CS 194-10, Fall 2011
Assignment 2 Solutions

1. (8 pts) *In this question we briefly review the expressiveness of kernels.*

   (a) *Construct a support vector machine that computes the XOR function. Use values of +1 and -1 (instead of 1 and 0) for both inputs and outputs, so that an example looks like* $([-1,1],1)$ *or* $([-1,-1],-1)$. *Map the input* $[x_1,x_2]$ *into a space consisting of* x_1 *and* x_1x_2. *Draw the four input points in this space, and the maximal margin separator. What is the margin? Now draw the separating line back in the original Euclidian input space.*
   The examples map from $[x_1,x_2]$ to $[x_1,x_1\,x_2]$ coordinates as follows:
   $[-1,-1]$ (negative) maps to $[-1,+1]$
   $[-1,+1]$ (positive) maps to $[-1,-1]$
   $[+1,-1]$ (positive) maps to $[+1,-1]$
   $[+1,+1]$ (negative) maps to $[+1,+1]$
   Thus, the positive examples have $x_1x_2 = -1$ and the negative examples have $x_1x_2 = +1$. The maximum margin separator is the line $x_1x_2 = 0$, with a margin of 1. The separator corresponds to the $x_1 = 0$ *and* $x_2 = 0$ axes in the original space—this can be thought of as the limit of a hyperbolic separator with two branches.

   (b) *Recall that the equation of the circle in the 2-dimensional plane is* $(x_1-a)^2+(x_2-b)^2-r^2=0$. *Expand out the formula and show that every circular region is linearly separable from the rest of the plane in the feature space* (x_1,x_2,x_1^2,x_2^2).
   The circle equation expands into five terms

   $$0 = x_1^2 + x_2^2 - 2ax_1 - 2bx_2 + (a^2+b^2-r^2)$$

   corresponding to weights $w=(2a,2b,1,1)$ and intercept $a^2+b^2-r^2$. This shows that a circular boundary is linear in this feature space, allowing linear separability.
   In fact, the three features $x_1, x_2, x_1^2+x_2^2$ suffice.

   (c) *Recall that the equation of an ellipse in the 2-dimensional plane is* $c(x_1-a)^2+d(x_2-b)^2-1=0$. *Show that an SVM using the polynomial kernel of degree 2,* $K(\mathbf{u},\mathbf{v})=(\mathbf{1}+\mathbf{u}\cdot\mathbf{v})^{\mathbf{2}}$, *is equivalent to a linear SVM in the feature space* $(1,x_1,x_2,x_1^2,x_2^2,x_1x_2)$ *and hence that SVMs with this kernel can separate any elliptic region from the rest of the plane.*
   The (axis-aligned) ellipse equation expands into six terms

   $$0 = cx_1^2 + dx_2^2 - 2acx_1 - 2bdx_2 + (a^2c+b^2d-1)$$

   corresponding to weights $w=(2ac,2bd,c,d,0)$ and intercept $a^2+b^2-r^2$. This shows that an elliptical boundary is linear in this feature space, allowing linear separability.
   In fact, the four features x_1, x_2, x_1^2, x_2^2 suffice for any axis-aligned ellipse.

2. *(12 pts) Logistic regression is a method of fitting a probabilistic classifier that gives soft linear thresholds. (See Russell & Norvig, Section 18.6.4.) It is common to use logistic regression with an objective function consisting of the negative log probability of the data plus an* L_2 *regularizer:*

   $$L(\mathbf{w}) = -\sum_{i=1}^{N} \log\left(\frac{1}{1+e^{y_i(w^T\mathbf{x}_i+b)}}\right) + \lambda||\mathbf{w}||_2^2$$

   (Here $\mathbf{w}$ *does not include the "extra" weight* w_0*.)*

(a) *Find the partial derivatives* $\frac{\partial L}{\partial w_j}$.

First define the function $g(z) = \frac{1}{1+e^z}$. Note that $\frac{\partial g(z)}{\partial z} = g(z)(1-g(z))$ and also $\frac{\partial \log(g(z))}{\partial z} = \frac{1}{g(z)} g(z)(1-g(z)) = (1-g(z))$. Then we get,

$$\frac{\partial L}{\partial w_j} = -\sum_{i=1}^{n} y_i x_{ij}(1 - g(y_i(w^T x_i + b))) + 2\lambda w_j$$

(b) *Find the partial second derivatives* $\frac{\partial^2 L}{\partial w_j \partial w_k}$.

$$\frac{\partial^2 L}{\partial w_j \partial w_k} = \sum_{i=1}^{n} y_i^2 x_{ij} x_{ik} g(y_i(w^T x_i + b))(1 - g(y_i(w^T x_i + b))) + 2\lambda \delta_{jk}$$

where $\delta_{jk} = 1$ if $i = j$, 0 if $i \neq j$.

