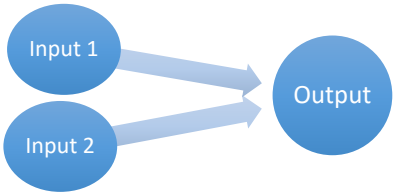
Problem 1

According to the description, there are 2 input nodes and 1 output node, without any hidden layer nodes.



Iteration 1:

Class 0:

Values of input node 1 & 2 are both: 1

Input to output node: $0 \times 1 + 0 \times 1 = 0$

Output from output node: $\frac{1}{1+e^{0}} = 0.5$

Error in output node: $\delta = (t_k - o_k)o_k(1 - o_k) = (0 - 0.5) \times 0.5 \times (1 - 0.5) = -0.125$

Update weights according to $w_{h,k} = \eta\delta_k o_k$

$w_{1,1} = 0 + (-0.125 \times 1) = -0.125$

$w_{2,1} = 0 + (-0.125 \times 1) = -0.125$

Class 1:

Output from input node 1: 2

Output from input node 2: 0

Input to output node: $-0.125 \times 2 + -0.125 \times 0 = -0.25$

Output from output node: $\frac{1}{1+e^{0.25}} = 0.4378$

Error in output node:

$\delta = (t_k - o_k)o_k(1 - o_k) = (1 - 0.4378) \times 0.4378 \times (1 - 0.4378) = 0.1384$

Update weights according to $w_{h,k} = \eta\delta_k o_k$

$w_{1,1} = -0.125 + (0.1384 \times 2) = 0.1518$

$w_{2,1} = -0.125 + (0.1384 \times 0) = -0.125$

Iteration 2:

Class 0:

Input to output node: $0.1518 \times 1 + -0.125 \times 1 = 0.0268$

Output from output node: $\frac{1}{1+e^{-0.0268}} = 0.5067$

Error in output node:$\delta = (t_k - o_k)o_k(1 - o_k) = -0.5067 \times 0.5067 \times 0.4933 = -0.1266$

Update weights according to $w_{h,k} = \eta\delta_k o_k$

$w_{1,1} = 0.1518 + (-0.1266 \times 1) = 0.0252$

$w_{2,1} = -0.125 + (-0.1266 \times 1) = -0.2516$

Class 1:

Input to output node: $0.0252 \times 2 - 0.2516 \times 0 = 0.0504$

Output from output node: $\frac{1}{1+e^{-0.0504}} = 0.5126$

Error in output node:

$\delta = (t_k - o_k)o_k(1 - o_k) = (1 - 0.5126) \times 0.5126(1 - 0.5126) = 0.1218$

Update weights according to $w_{h,k} = \eta \delta_k o_k$

$w_{1,1} = 0.0252 + (0.1218 \times 2) = 0.2688$

$w_{2,1} = -0.2516 + (0.1218 \times 0) = -0.2516$

Problem 2

During the 2nd iteration, output of class 0 sample is lower than output of class 1 sample, so that outputs become closer to their target values.

Problem 3

Introduce more features, like more parameters in sigmoid activation functions, or possibly perform more iterations. So that this ANN can be capable to process more features without adding hidden layers.

Problem 4

Given the conditions, the kernel function is the following:

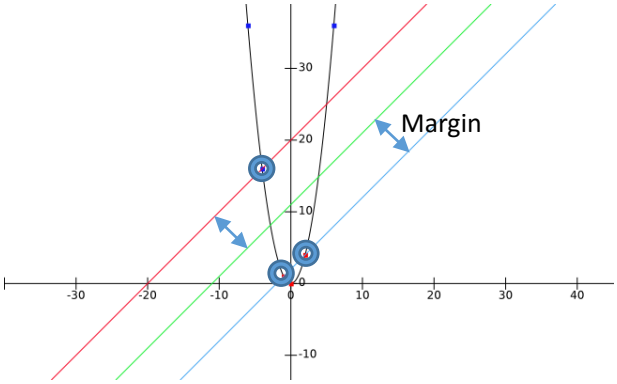
$$K(x, y) = \sum_{k=1}^{3} x_k y_k = x_1 y_1 + x_2 y_2 + x_3 y_3 = [x_1 \quad x_2 \quad x_3][y_1 \quad y_2 \quad y_3]^T = \phi(x)\phi(y)^T$$

So $\phi(\mathrm{u}) = [u_1 \quad u_2 \quad u_3]$

Problem 5

a. It cannot be separated by linear separator. Negative and positive samples are mixed on the line.
b. By introducing the feature map, we can get following answers:
   Positive examples: $\phi(-6) = (-6,36)$ $\phi(-4) = (-4,16)$ $\phi(6) = (6,36)$
   Positive examples: $\phi(-1) = (-1,1)$ $\phi(0) = (0,0)$ $\phi(2) = (2,4)$
c. Yes. They are not mixed any more, positive samples have greater values while negative ones are relatively smaller.

d. The plot is following:



The green line is hyperplane, red dots are negative samples while blue ones are positive.

Problem 6

Since a full dataset is split into several sets and trained by different classifiers respectively, and applying weighted majority voting based on training accuracy while performing classification, the averaged output will help eliminate the over-fitted classifiers which are affected by outliers too much.

Problem 7

The random forest algorithm randomly picks up features as roots and build a tree with remaining features. The algorithm will also repeat the procedure several times to build up a forest. Finally it will combine N fit trees by either calculating average or majority voting to avoid overfitting.

Problem 8

Random forests are consisted by several decision trees, while a decision tree itself is of low bias. The full model has a bias the same as a decision tree.

Problem 9

a. False. It can be, but not should be. In most cases it is a sigmoid function.
b. True. Multi-class SVM can be implemented by one class SVM.
c. False. It can have duplicate data, due to a bootstrap resample comes with replacement from data.
d. False. Linear kernel function is better when there are more features than data.