# Solutions Pattern Recognition 2015
# Optimization, Neural Networks and Support Vector Machines

## Question 1: Optimization/Linear Regression

(a) The error function is:

$$E(w_0, w_1) = (4 - w_0 - w_1)^2 + (8 - w_0 - 2w_1)^2 + (6 - w_0 - 3w_1)^2$$

(b) The partial derivatives are:

$$\frac{\partial E}{\partial w_0} = -2(18 - 3w_0 - 6w_1)$$
$$\frac{\partial E}{\partial w_1} = -2(38 - 6w_0 - 14w_1)$$

We get two linear equations with two unknowns:

$$18 - 3w_0 - 6w_1 = 0 \quad (1)$$
$$38 - 6w_0 - 14w_1 = 0 \quad (2)$$

Solving for $w_0$ and $w_1$ we find: $w_0 = 4$, $w_1 = 1$. So $y(x) = 4 + x$.

(c) The second derivatives are:

$$\frac{\partial^2 E}{\partial w_0^2} = 6 \qquad \frac{\partial^2 E}{\partial w_1^2} = 28 \qquad \frac{\partial^2 E}{\partial w_0 \partial w_1} = 12$$

Putting these in the Hessian matrix we get

$$H = \begin{bmatrix} 6 & 12 \\ 12 & 28 \end{bmatrix}$$

We find $H_{11} = 6 > 0$ and $\det(H) = 6 \cdot 28 - 12 \cdot 12 = 24 > 0$. Since both are positive, we conclude that $H$ is positive definite. This means the point $(w_0 = 4, w_1 = 1)$ is a (local) minimum. If fact, since the Hessian matrix is positive definite everywhere (the second derivatives do not depend on the values of $w_0$ and $w_1$), the error function is globally convex (or concave up) so that $(w_0 = 4, w_1 = 1)$ is the unique global minimum.

## Question 2: Optimization/Linear Regression

(a) The partial derivatives are:

$$\frac{\partial E}{\partial w_0} = -\sum_{n=1}^{N}(t_n - w_0 - w_1 x_n)$$
$$\frac{\partial E}{\partial w_1} = -\sum_{n=1}^{N} x_n(t_n - w_0 - w_1 x_n)$$

So we have:

$$\nabla E(\mathbf{w}) = \begin{bmatrix} \frac{\partial E}{\partial w_0} \\ \frac{\partial E}{\partial w_1} \end{bmatrix} = \begin{bmatrix} -\sum_{n=1}^{N}(t_n - w_0 - w_1 x_n) \\ -\sum_{n=1}^{N} x_n(t_n - w_0 - w_1 x_n) \end{bmatrix}$$

(b) For a single observation $(t_n, x_n)$ the gradient is:

$$\nabla E_n(\mathbf{w}) = \begin{bmatrix} \frac{\partial E_n}{\partial w_0} \\ \frac{\partial E_n}{\partial w_1} \end{bmatrix} = \begin{bmatrix} -(t_n - w_0 - w_1 x_n) \\ -x_n(t_n - w_0 - w_1 x_n) \end{bmatrix}$$

For the given data point and weight vector $\mathbf{w}^{(0)}$ we get:

$$\nabla E_n(\mathbf{w}^{(0)}) = \begin{bmatrix} -(3 - 1.6 - 0.8 \times 3) \\ -3(3 - 1.6 - 0.8 \times 3) \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

With $\eta = 0.1$, the new weights become:

$$\mathbf{w}^{(1)} = \begin{bmatrix} 1.6 \\ 0.8 \end{bmatrix} - 0.1 \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 0.5 \end{bmatrix}$$

(c) With $\mathbf{w}^{(0)}$ the prediction for $x_n = 3$ was

$$y(x_n = 3) = 1.6 + 0.8 \times 3 = 4$$

So the squared prediction error for the data point is $(y(x_n) - t_n)^2 = (4 - 3)^2 = 1$. With the new weight vector the prediction is:

$$y(x_n = 3) = 1.5 + 0.5 \times 3 = 3$$

This gives a prediction error of zero which is obviously an improvement.

