1. The following figure shows the city points when mapping the distance between each city into two dimensions, using classical MDS. The mapping matches intuition: we expect that cities far away from each other spatially will also be far away from each other in terms of feature inner products.

The code used implement this from the provided distance matrix $D$ was:

```r
# ONE is a vector of 1's, of length 10
G = -.5*(I.n - (1/10)*ONE %*% t(ONE)) %*% D %*% (I.n - (1/10)*ONE %*% t(ONE))

eig = eigen(G)

x1 = sqrt(eig$values[1]) * eig$vectors[,1]
x2 = sqrt(eig$values[2]) * eig$vectors[,2]
```
2. (a)

\[ R(f) = E(L(Y, f(x)) \]
\[ = E_x[E(L(Y, f(x))|X] \]
\[ = \sum_{k=1}^{K} E_x[L(k, f(x))P(Y = k|X = x)] \]
\[ = \sum_{k=1}^{K} E_x[I(k \neq f(x))\eta_k(x)] \]
\[ = E_x[\sum_{k=1}^{K} I(k \neq f(x))\eta_k(x)] \]
\[ = E_x[1 - \eta_{k_0}(x)] \]

Where \( k_0 \) is the class \( f \) predicted. Minimizing the last expression requires that \( \eta_{k_0}(x) \) be the largest of all the \( \eta_y \), so the optimal classifier \( f \) is as requested.

(b)

\[ R(f) - R(f^*) = E_x[\sum_{k=1}^{K} I(k \neq f(x))\eta_k(x) - \sum_{k=1}^{K} I(k \neq f^*(x))\eta_k(x)] \]
\[ = E_x[(1 - P(Y = f(X)|X = x)) - (1 - \max_{y \in Y} \eta_y(x))] \]
\[ = E_x[\max_{y \in Y} \eta_y(x) - P(Y = f(X)|X = x)] \]

(c) Let

\[ \hat{\eta}_y = \arg \max_{y \in Y} \hat{\eta}_y(x) \]

and

\[ \eta_y^* = \arg \max_{y \in Y} \eta_y(x) \]

and note that

\[ \hat{\eta}_y - \eta_y \geq 0 \]
Then we have:

\[
|\eta_y^* + \hat{\eta}_y| \leq |\eta_y - \eta_y^*| + |\hat{\eta}_y - \hat{\eta}_y^*| \\
= |\eta_y - \hat{\eta}_y^* + \hat{\eta}_y - \eta_y| \\
\leq |\eta_y - \hat{\eta}_y^*| + |\hat{\eta}_y - \eta_y| \\
\leq 2 \arg \max_{y \in Y} |\eta_y - \hat{\eta}_y|
\]

Where the last line is our upper bound.

3. (a), (b), (c), (d), (e), (f)

The classification errors for the methods were as follows:

- Generative (LDA)......0.22
- IRLS......................... 0.1066
- Stochastic...................0.1133
- linear regression.........0.18

So in this instance, the IRLS method predicted the test data most accurately. Note that when implementing the stochastic method, I chose a step size equal to 1, and initial \( \theta^{(0)} = 0 \). These settings gave a powerful classifier of the training data.
With less training instances, the line will invariably be less accurate.

More specifically, in the generative approach, the estimate of $\beta$ will not change too much, but $\gamma$ will increase from by a value close close to $\log\left(\frac{25}{25}\right) \approx 1.1$ Thus the classifier will be more likely to predict 1, which means the line will push away from the cluster of 1’s and into the cluster of 0’s.

In the IRLS approach, 0’s will be harder to predict, implying that $W$ will give more weight to those instances with 0, so the line should not change too much.

4. (a)

The following are the error rates for the various transformations under logistic regression:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>standardized</td>
<td>0.0717</td>
<td>0.0704</td>
<td>0.0577</td>
<td>0.0567</td>
<td>0.0571</td>
<td>0.0809</td>
</tr>
<tr>
<td>log transform</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>discrete transform</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b)

When applying the LDA function I built, I obtain the following error rates:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>standardized</td>
<td>0.1125</td>
<td>0.1070</td>
<td>0.0858</td>
<td>0.0776</td>
<td></td>
<td></td>
</tr>
<tr>
<td>log transform</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>discrete transform</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(c)

This table presents the errors rates amongst the testing sets for the various algorithms on the spam dataset.

<table>
<thead>
<tr>
<th>Data</th>
<th>LDA</th>
<th>Logistic</th>
<th>Linear Regression</th>
<th>Naive Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardized</td>
<td>.1070</td>
<td>.0704</td>
<td>.1096</td>
<td>-</td>
</tr>
<tr>
<td>Log Transform</td>
<td>.0776</td>
<td>.0567</td>
<td>.0691</td>
<td>-</td>
</tr>
<tr>
<td>Discrete Transform</td>
<td>.1141</td>
<td>.0809</td>
<td>.0861</td>
<td>.0769</td>
</tr>
</tbody>
</table>