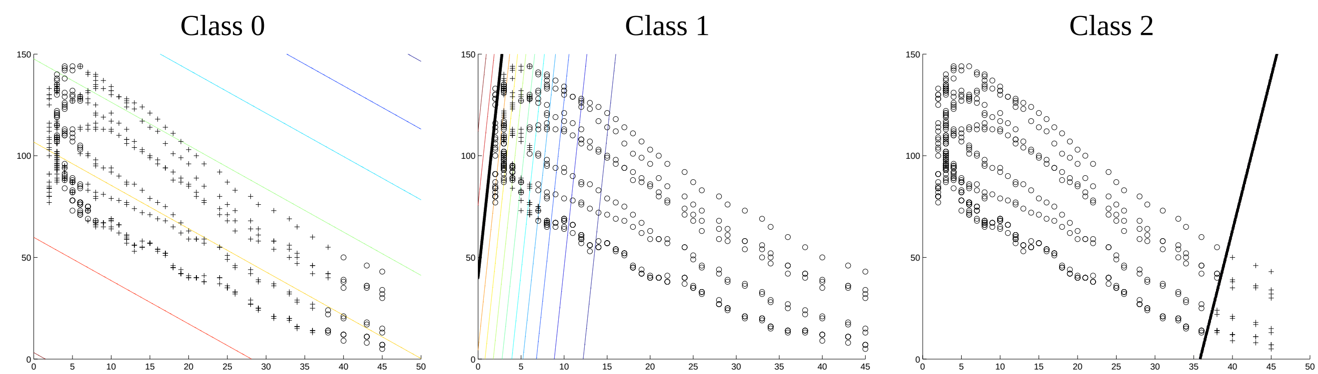
Homework 3 Sample Solutions
COMP 766-001 – Machine Learning for Bioinformatics

[1] (A) The weights, optimized by `fminsearch` in Matlab, are

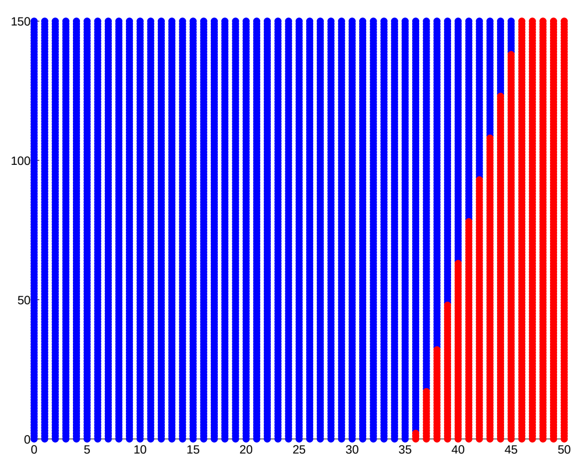
Class	w_0	w_1	w_2
0	1.7542	-0.0131	-0.0061
1	-0.2251	-0.2226	0.0056
2	-3637.8	101.5724	-6.7142

where w_0 is the offset term, w_1 multiplies x_1 (*bicoid* expression), and w_2 multiplies x_2 (*caudal* expression).

(B) The three plots below show the data again and the contours of the fit logstic surface for classes $i \in \{0, 1, 2\}$. For each i, the examples of that class are plotted as a "+" and the other examples as an "o". The dark line is the decision boundary, and the redder lines indicate regions with higher probability of membership in class i. (Note that the decision boundary for class $i = 0$ does not appear within the region plotted, so cannot be seen.)



(C) The plot below shows, by the color of the dots, which class is the most likely (i.e. which logistic evaluates highest) in different regions of the input space.



(D) The class 0 logistic is very broad, and not that informative. The logistic for class 1 captures the fact that the class 1 points are towards the upper left. Interestingly, the decision boundary occurs further to the left (small x_1) than all the actual class 1 examples. This, apparently, is because there are also a number of class 0 examples at the extreme left, preventing a clear decision. This is a case where we might want to add

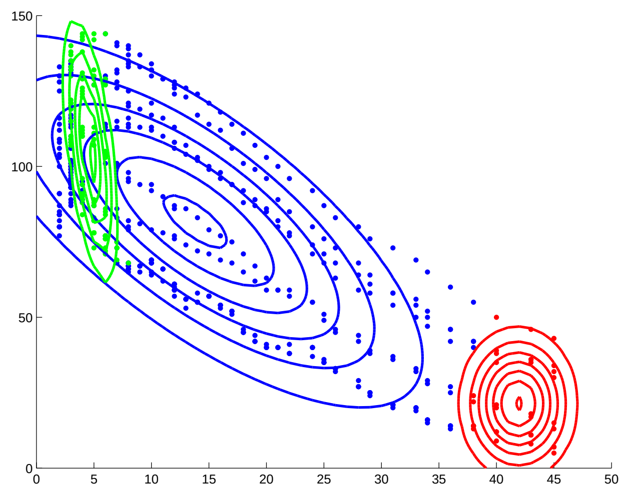
x_1^2 as a feature—alllowing the classification to be nonmonotonic in x_1, in particular, so that an intermediate range of x_1 could be selected as most likely for class 1. The logistic for class 2 is able to exactly separate the “+” and “-” points, resulting in a sharp decision boundary. This is what is called a *linearly separable* classification problem, and is not to be expected as the general case. In part C, we see that the difficulty in capturing class 1 makes it nowhere the most likely class. Class 2 is the most likely class on the right side of the corresponding logistic’s decision boundary.

