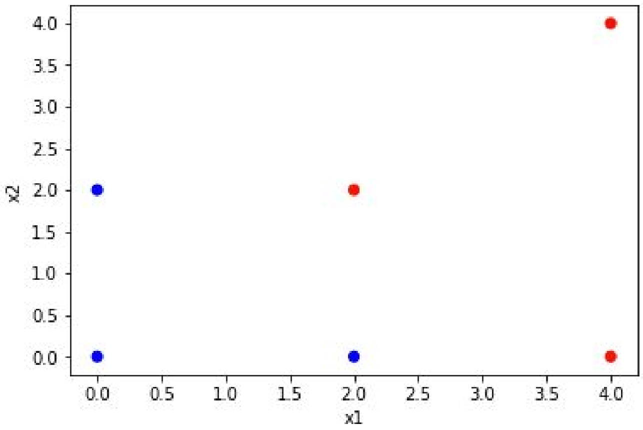
Machine Learning Problem Set 3- Support Vector Machines
Hannah Lyon

Problem 1:

a) Yes, these classes do appear to be linearly separable.



b) By inspection, we find that the separating hyperplane has a slope of -1 and when $x_1=0$, $x_2=3$. The equation is thus $x_1+x_2=3$ and the weight vector, **w**, is $(1,1)^T$.

c) If you were to remove one of the support vectors, especially points (2,2) and (2,0), then the optimal margin will increase, as there will be more space separating the points from the two classes.

Problem 2:

a) Given a separating hyperplane $w \bullet x + b = 0$ with margins of equations $w \bullet x + b =- 1$ and $w \bullet x + b = 1$, we take the nearest training example to the hyperplane (the support vectors). This point, x, lies on the margins of the maximum separating hyperplane and thus satisfies $w \bullet x + b = 1$. We define the closest point on the hyperplane to x as point z, which satisfies $w \bullet z + b = 0$. The distance between the two, or the margin, is $|x - z|$.

If we then normalize the distance by dividing it by the margin length, which we define as $\delta = |x - z|$, then we have the equality $\frac{|x-z|}{\delta} = 1$ and because 1 is the unit vector, $\frac{w}{|w|}$, we can form the equivalent expression $\frac{x-z}{\delta} = \frac{w}{|w|}$.

Isolating z, we get $z = x - \frac{w}{|w|}\delta$. Plugging this into $w \bullet z + b = 0$ gives us:

$$w\left(x - \frac{w}{|w|}\delta\right) + b = 0$$

$w \bullet x - \frac{|w|^2}{|w|}\delta + b = 0$
Isolating δ reveals $\delta = \frac{w \bullet x + b}{|w|}$ and because the support vector satisfies $w \bullet x + b = 1$, the numerator is equal to one and the margin width, $\delta = \frac{1}{|w|}$.

b) If z and d are both multiplied by $\frac{1}{|z|M}$, a factor of 1 because the margin, $M = \frac{1}{|Z|}$ (as demonstrated in part (a)), then the hard margin support vector machine problem is still feasible using z' and d'. First, because multiplying the factors by 1 does nothing, but also because the problem's restraints, $y^i(z \bullet x^i + d) \geq 1$, do not depend on magnitude so z and d can be multiplied by any constant. The resulting equation would be $y^i(\frac{z}{|z|M} \bullet x^i + \frac{d}{|w|M}) \geq 1$

Additionally, because $z' = \frac{z}{|z|M}$, then $|z|^2 = \frac{z}{|z|M} \bullet \frac{z}{|z|M}$. Further if we define the margin as $M = \frac{1}{|w|}$, then $|w|^2 = \frac{1}{M^2}$.
$\frac{1}{M^2} \leq \frac{z}{|z|M} \bullet \frac{z}{|z|M}$ so $|w|^2 \leq |z'|^2$ and therefore $|w| \leq |z'|$.

c) Given $z' = \frac{z}{|z|M}$, isolating M gives us $M = \frac{z}{z'|z|}$ which is the margin for z, d. We saw before that the margin of w, b was $\frac{1}{|w|}$. And because $\frac{z}{z'|z|} \leq \frac{1}{|w|}$, w, b thus form a maximizing margin for our support vector machine problem.

Problem 3:

a) There is no line that can completely separate the points of each class from one another (as stipulated in hard margin support vector machine problems), and the problem is therefore infeasible. Every possible line would misclassify at least one of the training points.

b) The soft margin support vector machine problem is an optimization problem with the goal of minimizing the equation $w \bullet w + C \sum_{i=1}^{4} \xi^i$ with respect to w, b, and ξ and subject to:
$(w \bullet x^i + b)y^i \geq 1 - \xi^i$ for all i with $\xi^i \geq 0$ for all i.

