Question – 1

Total loss funtion $L = C \sum[1- y_i.f(x_i)]_+ + ½ (w^T w)$

This is hinge loss function which is used for maximum-minimum margin classification. And here were will consider only those 'i' for which $1- y_i f(x_i)$ is positive.

Derivation of L wrt to w:

$$\partial L/\partial w = -C \sum x_i y_i + w$$

Derivation of L wrt to b:

$$\partial L/\partial b = -C \sum y_i$$

Stochastic gradient descent:

while not at local minimum
do
pick i
$x \leftarrow x - \eta\nabla x f i (x)$
end

condition for local minima is $\nabla x f i (x)$ should be zero.

$\nabla x f i (x)$ is here equivalent to $\partial L/\partial w = -C \sum x_i y_i + w$.
so at each point we calculate the descent and move towards it and at next point we do the same till we reach local minima.

Gradient descent:

while not at local minimum
do
pick i
$x \leftarrow x - \eta\nabla x f i (x)$
end

It is also very familiar with the stochastic gradient descent. But we dont pick point at every other stop in the direction of gradient descent. Instead we directly put value of $\partial L/\partial w = 0$ in the equation to get the local minima.

$\partial L/\partial w = w – C \sum x_i y_i$

so, $w \leftarrow w-\eta.\partial L/\partial w$

$w \leftarrow w – (w – C\sum x_i y_i)$

compared to perceptron: It does better classification because perceptron never classify with the best boundary and also if data are non -linear separable perceptron will fail because Perceptron will keep changing till the all datas are linearly spearable.