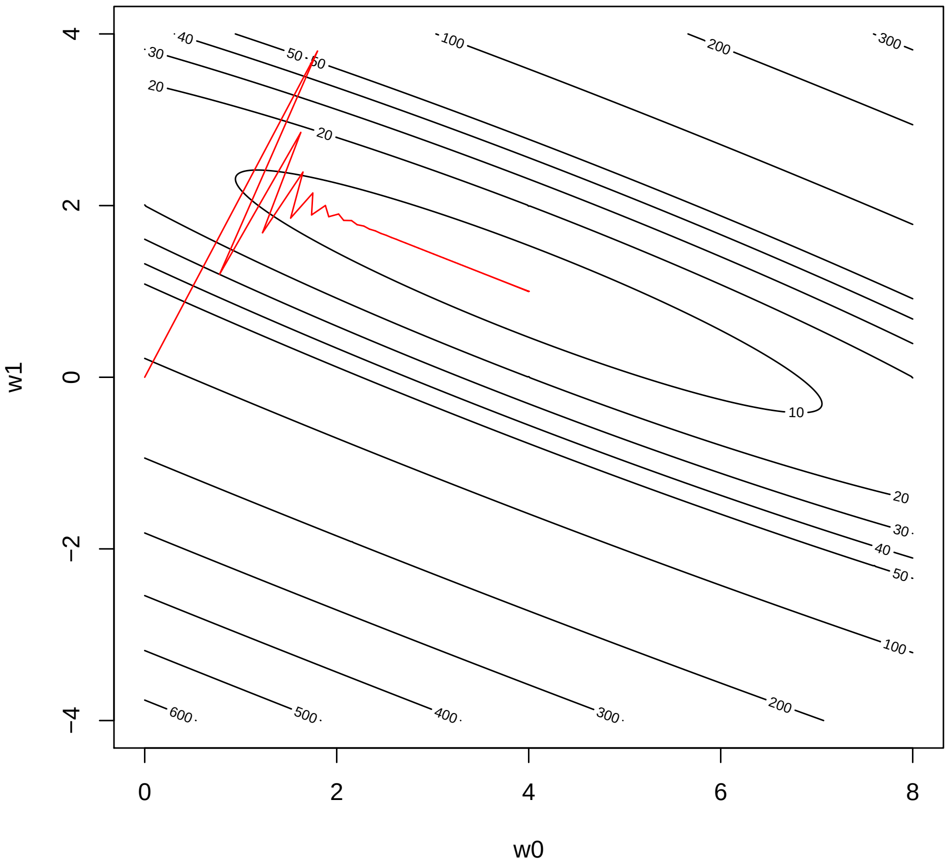
With $\eta = 0.2$ the new weight vector becomes:

$$\mathbf{w}^{(1)} = \begin{bmatrix} 1.6 \\ 0.8 \end{bmatrix} - 0.2 \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 1.4 \\ 0.2 \end{bmatrix}$$

The prediction becomes:

$$y(x_n = 3) = 1.4 + 0.2 \times 3 = 2$$

(d) You could have produced this picture with the code on the web site:




The gradient trajectory (the red line) starts in the point $(w_0 = 0, w_1 = 0)$ and with step size $\eta = 0.1$ converges to the global minimum $(w_0 = 4, w_1 = 1)$.

(e) P.M.

## Question 3: Neural Networks

(a) Output $y_5$:

$$y_5 = w_{50} + w_{53}\left(\frac{1}{1 + e^{-w_{30} - w_{31}x_1 - w_{32}x_2}}\right) + w_{54}\left(\frac{1}{1 + e^{-w_{40} - w_{41}x_1 - w_{42}x_2}}\right)$$

(b) Output units: $\delta_k = y_k - t_k$, $k = 5, 6, 7$. For the hidden units we have

$$\delta_3 = z_3(1 - z_3)(w_{53}\delta_5 + w_{63}\delta_6 + w_{73}\delta_7)$$
$$\delta_4 = z_4(1 - z_4)(w_{54}\delta_5 + w_{64}\delta_6 + w_{74}\delta_7)$$

(c) Activation of hidden units:

$$a_3 = 0.2 + 0.1x_1 + 0.2x_2 = 0.2 + 0.1 \times 2 + 0.2 \times 3 = 1$$
$$a_4 = 0.05 + 0.15x_1 + 0.1x_2 = 0.05 + 0.15 \times 2 + 0.1 \times 3 = 0.65$$