(c) *From these results, show that* $L(\mathbf{w})$ *is a convex function.*
Hint: A function L *is convex if its Hessian (the matrix* $\mathbf{H}$ *of second derivatives with elements* $H_{j,k} = \frac{\partial^2 L}{\partial w_j \partial w_k}$*) is positive semi-definite (PSD). A matrix* $\mathbf{H}$ *is PSD if and only if*

$$\mathbf{a}^T \mathbf{H} \mathbf{a} \equiv \sum_{j,k} a_j a_k H_{j,k} \geq 0$$

for all real vectors $\mathbf{a}$.
Applying the definition of PSD matrix to the Hessian of $L(\mathbf{w})$ we get,

$$\begin{aligned} \mathbf{a}^T \mathbf{H} \mathbf{a} &= \sum_{j,k} a_j a_k \frac{\partial^2 L}{\partial w_j \partial w_k} = \sum_{j,k} a_j a_k \left(\sum_i y_i^2 x_{ij} x_{ik} g(y_i(w^T x_i + b))(1 - g(y_i(w^T x_i + b))) + 2\lambda \delta_{jk} \right) \quad (1) \\ &= \sum_{j,k,i} a_j a_k y_i^2 x_{ij} x_{ik} g(y_i(w^T x_i + b))(1 - g(y_i(w^T x_i + b))) + 2\lambda \sum_j a_j^2 \quad (2) \end{aligned}$$

Define $P_i = g(y_i(w^T x_i + b))(1 - g(y_i(w^T x_i + b)))$ and $\rho_{ij} = y_i x_{ij} \sqrt{P_i}$. Then,

$$\mathbf{a}^T \mathbf{H} \mathbf{a} = \sum_{j,k,i} a_j a_k x_{ij} x_{ik} y_i^2 P_i + 2\lambda \sum_j a_j^2 = \sum_i a^T \rho_i \rho_i^T a + 2\lambda \sum_j a_j^2 = \sum_i (a^T \rho_i)^2 + 2\lambda \sum_j a_j^2 \geq 0$$

for $\lambda \geq 0$.

3. *(8 pts) Consider the following training data,*

class	x_1	x_2
+	1	1
+	2	2
+	2	0
−	0	0
−	1	0
−	0	1

(a) *Plot these six training points. Are the classes* $\{+, -\}$ *linearly separable?*
As seen in the plot the classes are linearly separable.

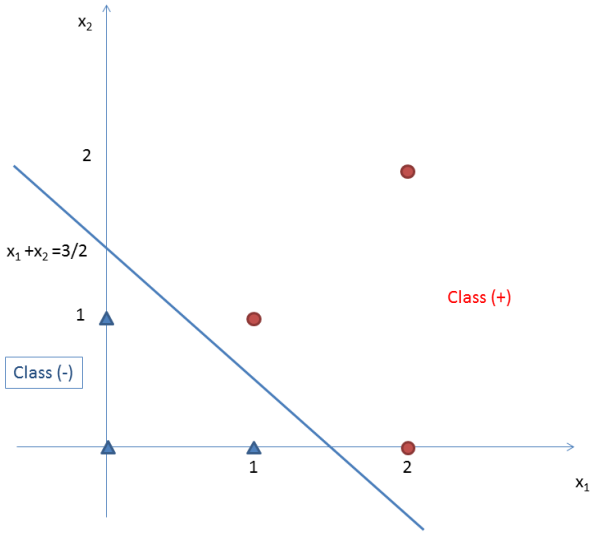

Figure 1: Question 3

(b) *Construct the weight vector of the maximum margin hyperplane by inspection and identify the support vectors.*
The maximum margin hyperplane should have a slope of -1 and should satisfy $x_1 = 3/2, x_2 = 0$. Therefore it's equation is $x_1 + x_2 = 3/2$, and the weight vector is $(1, 1)^T$.

(c) *If you remove one of the support vectors does the size of the optimal margin decrease, stay the same, or increase?* In this specific dataset the optimal margin increases when we remove the support vectors $(1, 0)$ or $(1, 1)$ and stays the same when we remove the other two.

