Contents

1 Gaussian Discriminant Analaysis

When using this method, we assume each class comes from its own normal distribution. In other terms, each class C has mean μ_c and variance σ_c^2. Recall the PDF for normally distributed random variables:

$$X \sim N(\mu, \sigma^2) : P(X) = \frac{1}{(\sqrt{2\pi}\sigma)^d} e^{-\frac{\|x-\mu\|^2}{2\sigma^2}}$$

Given a particular point x, the Bayes decision rule $r^*(x)$ will return the class C that maximizes the posterior probability $P(X = X|Y = C)\pi_c$, where π_c is the prior probability of drawing a sample from class c. This Bayes decision rule is not special for GDA, except that we model the posterior probability using a Gaussian.

In order to optimize the posterior probability, we want to take its derivative and find the critical points. In order to make this derivative easier, we take the logorithm of the likelihood function before attempting to derive it. Note that $ln(\omega)$ is monotomically increasing for $\omega > 0$, so maximizing $ln(\omega)$ is equivalent to maximizing ω. To further our goal of taking an easy derivative, let's also rewrite the PDF and prior in a slightly different form.

$$\begin{aligned} Q_C(x) =& ln((\sqrt{2\pi})^d P(X = x)\pi_C) \\ =& -\frac{\|x - \mu_c\|^2}{2\sigma_C^2} - dln(\sigma_C^2) + ln(\pi_C) \end{aligned}$$

Note that $P(X = x)$ in the above expression is shorthand for $P(X = x|Y = C)$, since we are talking about $Q_C(x)$. If we were to write the equation for $Q_D(x)$, it would mean $P(X = x|Y = D)$.

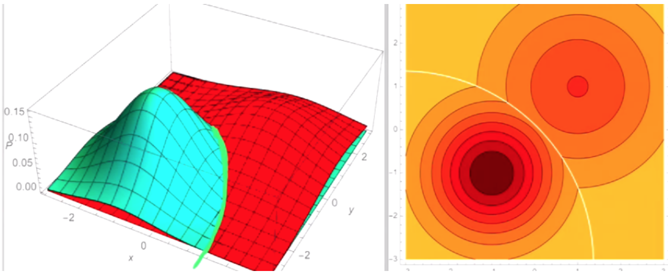
Our new function $Q_c(x)$ is useful, because it is quadratic in x (hence the name). Now that we have laid some groundwork, we can examine two methods of GDA. The first is called Quadratic Discriminant Analysis (QDA), and the second is called Linear Discriminant Analysis (LDA).

1.1 Quadratic Discriminant Analysis

Suppose there are only 2 classes C and D. The Bayes Optimal Decision Rule (BODR) $r^*(x)$ will be:

$$r^*(x) = \begin{cases} C & \text{if } Q_C(x) - Q_D(x) > 0 \\ D & \text{else} \end{cases}$$

We can think of $Q_C(x) - Q_D(x) > 0$ as the decision function, which is quadratic in x. The Bayes Optimal Decision Boundary (BODB) is the set of points where $\{x \mid Q_C(x) - Q_D(x) = 0\}$. In a single dimension, the BODB may have 0, 1, or 2 points. In d dimensions, the BODB will be a quadric (the higher-dimensional version of a conic section).



Sometimes, in addition to getting a prediction for the class of x, we may also want a probability that our prediction is correct.

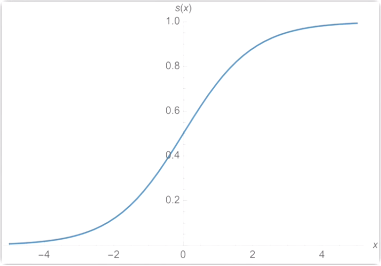
$$P(Y = C|X) = \frac{P(X|Y = C)\pi_C}{P(X|Y = C)\pi_C + P(X|Y = D)\pi_D}$$

This calculation is no different from how we made our class prediction in the first place, because we had to calculate the probability of each class in order to choose the maximum probability. Right at the decision boundary, the probability of correctness will be at a minimum of $\frac{1}{2}$ (or for n intersecting regions, $\frac{1}{n}$). Suppose we want to simplify the above expression. Recall that $e^{Q_C(x)} = ((\sqrt{2\pi})^d P(x)\pi_C$. We could rewrite the above expression as:

$$\begin{aligned} P(Y = C|X) =& \frac{e^{Q_c(x)}}{e^{Q_c(x)} + e^{Q_d(x)}} \\ =& \frac{1}{1 + e^{Q_D(x) - Q_C(x)}} \\ =& \frac{1}{1 + e^{-\gamma}} & \text{where } \gamma = Q_C(x) - Q_D(x) \end{aligned}$$

In the above expression, the result is just some scalar function applied to the BODR. Actually, it happens to be a special and famous function called the logistic function or sigmoid function.

$$s(\gamma) = \frac{1}{1 + e^{-\gamma}}$$



In the image above, the horizontal axis is γ and the vertical axis is $s(\gamma)$. Essentially, the logistic function "squishes" the number line between zero and one. This behavior is useful for modeling probabilities, which fall within that range. Note that $s(0)$ is $\frac{1}{2}$, which is consistant with our interpretation of the probability of C and D both being $\frac{1}{2}$ at the decision boundary $\gamma = 0$. Also note that s is monotomically increasing. Therefore larger values of γ imply higher likelihood of being in class C.