Output of hidden units:

$$z_3 = \frac{1}{1 + e^{-a_3}} = 0.73$$
$$z_4 = \frac{1}{1 + e^{-a_4}} = 0.66$$

Activation of output units:

$$y_5 = 3 + 8z_3 + 6z_4 = 3 + 8 \times 0.73 + 6 \times 0.66 = 12.8$$
$$y_6 = 1 + 4z_3 + 3z_4 = 1 + 4 \times 0.73 + 3 \times 0.66 = 5.9$$
$$y_7 = 2 + 2z_3 + 5z_4 = 2 + 2 \times 0.73 + 5 \times 0.66 = 6.76$$

The output of the output units is the same as their activation (*linear* output units: $h(a) = a$).

(d) Output units:

$$\delta_5 = y_5 - t_5 = 12.8 - 15 = -2.2$$
$$\delta_6 = y_6 - t_6 = 5.9 - 5 = 0.9$$
$$\delta_7 = y_7 - t_7 = 6.76 - 7 = -0.24$$

Hidden units:

$$\delta_3 = 0.73 \times 0.27 \times (8(-2.2) + 4(0.9) + 2(-0.24)) = -2.85$$
$$\delta_4 = 0.66 \times 0.34 \times (6(-2.2) + 3(0.9) + 5(-0.24)) = -2.63$$

(e) The partial derivatives are:

$$\frac{\partial E}{\partial w_{31}} = x_1 \times \delta_3 = 2 \times -2.85 = -5.7$$
$$\frac{\partial E}{\partial w_{73}} = z_3 \times \delta_7 = -0.24 \times 0.73 = -0.18$$

(f) The new weight values become:

$$w_{31}^{(\text{new})} = w_{31}^{(\text{old})} - \eta \frac{\partial E}{\partial w_{31}} = 0.1 - 0.01 \times -5.7 = 0.157$$
$$w_{73}^{(\text{new})} = w_{73}^{(\text{old})} - \eta \frac{\partial E}{\partial w_{73}} = 2 - 0.01 \times -0.18 = 2.0018$$

## Question 4: Support Vector Machines

(a) The support vectors are the attribute vectors with positive lagrange multiplier, so row 4, 6 and 7 in the data table:

$$\mathbf{x}_4 = \begin{bmatrix} 3 \\ 3 \end{bmatrix} \quad \mathbf{x}_6 = \begin{bmatrix} 4 \\ 6 \end{bmatrix} \quad \mathbf{x}_7 = \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

(b) To compute the value of the SVM bias term $b$, we use the formula

$$b = t_m - \sum_{n=1}^{N} a_n t_n \mathbf{x}_m^\top \mathbf{x}_n,$$

with any support vector, for example $\mathbf{x}_6 = [4 \;\; 6]^\top$. This yields:

$$b = 1 + 3\frac{1}{6}[4 \;\; 6]\begin{bmatrix} 3 \\ 3 \end{bmatrix} - [4 \;\; 6]\begin{bmatrix} 4 \\ 6 \end{bmatrix} - [4 \;\; 6]\begin{bmatrix} 6 \\ 4 \end{bmatrix} = -4$$

(c) To predict the class label for given attribute vectors, we use the formula

$$y(\mathbf{x}) = b + \sum_{n=1}^{N} a_n t_n \mathbf{x}^\top \mathbf{x}_n,$$

with $\mathbf{x} = [0 \;\; 7]^\top$. This yields:

$$y(\mathbf{x}) = -4 - 3\frac{1}{6}[0 \;\; 7]\begin{bmatrix} 3 \\ 3 \end{bmatrix} + [0 \;\; 7]\begin{bmatrix} 4 \\ 6 \end{bmatrix} + [0 \;\; 7]\begin{bmatrix} 6 \\ 4 \end{bmatrix} = -\frac{1}{2}$$

Since $y(\mathbf{x}) < 0$ we predict class $-1$.

(d) The weight vector is:

$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n = -3\frac{1}{6}\begin{bmatrix} 3 \\ 3 \end{bmatrix} + \begin{bmatrix} 4 \\ 6 \end{bmatrix} + \begin{bmatrix} 6 \\ 4 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$$

The equation for the maximum margin decision boundary is:

$$\frac{1}{2}x_1 + \frac{1}{2}x_2 - 4 = 0$$