \log P(X_i = x_i, z_i = k | w, \theta) \\
&= \sum_i \sum_k \gamma_{ik}^{t+1} \log P(z_i = k | w, \theta) P(X_i = x_i | z_i = k, w, \theta) \\
&= \sum_i \sum_k \gamma_{ik}^{t+1} \log(\binom{n}{x_i} \theta_k{}^{x_i} (1 - \theta_k)^{(n - x_i)} w_{ik}) \\
&= \sum_i \sum_k \gamma_{ik}^{t+1} (\log \binom{n}{x_i} + x_i \log \theta_k + (n - x_i) \log(1 - \theta_k) + \log w_{ik})
\end{aligned}
$$

Set partial derivative of the lagrangian to 0 with respect to w_{ik} and θ_k:

$$
\begin{aligned}
\frac{\partial L(w, \theta, w_t, \theta_t, \alpha, \beta)}{\partial w_{ik}} &= 0 = \frac{\partial}{\partial w_{ik}} (A(...) - \alpha) \\
\frac{\partial L(w, \theta, w_t, \theta_t, \alpha, \beta)}{\partial \theta_k} &= 0 = \frac{\partial}{\partial w_{ik}} (A(...) - \beta)
\end{aligned}
$$

And we know $\sum_k w_{ik} = w_i = 1$ and $\sum_k \theta_k = 1$. Solve to get:

$$
\begin{aligned}
w_{ik}^{t+1} &= \frac{\gamma_{ik}^{t+1}}{\sum_k \gamma_{ik}^{t+1}} \\
\theta_k^{t+1} &= \frac{\sum_i \gamma_{ik}^{t+1} x_i}{n \sum_i \gamma_{ik}^{t+1}}
\end{aligned}
$$

3 Clustering

(a)

(b)

(c)

Hierarchical agglomerative clustering performs better because the K-Means clusters the data by the Euclidean distance to the center of the cluster which always results in spherical cluster, where the dataset itself is not in spherical clusters.

(d)

We can add another dimension to the dataset, say y, such that y has very larg value when point is close to the center of the whole dataset and small value when point is spread out.