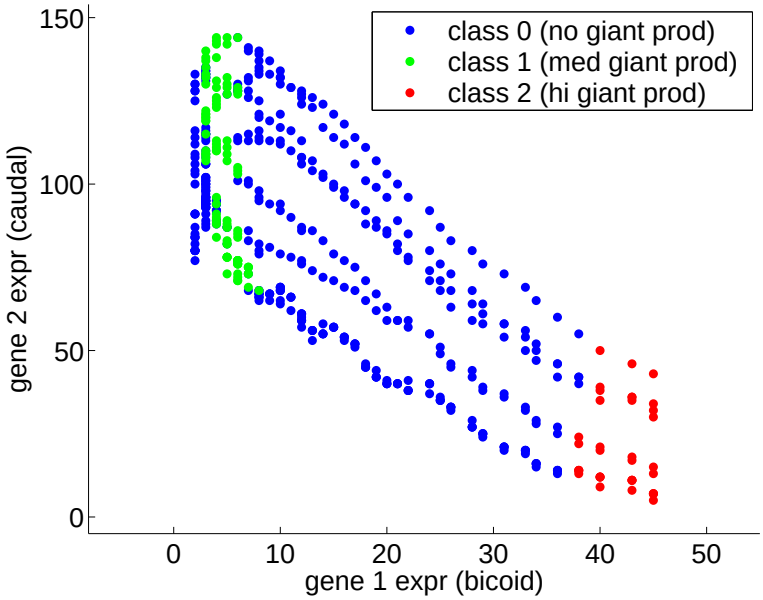
**Homework 3, due Thursday, Oct. 26, 2006**
**COMP 766-001 – Machine Learning for Bioinformatics**

In the files "Homework_03_X.txt" and "Homework_03_Y.txt" you will find a data set of 580 samples related to gene regulation in *Drosophila melanogaster*. In particular, the $X$ matrix has two columns with expression values for two genes (*bicoid* and *caudal*) that are believed to regulate a third gene, *giant*. The $Y$ vector has a discretized transcription rates for the *giant* gene, either 0, 1 or 2, which presumable depend in part on the expression levels of *bicoid* and *caudal*. These therefore comprise a three-class prediction problem, in which we wish to predict the transcription rate of *giant* based on the *bicoid* and *caudal* levels. For your convenience, a plot of the data is below.




**[1] (10pts) Logistic fits:** Fit three logistic regressors to the data: One that attempts to discriminate class 1 from the other two (classes 0 and 2), one that attempts to discriminate class 2 from the other two, and one that attempts to discriminate class 0 from the other two. If your programming language includes a routine for fitting logistic regressors, you may use it. Otherwise, you may use a general nonlinear optimization routine (such as `fminsearch` in Matlab). Or, you may code your own optimization routine, such as gradient descent or Newton's method. The weights should be optimized to minimize the cross-entropy error function, as discussed in class.

**(A)** In this case, of course, the form of a logistic regressor will be $f(x) = 1/(1+\exp(w_0 + w_1 x_1 + w_2 x_2))$. Briefly describe the method you used to fit the weights, and for each of the three fits, report the weights found.

**(B)** Plot the data and overlay on it the decision boundaries—that is, the lines in $x$-space where $f_i(x) = \frac{1}{2}$, where $f_i$ is the logistic regressor that discriminates class $i$ from the other two classes. If it is not clear, indicate in which side of the line $f_i$ is greater than one half.

**(C)** Finally, make a plot showing the most likely class (i.e. which of $f_0(x)$, $f_1(x)$ or $f_2(x)$) is greatest for *bicoid* expression in the range $\{0, 1, 2, \ldots, 50\}$ and *caudal* expression in the range $\{0, 1, 2, \ldots, 150\}$. This could look, for example, like the plot above, except it would have a point for each possible pairing of *bicoid* and *caudal* expression values, with the color of the dot indicating which class is most likely.

**(D)** Comment on your results.

**[2] (10pts) Gaussian discriminant analysis**

**(A)** Compute the maximum likelihood (bivariate) Gaussian fit to the $x$-samples for each of the three classes. Report the mean and covariance matrices for each class.

**(B)** Recall that Bayes's rule says the probability of class $y = i$ given the input values $x$ is:

$$P(y = i|x) = \frac{P(x|y = i)P(y = i)}{\sum_j P(x|y = j)P(y = j)} .$$

Assuming equal prior probabilities, $P(y = 0) = P(y = 1) = P(y = 2) = \frac{1}{3}$, plot the most likely class as a function of *bicoid* and *caudal* expression. (This is the same sort of plot described in problem 1D.)

**(C)** Repeat part B, but use for the prior probabilities, $P(y = i)$, the empirical frequencies of the different classes.

**(D)** Comment on your results.

**[3] (10pts) The shape of three-class decision boundaries** Suppose that we assume the covariance matrices for the three Gaussians are $\Sigma_0 = \Sigma_1 = \Sigma_2 = \sigma^2 I$ where $I$ is the identity matrix and $\sigma > 0$ is a given constant. (Or, we could think of estimating $\sigma$ from the data. It doesn't really matter for the purposes of this problem.) For any class $i$, describe the region of points $x$ for which class $i$ is the most likely class. (That is, describe the region of points for which $P(y = i|x) > P(y = j|x)$ for all $j \neq i$.) You may use formulas or words to "describe the regions", but your answer should be simple or else you're not getting it right!

Extra credit 1: Do you think such an assumption on the covariance matrices is justified for the Drosophila data? Try it, and plot the most likely classes that result (as parts 1D, 2B or 2C).

Extra credit 2: Describe the shape of the most-likely-class regions if $\Sigma_0 = \Sigma_1 = \Sigma_2$, but these are not restricted to any particular form. (No restriction except, of course, that they must be symmetric and positive [semi-]definite, as they are supposed to represent covariance.)