[2] (A) The means and covariance matrcies of the maximum likelihood Gaussian fits two the three classes are given below, alongside plots of the Gaussians.

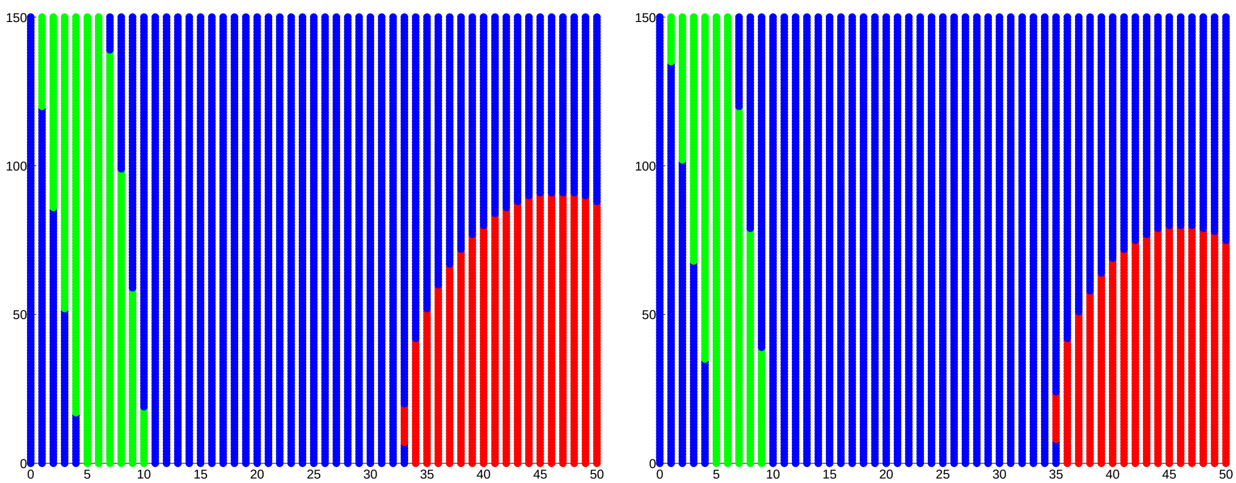
$$\mu_0 = \begin{bmatrix} 13.83 \\ 81.75 \end{bmatrix} \qquad \Sigma_0 = \begin{bmatrix} 106.51 & -244.20 \\ -244.20 & 1039.94 \end{bmatrix}$$

$$\mu_1 = \begin{bmatrix} 4.70 \\ 105.20 \end{bmatrix} \qquad \Sigma_1 = \begin{bmatrix} 1.59 & -17.60 \\ -17.60 & 557.72 \end{bmatrix}$$

$$\mu_2 = \begin{bmatrix} 41.89 \\ 21.42 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 6.73 & -0.18 \\ -0.18 & 166.19 \end{bmatrix}$$



(B), (C) The most likely class, as a function of the two input variables is shown below. The left plot assumes equal priors for the classes. The right plot assumes priors taken from the empirical frequencies.



(D) The three-class Gaussian fit is more satisfying than the logistic fits, mainly because all three classes have regions where they are most probable. By adding the quadratic terms to the logistic regression, one could probably get the same effect. However, the three-class Gaussian fits are also very computationally efficient. To me, it also just seems a more natural way of analyzing data with more than two possible output classes.

[3] The follow analysis assumes uniform priors, $P(y = i) = \frac{1}{3}$ for $i \in \{0, 1, 2\}$, and that x is a length d vector.

$$
\begin{aligned}
& P(y=i|x) > P(y=j|x) \\
\iff \quad & P(x|y=i)P(y=i)/P(x) > P(x|y=j)P(y=j)/P(x) \\
\iff \quad & P(x|y=i) > P(x|y=j) \\
\iff \quad & \frac{1}{(2\pi)^{d/2}|\Sigma_i|} \exp(-\frac{1}{2}(x-\mu_i)^T \sigma^{-2} I (x-\mu_i)) > \frac{1}{(2\pi)^{d/2}|\Sigma_j|} \exp(-\frac{1}{2}(x-\mu_j)^T \sigma^{-2} I (x-\mu_j)) \\
\iff \quad & \exp(-\frac{1}{2\sigma^2}\|x-\mu_i\|^2) > \exp(-\frac{1}{2\sigma^2}\|x-\mu_j\|^2) \\
\iff \quad & -\frac{1}{2\sigma^2}\|x-\mu_i\|^2 > -\frac{1}{2\sigma^2}\|x-\mu_j\|^2 \\
\iff \quad & \|x-\mu_i\|^2 > \|x-\mu_j\|^2
\end{aligned}
$$

So, a point x is more likely to be a member of class i than of class j if x is nearer, in term of Euclidean distance, to the class i mean, μ_i, than it is to the class j mean, μ_j. More generally, the most likely class for a point x depends simply on whichever class mean is nearest. The input space is thus divided into regions according to the Voronoi diagram based on the class means.

I don't need the extra credit, so here's just a hint of the solutions to those problems:

Extra credit 1: Visually, the data does not support equal covariance matrices—and particularly not of the form $\sigma^2 I$. So, no, I don't think such an assumption is justified.

Extra credit 2: Briefly, something like the Voronoi diagram will describe the regions—they will be separated by straight lines. But the exact location of those lines will depend on the covariance matrix.