Exercises Pattern Recognition 2015
Optimization, Neural Networks and Support Vector Machines

Question 1: Optimization/Linear Regression

Just for practice, let's solve a linear regression problem from first principles, that is, without using the formulas we derived for w_0 and w_1.

We are given the following three observations on x and t:

n	x_n	t_n
1	1	4
2	2	8
3	3	6

We want to fit a linear regression model

$$y(x) = w_0 + w_1 x$$

by the method of least-squares.

(a) Specify the sum of squared errors function $E(w_0, w_1)$ for this specific data set.

(b) Determine the partial derivatives $\frac{\partial E}{\partial w_0}$ and $\frac{\partial E}{\partial w_1}$ for the error function you found under (a). Equate both partial derivatives to zero and solve for w_0 and w_1.

(c) Determine the second order partial derivatives, and put them in the Hessian matrix. Verify that we have indeed found a minimum (rather than maximum or saddle point) by ascertaining that the Hessian matrix is positive definite for the values of w_0 and w_1 that you found under (b).

Question 2: Optimization/Linear Regression

In simple linear regression, the sum of squared errors is given by:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - w_0 - w_1 x_n)^2,$$

where $\mathbf{w} = [w_0 \;\; w_1]^\top$ denotes the weight vector to be estimated from the data.
Suppose we want to minimize this error function using the method of gradient descent.

(a) Derive an expression for the gradient $\nabla E(\mathbf{w})$.

(b) Let $\mathbf{w}^{(0)} = [1.6 \;\; 0.8]^\top$, and the step size (learning rate) $\eta = 0.1$.
Use the single data point $t_n = 3, x_n = 3$ to update the weight vector.

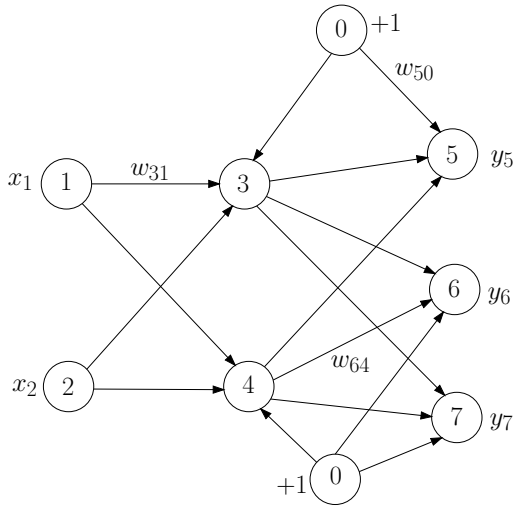
(c) Verify whether the update has decreased the squared prediction error for the data point used. What if η were 0.2 instead of 0.1?

(d) If you have your laptop with you, play around a little with the R code for gradient descent for linear regression (see the course web page under *schedule* at lecture 50B).

(e) (Continuation of (a); hard) Derive a general expression for the Hessian matrix for simple linear regression and prove that it is positive definite.

Question 3: Neural Networks

Given is the following feed-forward neural network

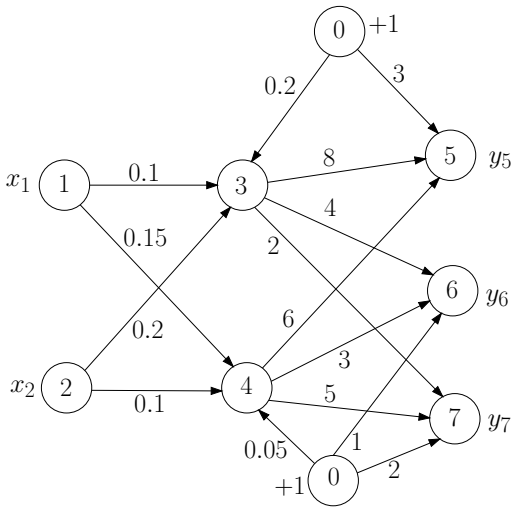


The activation function in the hidden units (3 and 4) is the logistic function ($h(a) = \sigma(a)$). The activation function in the output units (5,6 and 7) is the identity function function ($h(a) = a$). As usual, node 0 denotes the bias unit that always outputs the value +1. To avoid clutter, two incarnations of the bias unit are drawn in the picture. As error function we use squared error at all output units.

