# Linear Discriminant Analysis

Lab Assignment

By: Kenneth Marenco

1. Suppose that $p(x)$ is a function such that $\ln\left\{\dfrac{p(x)}{1-p(x)}\right\} = \beta_0 + \beta_1 x$ for some real numbers $\beta_0, \beta_1$.

   Solve for $p(x)$.

# LDA

Kenneth Marenco

1] $\ln\left\{\frac{p(x)}{1-p(x)}\right\} = \beta_0 + \beta_1 x$

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$$

$$p(x) = e^{\beta_0+\beta_1 x} - p(x)e^{\beta_0+\beta_1 x}$$

$$p(x)\left(1+e^{\beta_0+\beta_1 x}\right) = e^{\beta_0+\beta_1 x}$$

$$\boxed{p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}$$

2. Suppose that $f_1(x)$ and $f_0(x)$ are two one-dimensional Gaussian pdfs, with distinct means $\mu_1$ and $\mu_0$ but with the same variance $\sigma^2$. Use algebraic manipulation to prove the following claim from lecture that the log-odds ratio gives a linear function in $x$:

$$\ln\left\{\frac{f_1(x)\pi_1}{f_0(x)\pi_0}\right\} = \frac{\mu_1-\mu_0}{\sigma^2}x + \left[\ln\pi_1 - \ln\pi_0 - \frac{1}{2\sigma^2}(\mu_1^2-\mu_0^2)\right]$$

2] $$\ln\left\{\frac{f_1(x)\,\pi_1}{f_0(x)\,\pi_0}\right\} = \frac{\mu_1-\mu_0}{\sigma^2}x + \left[\ln\pi_1 - \ln\pi_0 - \frac{1}{2\sigma^2}(\mu_1^2-\mu_0^2)\right]$$

Gaussian : $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

$$\ln\left\{\frac{\frac{\pi_1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2}}{\frac{\pi_0}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu_0}{\sigma}\right)^2}}\right\} = \ln\left\{\frac{\frac{\pi_1}{\sigma\sqrt{2\pi}}}{\frac{\pi_0}{\sigma\sqrt{2\pi}}}\, e^{-\frac{1}{2}\left(\left(\frac{x-\mu_1}{\sigma}\right)^2-\left(\frac{x-\mu_0}{\sigma}\right)^2\right)}\right\}$$

$$= \ln\left\{\frac{\pi_1\sqrt{2\pi}}{\pi_0\sqrt{2\pi}}\right\} + \ln\left\{e^{-\frac{1}{2}\left(\left(\frac{x-\mu_1}{\sigma}\right)^2-\left(\frac{x-\mu_0}{\sigma}\right)^2\right)}\right\}$$

$$= \left[\ln\pi_1 - \ln\pi_0 + \ln\left(\frac{2\pi_0}{2\pi_1}\right)^{\frac{1}{2}}\right] + \frac{-1}{2}\left[\left(\frac{x-\mu_1}{\sigma}\right)^2-\left(\frac{x-\mu_0}{\sigma}\right)^2\right]$$

$$= \left[\ln\pi_1 - \ln\pi_0 + \frac{1}{2}(\ln\pi_0 - \ln\pi_1)\right] + \frac{-1}{2}\left[\frac{x^2-2\mu_1x+\mu_1^2}{\sigma^2} - \frac{x^2-2\mu_0x+\mu_0^2}{\sigma^2}\right]$$

$$= \left[\ln\pi_1 - \ln\pi_0\right] + \frac{-1}{2}\left[\frac{2\mu_0x-2\mu_1x+\mu_1^2-\mu_0^2}{\sigma^2}\right]$$

$$= \left[\ln\pi_1 - \ln\pi_0\right] + \frac{1}{2}\left(\frac{2\mu_1-2\mu_0}{\sigma^2}\right)x + \left(\frac{-\mu_1^2+\mu_0^2}{2\sigma^2}\right)$$

$$= \frac{\mu_1-\mu_0}{\sigma^2}x + \left[\ln\pi_1 - \ln\pi_0 - \frac{1}{2\sigma^2}(\mu_1^2-\mu_0^2)\right] = RHS$$

3. Suppose that $f_1(x)$ and $f_0(x)$ are two one-dimensional Gaussians, with distinct means $\mu_1$ and $\mu_0$ *and* distinct variances $\sigma_1^2$ and $\sigma_0^2$. Show that the log-odds ratio now gives a quadratic function in $x$.

