EE645 Spring 2007
Problem Set 2 Solutions

1) For this problem the more general case of kernels was considered. Inputs are
transformed into feature space via functions $\phi(x)$ with $K(x, y) = \langle \phi(x), \phi(y) \rangle$. For
linear SVM the kernels are given by $K(x, y) = \langle x, y \rangle$.

The cost function are given by

$$
\min J(w, \xi) = \min .5||w||^2 + C \sum_{i=1}^{l} \xi_i
$$

subject to

$$y_i((w \cdot \phi(x_i)) + b) \geq 1 - \xi_i, \quad 1 \leq i \leq l$$

and $\xi_i \geq 0$ we first modify the function to include constraints. The Lagrangian
function in primal space is then

$$L_P = .5||w||^2 + C \sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \alpha_i(y_i((w \cdot \phi(x_i)) + b) \geq 1 - \xi_i) - \sum_{i=1}^{l} \mu_i \xi_i$$

where $\alpha_i \geq 0$ are Lagrangian multipliers to enforce the constraints on the margins
and $\mu_i \geq 0$ are Lagrangian multipliers to enforce positivity of the $\xi_i$. We then find
partial derivatives with respect to $w$, $b$, $\xi_i$ and set to zero to get

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i \tag{1}$$

and

$$0 = \sum_{i=1}^{l} \alpha_i y_i \tag{2}$$

$$C = \alpha_i + \mu_i$$

By substituting for $w$ into the Lagrangian equation we get the dual programming
problem

$$\max W(\alpha) = \max \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j).$$

where $K(x_i, x_j) = \phi(x_i)\phi(x_j)$ subject to equation (2) and the constraints on $\alpha_i$ that
$0 \leq \alpha_i \leq C$. A solution satisfies the KKT conditions:

$$\alpha_i (y_i ((w \cdot \phi(x_i)) + b) \geq 1 - \xi_i) = 0$$
\[ \mu_i \xi_i = 0 \]

2) For the iris data set, it has three classes of data: the first set is linearly separable from the other sets and the third set is almost linearly separable from the first two sets (a classifier can be constructed with one error). The second set cannot be separated from the first and third sets via a linear threshold function. The matlab quadprog routine was used to find the SVM solution. Solution found in the dual space.

a) Weight is \( w = [-.046, .5217, -1.032, .4642]^T \) and threshold \( b = 1.4506 \). Cost function is .7481 and there are three support vectors associated with \( x(24) = [5.1, 3.3, 1.7, .5]^T \), \( x(42) = [4.5, 2.3, 1.3, .3]^T \), and \( x(99) = [5.1, 2.5, 3.0, 1.1] \). The respective values of \( \alpha \) are .6715, .0717, .7481. Regularization term of \( \gamma = 1 \) was used. All points are correctly classified.

b) Weight is \( w = [-.0275, -1.1588, .5583, -1.2434]^T \) and threshold \( b = 2.7998 \). Cost function is \(-171.1057 \). There are 89 support vectors with five support vectors on margin \( x(46), x(54), x(73), x(81), x(139) \). Regularization term of \( \gamma = 2 \) was used. Error rate was .2333. Here a nonlinear classifier would do much better.

c) Weight is \( w = [-.5955, -.9759, 2.0322, 2.0061]^T \) and threshold \( b = -6.7811 \). Cost function is \(-15.7599 \). There are 23 support vectors with four support vectors on margin \( x(77), x(130), x(147), x(148) \). Regularization term of \( \gamma = 1 \) was used. Error rate was .0067 with \( x(84) \) misclassified.

3) For the Wisconsin Breast Cancer data the linear SVM did quite well. For linear SVM the average test error rate was about 2.7% For this problem good regularization values are when \( C \approx 1.4 \). This problem is close to being linearly separable as we are working in 30 dimensional input space and we usually usually have two to eight misclassified points on the 189 test data. We used SVM light which works reasonably quickly. For this data set we wrote separate C code to input data into SVM light format and randomly choose training and test data.

4)

a) For one observation the log-likelihood ratio gives

\[ l(x) = \log(f_{X|D}(x|1)/f_{X|D}(x|-1)) = 2x_1 + 4x_2 \]

as we take the ratio of two Gaussian pdfs with different means and the same covariance. For minimum error probability when \( P(D = 1) = P(D = -1) = .5 \) we get a linear discriminant described by \( f(x) = \text{sign}(x_1 + 2x_2) \). The sufficient statistic is given by \( l(x) \). When \( D = 1 \), \( l(x) \) is Gaussian with mean 5 and
variance 5. When $D = -1$, $l(x)$ is Gaussian with mean -5 and variance 5. The probability of error is then given by

$$P_e = Q(\sqrt{5}) = \int_{\sqrt{5}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-.5x^2)dx \approx .0125$$

For $m$ observations we get that

$$P_e = Q(\sqrt{5m}) \approx 1/\sqrt{10\pi m} \exp(-2.5m)$$

b) For the SVM we construct here we compare to the optimal Bayesian error for one observation which is approximately .0125. For SVM simulations we found the following results.

<table>
<thead>
<tr>
<th>SVM Error (100 simulations)</th>
<th>m=10</th>
<th>m=25</th>
<th>m=50</th>
<th>m=100</th>
<th>m=200</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>average MSE</td>
<td>.0310</td>
<td>.0207</td>
<td>.0187</td>
<td>.0169</td>
<td>.0146</td>
</tr>
<tr>
<td>dev. of MSE</td>
<td>.0219</td>
<td>.00105</td>
<td>.0092</td>
<td>.0056</td>
<td>.0044</td>
</tr>
</tbody>
</table>

As we choose more training samples the SVM classifier gets closer to the Bayesian optimal classifier.

5) For Fisher Linear Discriminant Analysis (FLDA) results are generally not as good as SVM results. FLDA results can be improved by adding regularization parameter. We will show how to do this later.

a) For iris data set:


second part $w = [-1311, -1.4928, .7129, -.7580]^T$ and threshold $b = 3.5857$.
Threshold was not set at midpoint between two means, but set at a point to minimize errors. Error rate was .3133.

Threshold was not set at midpoint between two means, but set at a point to minimize errors. Error rate was .0533.

b) We used a LS SVM which is very similar to the Fisher Linear Discriminant. We got similar results to the linear SVM when using a linear LS SVM. The LS SVM was implemented on matlab and the code ran much quicker than the SVM code which was implemented using C. LS SVM involves finding a solution to $m + 1$ linear equations as opposed to solving a QP with $m$ inequality constraints.