Explicitly, given $w = (w_o, w_1)$, $\underline{x}^1 = (x_1{}^1, x_2{}^1)$, ..., $\underline{x}^4 = (x_1{}^4, x_2{}^4)$, $y^1, \ldots, y^4$, and $\xi^1, \ldots, \xi^4$, this calls for minimizing the equation $w_0{}^2 + w_1{}^2 + C(\xi^1 + \xi^2 + \xi^3 + \xi^4)$ subject to the following equations:
$(w_0 x_1{}^1 + w_1 x_2{}^1 + b)y^1 \geq 1 - \xi^1$
$(w_0 x_1{}^2 + w_1 x_2{}^2 + b)y^2 \geq 1 - \xi^2$
$(w_0 x_1{}^3 + w_1 x_2{}^3 + b)y^3 \geq 1 - \xi^3$
$(w_0 x_1{}^4 + w_1 x_2{}^4 + b)y^4 \geq 1 - \xi^4$
$\xi^1, \xi^2, \xi^3, \xi^4 \geq 0$

So a potential solution could be the hyperplane $w = (1.5,\ 0)$ and $b = 0$. If the margin is 1.5 and C=1, then the minimizing equation is $1.5^2 + (3 + 3)\ = 8.25$

Problem 4:

a) Yes, a function that finds the intersection of two sets of words would be a valid kernel. This function would consist of feature vectors with a feature for each unique word in each document and the complete set of interaction terms between the two. The kernel denoting this would be quadratic kernel, $K(x, z) = (x^T z)^{\ 2}$, which is equivalent to $\phi(x)^T \phi(z)$ and the feature mapping of $\phi(x)\ =\ ($x$_1$x$_1$, x$_1$x$_2$, …, x$_1$x$_n$, x$_2$x$_1$, …, x$_n$x$_n$).

b) $K(x, z) = (1 + \beta x \bullet z)^{\ 2} - 1$

$x = [x_1,\ x_2]\ \mathit{and}\ z = [z_1, z_2]$

$$K(x, z) = \left[\beta \sum_{j=1}^{2} (x^j z^j) + 1\right]^2 - 1$$

$$= \beta^2 \sum_{j,\,l=1}^{2} (x^j x^l)(z^j z^l) + 2\beta \sum_{j=1}^{2} (x^j z^j) + 1 - 1$$

$$= \sum_{j,\,l=1}^{2} (\beta x^j x^l)(\beta z^j z^l) + \sum_{j=1}^{2} (\sqrt{2\beta}\, x^j)\,(\sqrt{2\beta}\, z^j)$$

The feature vector is $\Phi(x) = \left[\beta x^{(1)2},\ \beta x^{(1)} x^{(2)},\ \beta x^{(2)} x^{(1)},\ \beta x^{(2)2},\ \sqrt{2\beta}\, x^{(1)},\ \sqrt{2\beta}\, x^{(2)}\right]$

Problem 5:

a) See code.

b) The test error of the model is 0.078, or 7.8%.

c) The five fold cross validation scores were[0.91, 0.9225, 0.9175, 0.9225, 0.925] given default C and γ values, so the errors were [0.09, 0.0775, 0.0825, 0.0775, 0.075]. Next, in trying out different gamma and C values, I tested all combinations of gamma = [0.005, 0.01, 0.1, 0.5, 1] and C = [0.01, 1, 3, 5, 10] and noticed that the best gamma score across all the C values was gamma = 0.005. After finding this, I used tested more C values and found that a similar error rate was found for all C values above 1, so I left C as 1. The five fold cross validation scores for this combination was [0.9425, 0.935, 0.945, 0.9525, 0.9475] and the test error was 0.053.

Problem 6:

Show that the Lagrangian optimization problem for support vector machines is equal to the maximization of the dual problem.

We start with the Lagrangian: $L(w, b, \alpha) = \frac{1}{2}\, \|w\|^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w \bullet x^{(i)} + b) - 1\right]$

Then, we fix α and the find the partial derivatives of the Lagrangian with respect to w and b.

$$\partial L(w, b, \alpha) / \partial w = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0$$

$$\partial L(w, b, \alpha)\ / \partial b = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

If we then isolate w, we can plug it back into the original Lagrangian equation.

$$w = \sum_{i=1}^{m} \alpha_i\, y^{(i)} x^{(i)}$$

$$L(w, b, \alpha) = \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)} \bullet x^{(j)}) - \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)} x^{(j)} - b\alpha_i y^{(i)} + \alpha_i$$

$$L(w, b, \alpha) = \ -\frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)} \bullet x^{(j)}) - b \sum_{i=1}^{m} \alpha_i y^{(i)} + \sum_{i=1}^{m} \alpha$$

If we then reorganize the terms and substitute 0 for the $-b \sum_{i=1}^{m} \alpha_i y^{(i)}$ term as we found when solving for the partial derivative of b, we get the dual problem that, when maximized in regards to α, solves the support vector machine problem.

$$L(w, b, \alpha) = \sum_{i=1}^{m} \alpha \ -\frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)} \bullet x^{(j)})$$

$$\max_{\alpha} \ \sum_{i=1}^{m} \alpha \ -\frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)} \bullet x^{(j)})$$

subject to $\alpha_i \geq 0,\ i = 1,\ \ldots,\ m$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$