(d) *(Extra Credit) Is your answer to (c) also true for any dataset? Provide a counterexample or give a short proof.*
When we drop some constraints in a constrained maximization problem, we get an optimal value which is *at least as good* the previous one. It is because the set of candidates satisfying the original (larger, stronger) set of contraints is a subset of the candidates satisfying the new (smaller, weaker) set of constraints. So, for the weaker constraints, the oldoptimal solution is still available and there may be additions soltons that are even better. In mathematical form:

$$\max_{x \in A, x \in B} f(x) \leq \max_{x \in A} f(x) \ .$$

Finally, note that in SVM problems we are maximizing the margin subject to the constraints given by training points. When we drop any of the constraints the margin can increase or stay the same depending on the dataset. In general problems with realistic datasets it is expected that the margin increases when we drop support vectors. The data in this problem is constructed to demonstrate that when removing some constraints the margin can stay the same or increase depending on the geometry.

4. (12 pts) *Consider a dataset with 3 points in 1-D:*

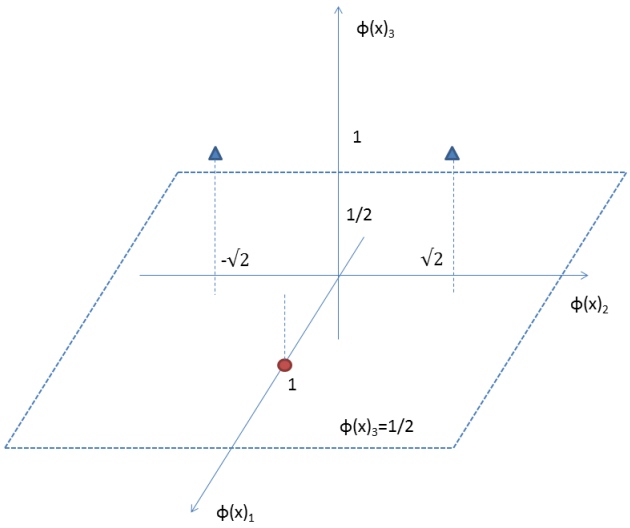

Figure 2: Question 4

(class)	x
+	0
−	−1
−	+1

(a) *Are the classes* $\{+,-\}$ *linearly separable?*
Clearly the classes are not separable in 1 dimension.

(b) *Consider mapping each point to 3-D using new feature vectors* $\boldsymbol{\phi}(x) = [1, \sqrt{2}x, x^2]^T$. *Are the classes now linearly separable? If so, find a separating hyperplane.*
The points are mapped to $(1,0,0), (1,-\sqrt{2},1), (1,\sqrt{2},1)$ respectively. The points are now separable in 3-dimensional space. A separating hyperplane is given by the weight vector (0,0,1) in the new space as seen in the figure.

(c) *Define a class variable* $y_i \in \{-1,+1\}$ *which denotes the class of* x_i *and let* $\mathbf{w} = (w_1, w_2, w_3)^T$. *The max-margin SVM classifier solves the following problem*

$$\min_{\boldsymbol{w},b} \ \tfrac{1}{2}||\mathbf{w}||_2^2 \ \textit{s.t.} \tag{3}$$

$$y_i(\mathbf{w}^T\boldsymbol{\phi}(x_i) + b) \geq 1, \ \ i = 1,2,3 \tag{4}$$

Using the method of Lagrange multipliers show that the solution is $\hat{\mathbf{w}} = (0,0,-2)^T$, $b = 1$ *and the margin is* $\frac{1}{||\hat{\mathbf{w}}||_2}$.

For optimization problems with inequality constraints such as the above, we should apply KKT conditions which is a generalization of Lagrange multipliers. However this problem can be solved easier by noting that we have three vectors in the 3-dimensional space and all of them are support vectors. Hence the all 3 constraints hold with equality. Therefore we can apply the method of Lagrange multipliers to,

$$\min_{\boldsymbol{w},b} \ \tfrac{1}{2}||\mathbf{w}||_2^2 \ \text{s.t.} \tag{5}$$

$$y_i(\mathbf{w}^T\boldsymbol{\phi}(x_i) + b) = 1, \ \ i = 1,2,3 \tag{6}$$

We have 3 constraints, and should have 3 Lagrange multipliers. We first form the Lagrangian function $L(\mathbf{w}, \boldsymbol{\lambda})$ where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ as follows

