ELEG 5040: Homework #1

Due on Monday, March 2, 2015, 7:30pm

Xiaogang Wang

Problem 1

[**20 points**]

Design a three layer neural network whose decision boundary is as shown in Figure 1. The gray region belongs to class 1 and other region belongs to class 0. Show your network structure, weights and nonlinear activation function.

Problem 2

[**20 points**]

The decision function of support vector machine $f(\mathbf{x})$ can be expressed as $f(\mathbf{x}) = 1$, if $\sum_{i=1}^{n} \alpha_i K(\mathbf{x}, \mathbf{x}_i) \geq 0$; otherwise $f(\mathbf{x}) = 0$. $\mathbf{x}$ is a test sample whose dimensionality is d, and $\{\mathbf{x}_i\}$ are support vectors autmatically selected from the training set. $K(\mathbf{x}, \mathbf{x}_i)$ is the kernel function to measure the similarity between $\mathbf{x}$ and $\mathbf{x}_i$. In this problem, we assume $K(\mathbf{x}, \mathbf{x}_i) = \exp(-||\mathbf{x} - \mathbf{x}_i||_2^2/\sigma^2)$. Show that this support vector machine can be implemented with a 3-layer neural network. Show the network structure, weights, and nonlinear activation functions at each layer.

Problem 3

[**20 points**]

Let $\mathbf{x}$ be an image and $f(\cdot)$ is a convolution operation. $g(\cdot)$ is spatial translation applied to an image. Prove that convolution has equivariance to translation, i.e. $f(g(\mathbf{x})) = g(f(\mathbf{x}))$. Is convolution equivariant to downsampling? Explain why.

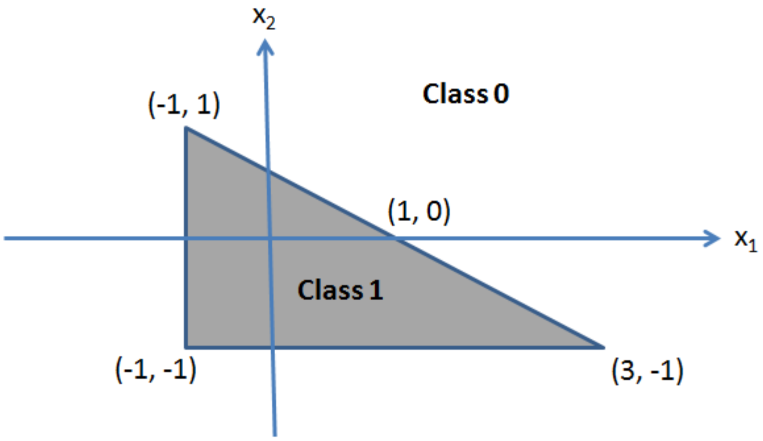

Figure 1:

Problem 4

[**20 points**]

In the lecture, we showed that gradient of a filter weight in a covolutional layer can be calculated from the correlation between the sensitivity map and the input map. It assumed one dimensional data sequence and the convolutional stride is one. Derive the equation when the convolution stride is k and also extend it to the 2D case.

Problem 5

[**20 points**]

Consider a three layer neural network whose structure is shown in Figure 2. You are required to calculate the sensitivty $\delta_k = -\frac{\partial J}{\partial net_k}$ at the output node k, where J is the objective function to be minimized and net_k is the net activation of the output node k. We consider two cases, where the objective function J and the nonlinear activation function at the output layer are chosen differently.

- In the first case, J is chosen as the squared error $J(\mathbf{W}) = \frac{1}{2}||\mathbf{t} - \mathbf{z}||_2^2 = \frac{1}{2}\sum_{t_k - z_k}^2$, where z_k is the prediction at the output node k and t_k is the corresponding target value. In the classification problem, only one t_k equals to 1 (corresponding the ground truth class) and all the other t_ks are all zeros. Sigmoid $f(net_k) = 1/(1 + e^{-net_k})$ is chosen as the activation function at the output layer. Calculate the sensitivity δ_k in terms t_k, z_k and net_k. Show that all the δ_k could be close to zero even if the prediction error is large and explain why this is bad.
- In the second case, the objective function is chosen as cross entropy $J(\mathbf{W}) = -\sum_{k=1}^{c} t_k \log(z_k)$ and the nonlinear activation function at the output layer is chosen as softmax $f(net_k) = \frac{e^{net_k}}{\sum_{k'=1}^{C} e^{net_{k'}}}$. Calculate the sensitivity t_k again. Prove that if the prediction error is large, at least one of the δ_k will be large.