(a) Write network output y_5 as a function of the inputs $x_0 \equiv 1$, x_1, x_2, and the network weights w_{ji}.

(b) Give expressions for the local errors (δ) of the hidden units and the output units, using the given error function and activation functions.

Suppose the network is being trained, using some form of gradient descent algorithm. The current weight values are



We are processing an observation with input values $x_1 = 2$, $x_2 = 3$, and target values $t_5 = 15$, $t_6 = 5$, and $t_7 = 7$ (we numbered the targets to match with the numbers of the output units, for example y_6 is the prediction for t_6).

(c) Compute the predicted values for t_5, t_6, and t_7 for the given input values by "forward propagation" through the network.

(d) Compute the values of the local error δ_j for each output unit and hidden unit.

(e) Compute the value of the derivative of the error function with respect to w_{31} and w_{73} for the current weight values.

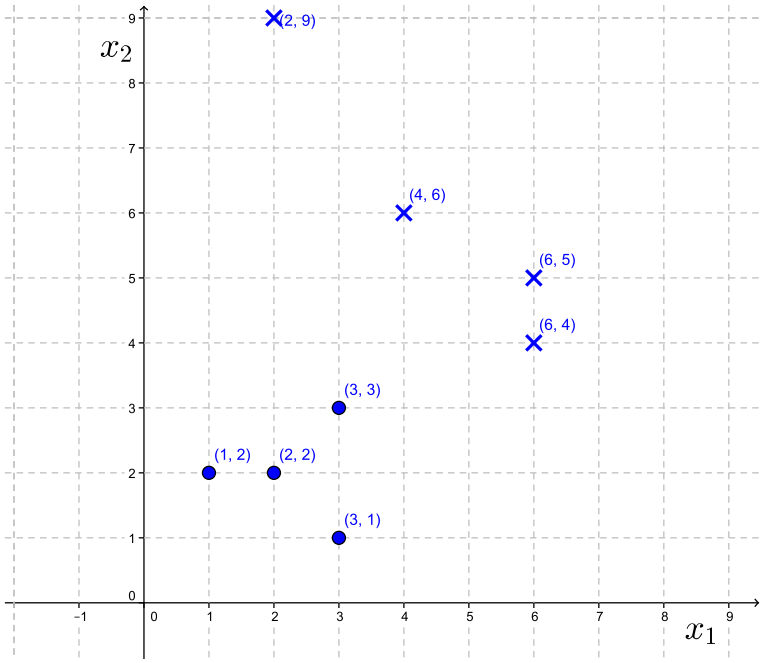
(f) Suppose we apply a gradient descent algorithm with learning rate $\eta = 0.01$. Give the new values for w_{31} and w_{73}.

Question 4: Support Vector Machines

We receive the following output from the optimization software for fitting a support vector machine with linear kernel and perfect separation of the training data:

n	$x_{n,1}$	$x_{n,2}$	t_n	a_n
1	1	2	-1	0
2	2	2	-1	0
3	3	1	-1	0
4	3	3	-1	$3\frac{1}{6}$
5	2	9	$+1$	0
6	4	6	$+1$	1
7	6	4	$+1$	1
8	6	5	$+1$	0

Here $x_{n,1}$ denotes the value of x_1 for the n-th observation, a_n is the value of the Lagrange multiplier for the n-th observation, etc. The figure below is a plot of the same data set, where the dots represent points with class -1, and the crosses points with class $+1$.



You are given the following formulas:

$$b = t_m - \sum_{n=1}^{N} a_n t_n \mathbf{x}_m^\top \mathbf{x}_n \qquad \text{(for any support vector } \mathbf{x}_m\text{)}$$

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

Answer the following questions:

(a) Give the support vectors for this problem.

(b) Compute the value of the SVM bias term b.

(c) Which class does the SVM predict for the data point $x_1 = 0, x_2 = 7$?

(d) Use equation 7.8 from Bishop (and the lecture slides) to determine the value of the weight vector $\mathbf{w}$. Give the equation of the maximum margin decision boundary. Draw it in the graph.