$$L(\mathbf{w}, \boldsymbol{\lambda}) = \frac{1}{2}||\mathbf{w}||_2^2 + \sum_{i=1}^{3} \lambda_i (y_i(\mathbf{w}^T \boldsymbol{\phi}(x_i) + b) - 1) \tag{7}$$

and differentiate with respect to optimization variables $\mathbf{w}$ and b and equate to zero,

$$\frac{\partial L(\mathbf{w}, \boldsymbol{\lambda})}{\partial \mathbf{w}} = \mathbf{w} + \sum_{i=1}^{3} \lambda_i y_i \boldsymbol{\phi}(x_i) = 0 \tag{8}$$

$$\frac{\partial L(\mathbf{w}, \boldsymbol{\lambda})}{\partial b} = \sum_{i=1}^{3} \lambda_i y_i = 0 \quad . \tag{9}$$

Using the data points $\boldsymbol{\phi}(x_i)$, we get the following equations from the above lines,

$$w_1 \quad +\lambda_1 - \lambda_2 - \lambda_3 = 0 \tag{10}$$

$$w_2 \quad +\sqrt{2}\lambda_2 - \sqrt{2}\lambda_3 = 0 \tag{11}$$

$$w_3 \quad -\lambda_2 - \lambda_3 = 0 \tag{12}$$

$$\lambda_1 \quad -\lambda_2 - \lambda_3 = 0 \tag{13}$$

$$\tag{14}$$

Using (10) and (14) we get $w_1 = 0$. Then plugging this to equality constraints in the optimization problem, we get

$$b = 1 \tag{15}$$

$$-\sqrt{2}w_2 + w_3 + b = -1 \tag{16}$$

$$+\sqrt{2}w_2 + w_3 + b = -1 \tag{17}$$

$$\tag{18}$$

(16) and (17) imply that $w_2 = 0$, and $w_3 = -2$. Therefore the optimal weights are $\hat{\mathbf{w}} = (0, 0, -2)^T$ and $b = 1$. And the margin is $1/2$.

(d) *Show that the solution remains the same if the constraints are changed to*

$$y_i(\mathbf{w}^T \boldsymbol{\phi}(x_i) + b) \geq \rho, \ i = 1, 2, 3$$

for any $\rho \geq 1$.
Note that changing the constraints in the solution of part (c) only changes equation (15-17) and we get $b = \rho$, and $\hat{\mathbf{w}} = (0, 0, -2\rho)^T$. However the hyperplane described by the equation, $\hat{\mathbf{w}}^T\mathbf{x} + b = 0$, remains the same as before: $\{\mathbf{x} : -2\rho x_3 + \rho = 0\} \equiv \{\mathbf{x} : -2x_3 + 1 = 0\}$. Hence, we have the same classifier in two cases: assign class label + if $\hat{\mathbf{w}}^T\mathbf{x} + b \geq 0$ and assign class − otherwise.

(e) *(Extra Credit) Is your answer to (d) also true for any dataset and* $\rho \geq 1$*? Provide a counterexample or give a short proof.*
This is true for any dataset and it follows from the homogeneity of the optimization problem. For constraints $y_i(\mathbf{w}^T \boldsymbol{\phi}(x_i) + b) \geq \rho$, we can define new weight vectors $\tilde{\mathbf{w}} = \mathbf{w}/\rho$ and $\tilde{b} = b/\rho$. So that the constraints in the new variables are $y_i(\tilde{\mathbf{w}}^T \boldsymbol{\phi}(x_i) + \tilde{b}) = 1$. And equivalently optimize the following,

$$\min_{\tilde{w}, \tilde{b}} \ \tfrac{1}{2}\rho^2 ||\tilde{\mathbf{w}}||_2^2 \text{ s.t.} \tag{19}$$

$$y_i(\tilde{\mathbf{w}}^T \boldsymbol{\phi}(x_i) + \tilde{b}) \geq 1, \ i = 1, 2, 3 \tag{20}$$

Since ρ^2 is a constant multiplying the objective function $||\tilde{\mathbf{w}}||_2^2$, it does not change the optimal value, and the two solutions describe the same classifier $\mathbf{w}^T\mathbf{x} + b \geq 0 \equiv \rho\tilde{\mathbf{w}}^T\mathbf{x} + \rho\tilde{b} \geq 0$.