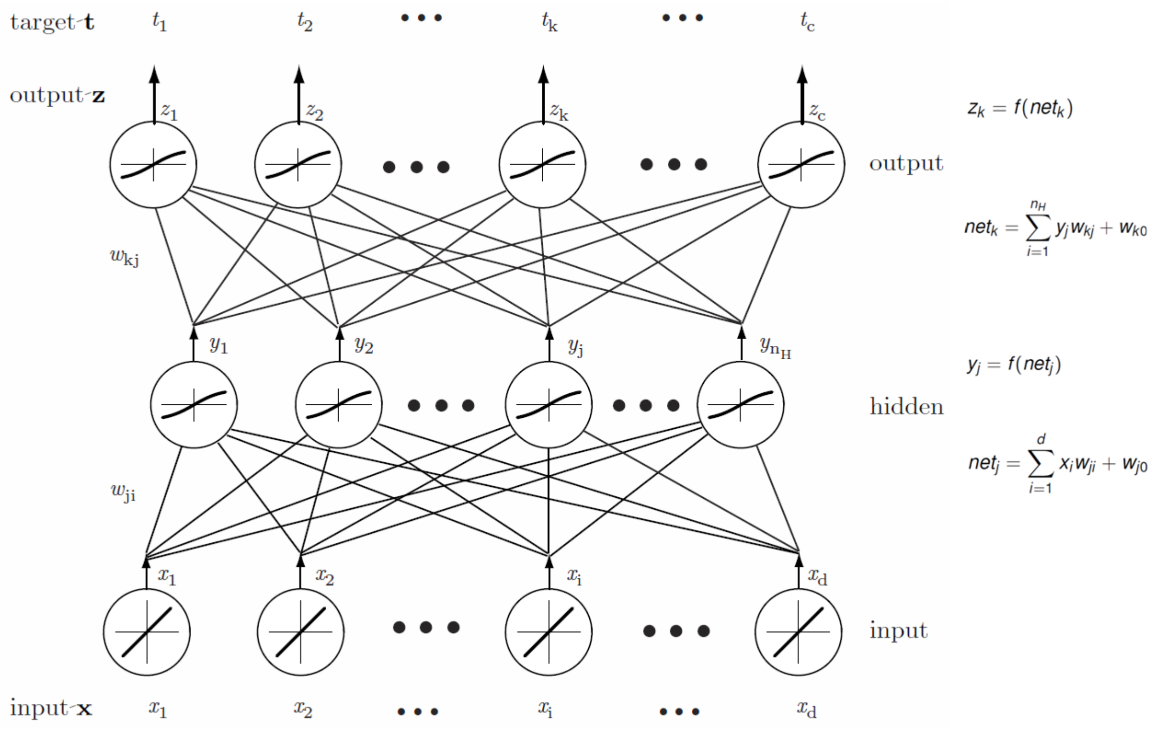
target $\mathbf{t}$
t_1
t_2
t_k
t_c
output $\mathbf{z}$
z_1
z_2
z_k
z_c
$z_k = f(net_k)$
output
w_{kj}
$net_k = \sum_{i=1}^{n_H} y_j w_{kj} + w_{k0}$
y_1
y_2
y_j
y_{n_H}
$y_j = f(net_j)$
hidden
w_{ji}
$net_j = \sum_{i=1}^{d} x_i w_{ji} + w_{j0}$
x_1
x_2
x_i
x_d
input
input $\mathbf{x}$
x_1
x_2
x_i
x_d

Figure 2: