582669 Supervised Machine Learning (Spring 2014)
Homework 5, sample solutions

Credit for the solutions goes to mainly to Panu Luosto and Joonas Paalasmaa, with some additional contributions by Jyrki Kivinen

Problem 1

This exercise is of course a drill: there are simpler ways of finding the minimum of \( x^2 \) in a closed interval.

First part. Let \( f(x) = x^2 \) and \( p(x) = (x-2)^2 - 1 \). We are asked to minimize \( f(x) \) with the constraint \( p(x) \leq 0 \). Both functions are clearly convex, and the Karush-Kuhn-Tucker conditions of this convex optimization problem are

\[
\begin{cases}
    f'(x) + \lambda p'(x) = 0 \\
    p(x) \leq 0 \\
    \lambda \geq 0 \\
    \lambda p(x) = 0 ,
\end{cases}
\]

or equivalently

\[
\begin{cases}
    2x + \lambda(2x - 4) = 0 & \text{(i)} \\
    (x-2)^2 - 1 \leq 0 & \text{(ii)} \\
    \lambda \geq 0 & \text{(iii)} \\
    \lambda((x-2)^2 - 1) = 0. & \text{(iv)}
\end{cases}
\]

The solutions of (iv) are \( \lambda = 0 \), \( x = 3 \) or \( x = 1 \). If \( \lambda = 0 \), we get \( x = 0 \) from (i), and the condition (ii) is not true. Otherwise if \( x = 3 \), equation (i) has the form \( 6 + 2\lambda = 0 \), which means that condition (iii) is violated. We are left with \( x = 1 \). Then (i) yields \( 2 - 2\lambda = 0 \leftrightarrow \lambda = 1 \), and the inequalities (ii) and (iii) are true. In other words, \( f(x) \) is minimized subject to \( p(x) \leq 0 \) when \( x = 1 \).

The Lagrangian of the problem is

\[
L(x, \lambda) = f(x) + \lambda p(x) \\
= x^2 + \lambda((x-2)^2 - 1) \\
= (1 + \lambda)x^2 - 4\lambda x + 3\lambda ,
\]

and the dual is

\[
g(\lambda) = \inf_{x \in \mathbb{R}} L(x, \lambda) \\
= \inf_{x \in \mathbb{R}} \left( (1 + \lambda)x^2 - 4\lambda x + 3\lambda \right) .
\]

The dual problem is to maximize \( g(\lambda) \) in the set \( \lambda \in [0, \infty] \). Let \( h(x) = (1 + \lambda)x^2 - 4\lambda x + 3\lambda \). Because \( \lambda \geq 0 \), the graph of \( h \) is a convex parabola, and the minimum of \( h \) is at the point where \( h'(x) = (2 + 2\lambda)x - 4\lambda = 0 \Rightarrow x = 2\lambda/(1 + \lambda) \). Thus

\[
g(\lambda) = (1 + \lambda)\frac{4\lambda^2}{(1 + \lambda)^2} - \frac{8\lambda^2}{1 + \lambda} + 3\lambda \\
= \frac{4\lambda^2 - 8\lambda^2 + 3\lambda + 3\lambda^2}{1 + \lambda} \\
= \frac{3\lambda - \lambda^2}{1 + \lambda} .
\]
With some more calculations, we could show that the dual \( g \) is maximized with \( \lambda = 1 \), and \( g(1) = 1 \). Our optimization problem is convex, and \( p(2) < 0 \), in other words, there is \( x \in \mathbb{R} \) such that \( p(x) \) is strictly less than zero. Therefore, according to Theorem 3.4 of the lecture slides, the strong duality holds, and \( g(\lambda^*) = f(x^*) \), where \( \lambda^* \) and \( x^* \) are the solutions of the dual and primal optimization problems, respectively.

**Second part.** Let \( f(x) = x^2 \) and let \( p(x) = (x - 2)^2 - 8 \). Again, we want to minimize the convex function \( f \) subject to \( p(x) \leq 0 \), where \( p \) is also a convex function. The KKT conditions of this convex optimization problem are

\[
\begin{align*}
2x + \lambda(2x - 4) &= 0 \quad (i) \\
(x - 2)^2 - 8 &\leq 0 \quad (ii) \\
\lambda &\geq 0 \quad (iii) \\
\lambda((x - 2)^2 - 8) &= 0. \quad (iv)
\end{align*}
\]

From (iv) and (i) we find easily the solution \( x = \lambda = 0 \). In order to check for other possible solutions (for practice only), we get \( x = 2 + 2\sqrt{2} \) and \( x = 2 - 2\sqrt{2} \) from (iv). But then solving \( \lambda = x/(2 - x) \) from (i), shows that the condition (iii) is violated in these two cases. So the solution of the problem is \( x = 0 \).

The Lagrangian of \( f \) is

\[
L(x, \lambda) = f(x) + \lambda p(x) = x^2 + \lambda((x - 2)^2 - 8),
\]

and the dual is

\[
g(\lambda) = \inf_{x \in \mathbb{R}} L(x, \lambda) = \inf_{x \in \mathbb{R}} \left((1 + \lambda)x^2 - 4\lambda x - 4\lambda\right).
\]

Finding the infimum goes along similar steps as in the first part of the exercise, and

\[
g(\lambda) = (1 + \lambda)\left(\frac{2\lambda}{1 + \lambda}\right)^2 - 4\lambda \frac{2\lambda}{1 + \lambda} - 4\lambda = \frac{4\lambda^2 - 8\lambda^2 - 4\lambda - 4\lambda^2}{1 + \lambda} = -2\lambda \frac{2 + 4\lambda}{1 + \lambda}.
\]

In the dual problem, we maximize \( g(\lambda) \) in the set \( \lambda \in [0, \infty] \). Because \( g(\lambda) \leq 0 \) everywhere in \( \mathbb{R}_+ \), and \( g(\lambda) = 0 \) only when \( \lambda = 0 \), the maximum is achieved when \( \lambda = 0 \). Then \( x = (2 \cdot 0)/(1 + 0) = 0 \). Strong duality holds also in this case, which can be be seen also from the fact that all the KKT conditions are satisfied when \( \lambda = x = 0 \).

**Problem 2**

Let \( f : \mathbb{R}^d \to \mathbb{R} \), \( f(w) = \|w - w_t\|^2 = (w - w_t) \cdot (w - w_t) = w \cdot w - 2w \cdot w_t + w_t \cdot w_t \), and let \( p(w) = 1 - y_t w \cdot x_t \). Notice that this is the revised version of the exercise 2, in the original version \( p \) was different. We prove as a warm-up formally that \( f \) is a convex function in \( \mathbb{R}^d \).

A twice differentiable function is convex in a convex set if and only if its Hessian matrix is positive semidefinite on the interior of the convex set (compare with the non-negativeness of the second derivative of a function of type \( \mathbb{R} \to \mathbb{R} \)). Because \( f(w) = \sum_{i=1}^d (w_i - w_{t,i})^2 \), it is easy to see that

\[
\frac{\partial f(w)}{\partial w_i} = 2(w_i - w_{t,i}) \quad (i \in \{1, 2, \ldots, d\}),
\]
and for all $i, j \in \{1, 2, \ldots, d\}$
\[
\frac{\partial^2 f(w)}{\partial w_i \partial w_j} = 2 \quad \text{if} \quad i = j \quad \text{and} \quad \frac{\partial^2 f(w)}{\partial w_i \partial w_j} = 0 \quad \text{if} \quad i \neq j .
\]
So the Hessian matrix of $f$ is $H(f) = 2I_d$. It holds for all $z \in \mathbb{R}^{d \times 1}$ that $z^T (2I_d) z = 2 \| z \|^2 \geq 0$, which means $H(f)$ is positive semidefinite.

The optimization problem is to minimize $f(w)$ subject to $p(w) \leq 0$. The Karush-Kuhn-Tucker conditions are
\[
\begin{cases}
\nabla f(w) + \lambda \nabla p(w) = 0 \\
p(w) \leq 0 \\
\lambda \geq 0 \\
\lambda p(w) = 0,
\end{cases}
\]
or equivalently
\[
\begin{cases}
2w - 2w_t - \lambda y_t x_t = 0 \quad (i) \\
1 - y_t w \cdot x_t \leq 0 \quad (ii) \\
\lambda \geq 0 \quad (iii) \\
\lambda (1 - y_t w \cdot x_t) = 0 . \quad (iv)
\end{cases}
\]
Equation (i) yields
\[
w = w_t + (1/2) \lambda y_t x_t . \quad (1)
\]
From equality (iv) we get two solutions. In the first case, $\lambda = 0$, and inequality (iii) is true. Also, $w = w_t$, and $1 - y_t w_t \cdot x_t \leq 0$ has to hold.

In the second solution $\lambda > 0$, and we get from (iv)
\[
1 - y_t w \cdot x_t = 1 - y_t (w_t + \frac{1}{2} \lambda y_t x_t) \cdot x_t \\
= 1 - y_t w_t \cdot x_t - \frac{1}{2} \lambda x_t \cdot x_t \\
= 0
\]
where we used $y_t^2 = 1$. In this case (ii) holds with equality. Solving $\lambda$ from (2) and using (iii) gives
\[
\lambda = 2 \cdot \frac{1 - y_t w_t \cdot x_t}{\| x_t \|^2} \geq 0 \quad (3)
\]
which implies $1 - y_t w_t \cdot x_t \geq 0$. Plugging $\lambda$ from (3) into (1) yields
\[
w = w_t + \frac{1}{2} \cdot 2 \cdot \frac{1 - y_t w_t \cdot x_t}{\| x_t \|^2} y_t x_t \\
= w_t + y_t \frac{1 - y_t w_t \cdot x_t}{\| x_t \|^2} x_t.
\]
Putting finally all things together, we have the update rule
\[
w_{t+1} = w_t + \sigma_t y_t \frac{1 - y_t w_t \cdot x_t}{\| x_t \|^2} x_t
\]
where
\[
\sigma_t = \begin{cases} 
0 & \text{if } y_t w_t \cdot x_t \geq 1 \\
1 & \text{if } y_t w_t \cdot x_t < 1 .
\end{cases}
\]
Additionally, for all $\beta \in \mathbb{R}$, where $\beta \geq 0$, the idea of this exercise was to show that if there is a solution to the given minimization problem, then the coordinate $i$ of the new weight vector satisfies

$$w_{t+1,i} = \frac{1}{2} \frac{\partial f(w)}{\partial w_i} \beta_i x_i,$$

where $\beta \geq 1$. Notice that the formulation of the exercise is slightly revised.

Let $f : \mathbb{R}_+^d \to \mathbb{R}, f(w) = d_{+}(w,w_t)$. See some discussion about defining $f$ on the boundary of set $\mathbb{R}_+^d$ in the solution for the exercise 2b of the third homework. Similarly as in the previous exercise, we start by proving that $f$ is convex in $\mathbb{R}_+^d$, which we do by examining the Hessian matrix of $f$ on the interior of $\mathbb{R}_+$. Let $w = (w_1, w_2, \ldots, w_d) \in (\mathbb{R}_+ \setminus \{0\})^d$. For all $i \in \{1, 2, \ldots, d\}$

$$\frac{\partial f(w)}{\partial w_i} = \frac{\partial \sum_{i=1}^d w_i \ln(w_i/w_{t,i})}{\partial w_i} = \ln w_i - \ln w_{t,i} + 1,$$

and

$$\frac{\partial^2 f(w)}{\partial w_i \partial w_j} = \frac{\partial (\ln w_i - \ln w_{t,i} + 1)}{\partial w_i} = \frac{1}{w_i}.$$

Additionally, for all $i, j \in \{1, 2, \ldots, d\}, i \neq j$,

$$\frac{\partial^2 f(w)}{\partial w_i \partial w_j} = 0.$$
Let $z \in \mathbb{R}^{d \times 1}$. The Hessian of $f$ is positively semidefinite because

$$z^T H(f) z = ((z_1/w_1), (z_2/w_2), \ldots, (z_d/w_d)) z$$

$$= \sum_{i=1}^{d} \frac{z_i^2}{w_i}$$

$$\geq 0.$$  

We have now proved that $f$ is convex in $\mathbb{R}_+^d$.

This time we have in addition to inequality constrains also an equality constraint. Let

$$p(w) = -y_t \mathbf{w} \cdot \mathbf{x}_t,$$

$$q(w) = \sum_{i=1}^{d} w_i - 1.$$  

The conditions $w_i \geq 0$ are left out, because of the definition of $f$. The optimization problem is to minimize $f$ in the set $\mathbb{R}_+^d$ subject to

$$\begin{cases}
    p(w) \leq 0 \\
    q(w) = 0.
\end{cases}$$

The function $p$ is linear, and therefore convex. Let $a = (1, 1, \ldots, 1)$. We see that $q(w) = a \cdot w - 1$ is affine. So our minimization problem is convex.

The KKT conditions for the optimization problem are

$$\begin{cases}
    \nabla f(w) + \lambda \nabla p(w) + \gamma \nabla q(w) = 0 \\
    p(w) \leq 0 \\
    q(w) = 0 \\
    \lambda \geq 0 \\
    \lambda p(w) = 0.
\end{cases}$$

where $\gamma \in \mathbb{R}$. For all $i \in \{1, 2, \ldots, d\}$ it holds that

$$\frac{\partial f(w)}{\partial w_i} = \ln w_i - \ln w_{t,i} + 1$$

$$\frac{\partial p(w)}{\partial w_i} = y_t x_{t,i}$$

$$\frac{\partial q(w)}{\partial w_i} = 1$$

and the first KKT condition yields thus

$$(\ln w_i - \ln w_{t,i} + 1) - \lambda y_t x_{t,i} + \gamma = 0.$$  

This implies

$$\ln w_i = \ln w_{t,i} - 1 + \lambda y_t x_{t,i} - \gamma,$$

and further

$$w_i = w_{t,i} \exp(-1 + \lambda y_t x_{t,i} - \gamma) = \exp(-1 - \gamma) w_{t,i} \exp(\lambda y_t x_{t,i}) = \frac{1}{Z} w_{t,i} e^{\beta_t x_{t,i}}.$$
where $1/Z = \exp(-1 - \gamma)$ and $\beta_t = \exp(\lambda)$. The condition $q(w) = 0$ yields now

$$\sum_{i=1}^{d} \frac{1}{Z} w_{t,i} \beta_{t,i}^{x_{t,i}} - 1 = 0 \Rightarrow Z = \sum_{i=1}^{d} w_{t,i} \beta_{t,i}^{x_{t,i}},$$

and because of the condition $\lambda \geq 0$ we know that $\beta_t = \exp(\lambda) \geq 1$.

**Problem 4**

The accuracy with different choices of the kernel width and parameter $C$ are shown below. For both parameters, the values 1, 10 and 100 are tried.

Note that the smallest test set error is attained with $C = 1.0$ when $width = 1.0$ and with $C = 100.0$ when $width = 10.0$. When $width = 100.0$, the classifier breaks down completely, because the kernel size is much larger than the extent of the data set.

data = dlmread( 'hw05data.txt' );
X = data(:,1);
Y = data(:,2);
label = data(:,3);
N = size(data, 1);
negative_mask = find(label==-1);
positive_mask = find(label==1);
X_train = X(1:N/2);
Y_train = Y(1:N/2);
label_train = label(1:N/2);
X_test = X(N/2+1:end);
Y_test = Y(N/2+1:end);
label_test = label(N/2+1:end);
train_data=data(1:N/2, 1:2);
test_data=data(N/2+1:end, 1:2);
sigma = [1 10 100];
C = [1 10 100];
for sigma_i=1:3
    for C_i=1:3
        figure;
        model = svmtrain(train_data, label_train, 'kernel_function', 'rbf', 'rbf_sigma', sigma(sigma_i), ...
                       'boxconstraint', C(C_i), 'showplot', true)
        label_test_data_result = svmclassify(model, test_data, 'showplot', true);
        label_train_data_result = svmclassify(model, train_data);
        train_error = mean(label_train~=label_train_data_result)
        test_error = mean(label_test~=label_test_data_result)
    end
end
\texttt{title(sprintf('width=%.1f, C=%.1f, testerr.=%.3f, trainerr.=%.3f', \ldots \sigma(\sigma_i), C(C_i), test_error, train_error));}

\texttt{print \textasciitilde deps}

end

end