3]

$$\ln\left\{\frac{\frac{\pi_1}{\sigma_1\sqrt{2\pi}}e^{-1/2\left(\frac{x-\mu_1}{\sigma_1}\right)^2}}{\frac{\pi_0}{\sigma_0\sqrt{2\pi}}e^{-1/2\left(\frac{x-\mu_0}{\sigma_0}\right)^2}}\right\} = \ln\left\{\frac{\pi_1\sigma_0\sqrt{2\pi}}{\pi_0\sigma_1\sqrt{2\pi}}\right\} + \frac{-1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_0}{\sigma_0}\right)^2\right]$$

$$= \left[\ln \pi_1\sigma_0 - \ln \pi_0\sigma_1\right] + \frac{1}{2}\left[\frac{x^2-2\mu_1x+\mu_1^2}{\sigma_1^2} - \frac{x^2-2\mu_0x+\mu_0^2}{\sigma_0^2}\right]$$

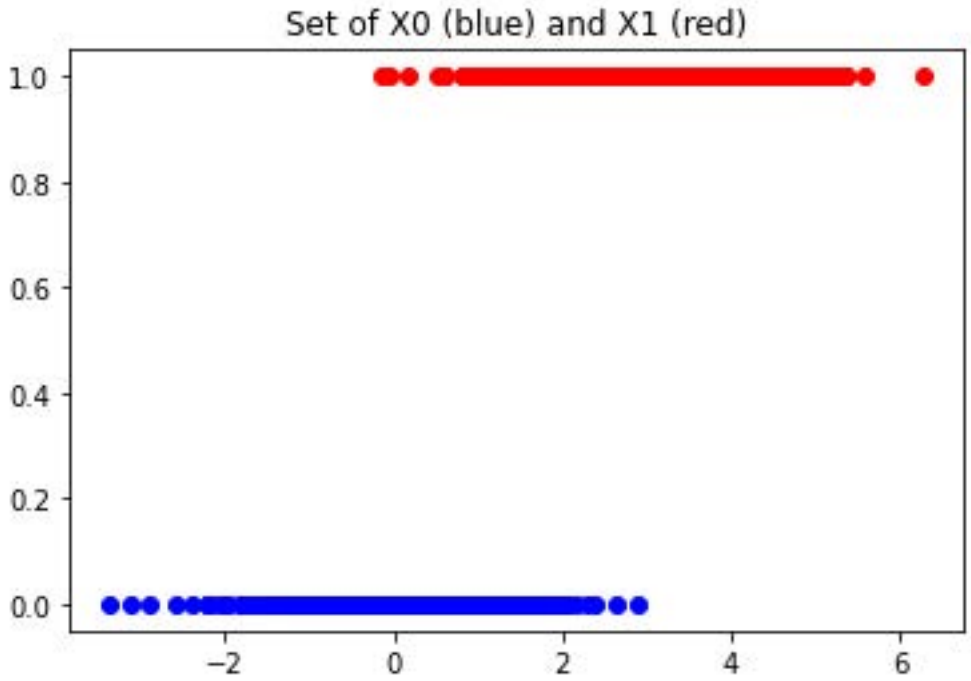
$$= \left[\ln \pi_1\sigma_0 - \ln \pi_0\sigma_1\right] + \frac{-1}{2}\left[\frac{\sigma_0^2(x^2-2\mu_1x+\mu_1^2) - \sigma_1^2(x^2-2\mu_0x+\mu_0^2)}{\sigma_1^2\sigma_0^2}\right]$$

$$= \left[\ln \pi_1\sigma_0 - \ln \pi_0\sigma_1\right] + \frac{-(\sigma_0^2-\sigma_1^2)}{2\sigma_1^2\sigma_0^2}x^2 - \frac{(-2\sigma_0^2\mu_1+2\sigma_1^2\mu_0)}{2\sigma_1^2\sigma_0^2}x - \left(\frac{\sigma_0^2\mu_1^2-\sigma_1^2\mu_0^2}{2\sigma_1^2\sigma_0^2}\right)$$

$$= \frac{\sigma_1^2-\sigma_0^2}{2\sigma_1^2\sigma_0^2}x^2 + \frac{\sigma_0^2\mu_1-\sigma_1^2\mu_0}{\sigma_1^2\sigma_0^2}x + \left[\ln \pi_1\sigma_0 - \ln \pi_0\sigma_1 - \frac{\sigma_0^2\mu_1^2-\sigma_1^2\mu_0^2}{2\sigma_1^2\sigma_0^2}\right]$$