1.2 Linear Discriminant Analysis

This method is very similar to QDA, but the decision boundaries generated by LDA will always be *linear*. This method is primarily useful to avoid overfitting that might occur in QDA. We make an extra assumption in LDA: all of the Gaussians have identical variance σ^2. This extra assumption greatly simplifies our calculation. Now each class is fully identified by its mean μ_C

and its prior π_C. If the priors are uniform, then the optimal decision rule simply chooses the class whose mean is closest to x.

$$
\begin{aligned}
Q_C(x) - Q_D(x) =& \frac{(\mu_C - \mu_D) \cdot x}{\sigma^2} - \frac{|\mu_C|^2 - |\mu_D|^2}{2\sigma^2} + ln(\pi_C) - ln(\pi_D) \\
=& w \cdot x + \alpha
\end{aligned}
$$

Where $w = \frac{\mu_C - \mu_d}{\sigma^2}$, and $\alpha = -\frac{|\mu_C|^2 - |\mu_D|^2}{2\sigma^2} + ln(\pi_C) - ln(\pi_D)$. Since we were able to write the decision rule in the standard form $w \cdot x + \alpha$, we have found a linear classifier.

Instead of looking at the sign of $Q_C(x) - Q_D(x)$, we can express the exact same linear classifier in a different way. We can choose the class C which maximizes the linear discriminant function (LDF).

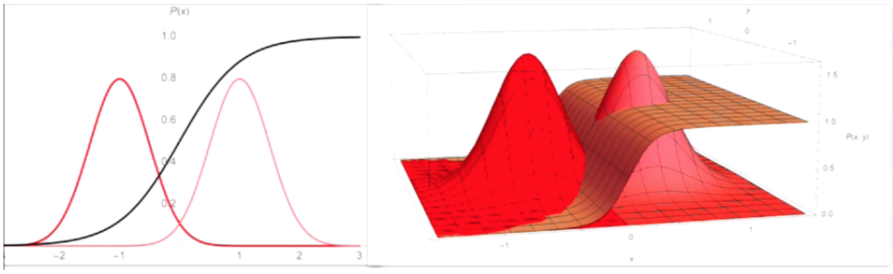
$$LDF_C(x) = \frac{\mu_C \cdot x}{\sigma^2} - \frac{|\mu_C|^2}{2\sigma^2} + ln(\pi_C)$$

If we subtract the analogous expression for class D from the one above, we get an expression that is equivalent to $Q_C(x) - Q_D(x)$. This slight variation on checking the sign of $Q_C(x) - Q_D(x)$ is useful, because we can easily generalize to several classes. We simply compute the LDF for each class, and our decision rule chooses the largest one for each x.

$$r^*(x) = \text{class } C \text{ which maximizes } LDF_C(x)$$

This rule is equivalent to computing all of the pairwise combinations of $Q_C(x) - Q_D(x)$ for each pair of classes, then running a simple algorithm to find the most likely class. Computing a single value for each class and taking the maximum is obviously a more efficient approach.

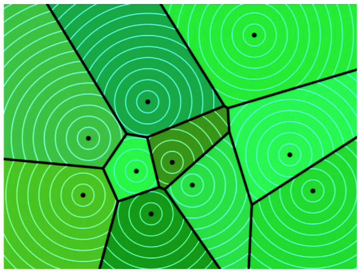
In the 2-class case, the decision boundary is $w \cdot x + \alpha = 0$ as always. Moreover, the Bayes posterior $P(Y = C|X = x)$ is equal to $s(w_C \cdot x + \alpha_D)$.



The curve in black (on the left) is the posterior probability, which is calculated by $\frac{\text{light curve}}{\text{light curve+dark curve}}$. Note that the posterior probability is a linearly transformed logistic function.

The curve in orange (on the right) is the posterior probability, calculated similarly to the previous case. Note that the posterior probability in this case is still a one-dimensional, linearly transformed logistic function. It has simply been extended through another dimension. In a higher-dimensional space, the posterior probability would behave the same way: extending "flatly" through each new dimension, remaining linear.

Below is an example of a multi-class LDA classifier, in the special case where each class has a uniform prior. As shown in the image below, the closest mean defines each region of classification. This classifier turns out to be equivalent to the centroid method, which seems naive yet works well for data under assumptions of Gaussian distributions, uniform variance, and uniform priors.



2 Maximum Likelihood Estimation