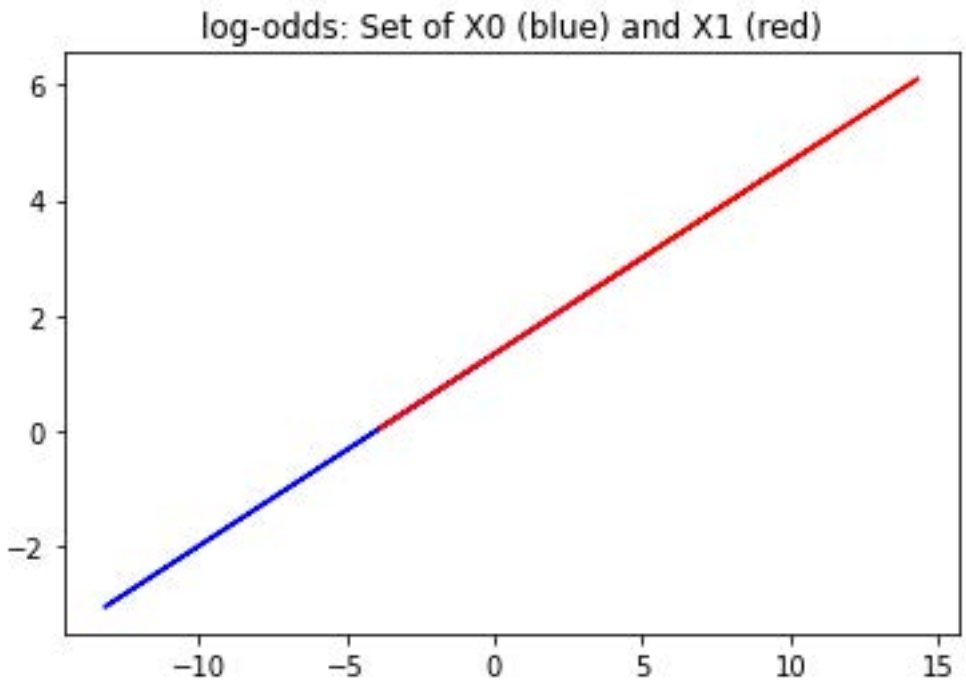
4. Create two sets of real numbers. Let $X_0$ consist of 500 real numbers drawn from a normal distribution with mean 0 and variance 1, and let $X_1$ consist of 500 real numbers drawn from a normal distribution with mean 3 and variance 1. Points in $X_0$ correspond to the classification (target) variable $G = 0$ and in $X_1$ to $G = 1$.

   (a) Plot the two sets of real numbers, using both color and symbol to distinguish them.[1]




[1] Recall it is good practice to attempt choose colors which may be more distinguishable by individuals who are colorblind.

(b) Perform linear discriminant analysis (LDA) on this dataset (note: this requires estimating the various parameters from your dataset), and plot both the log-odds as a function of $x$ and the probability of being in class 1 as a function of $x$.




(c) Use LDA as a classification method on this dataset, using $\frac{4}{5}$ of the data for training and the rest for testing. Output the confusion matrices for a few decision thresholds, and also output the ROC curve.

Attempts in the code file and output

(d) Use LDA as a classification method on this dataset, now using 5-fold cross-validation.

Attempted in the code file

(e) Repeat the experiment of running LDA, using 5-fold cross-validation, for a few different choices of means and variances (usually keep the two variances the same, but at least once choose two different variances). Also experiment with what happens when the two classes are not equally balanced in your dataset (be sure to account for that in your parameter estimation steps for LDA!)

5. This exercise concerns the 'Auto Data'[2] (again!), with the goal of determining which combinations of numerical variables (displacement, horsepower, weight, acceleration) best predict whether the car is 'old' or 'new'. For the purposes of this exercise, 'old' will mean model year 1974 or earlier, and 'new' will mean model year 1975 or later. [3]

   (a) Plot the displacement values with colors (and symbols) to indicate 'new' or 'old'.[4]

[2]Data sourced from `http://www-bcf.usc.edu/~gareth/ISL/data.html`.

[3]The first Clean Air Act was passed in 1970, and required a 90% reduction in emissions from new vehicles by 1975. `https://www.epa.gov/transportation-air-pollution-and-climate-change/timeline-major-accomplishments-transportation-air`

[4]It is good practice to attempt choose colors which may be more distinguishable by individuals who are colorblind.

(b) Run LDA on this feature, with 5-fold cross-validation, and provide your ROC curve.

(c) Repeat the above steps for all other possible choices of predictor variable.