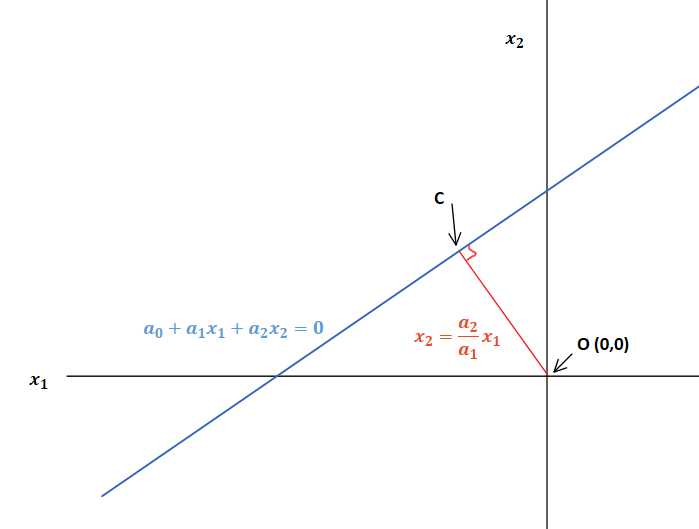
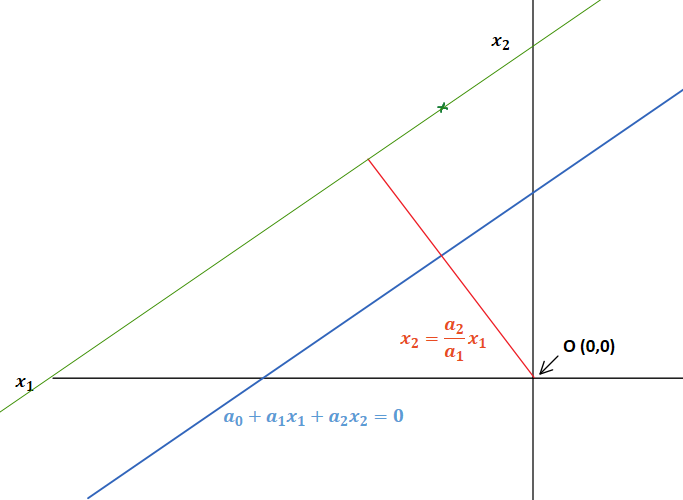
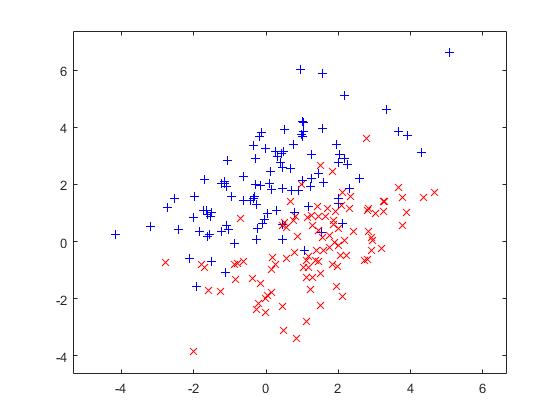
# Question 1.

The general equation for a straight line in two dimensions is . The gradient of this line is . By considering the perpendicular line to this which passes through the origin we can calculate the perpendicular distance from the origin to the straight line. The line has an equation of: (known because of known intercept (0,0) and gradient (perpendicular to ). Then the point where they cross can be calculated by equating these two equations. This gives. The distance to this point from the origin is the perpendicular distance from the line to the origin. Using Pythagoras’ theorem we can show that this distance is 

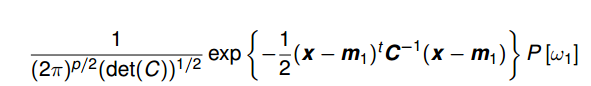
This result can be generalised to show the perpendicular distance from any point (P (m,n)) and a line by considering a line parallel to , so the distance between this line and is equal to the distance from P 🡪 . Calculating the distance between this line and the origin along the same perpendicular line used to calculate the distance allows us to compare the distances directly, giving the general solution: , http://www.intmath.com/plane-analytic-geometry/perpendicular-distance-point-line.php



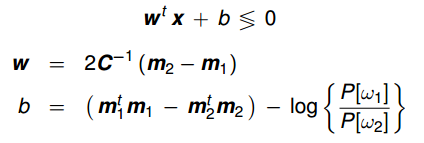
# Question 2.

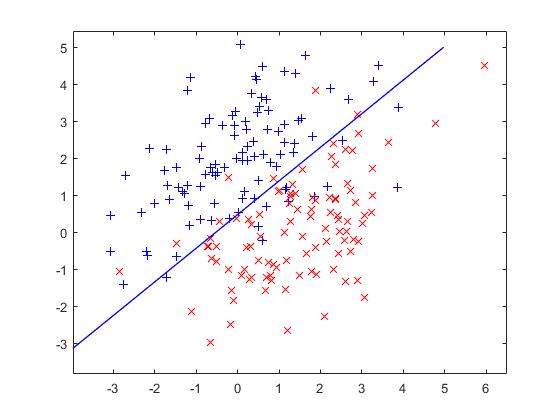
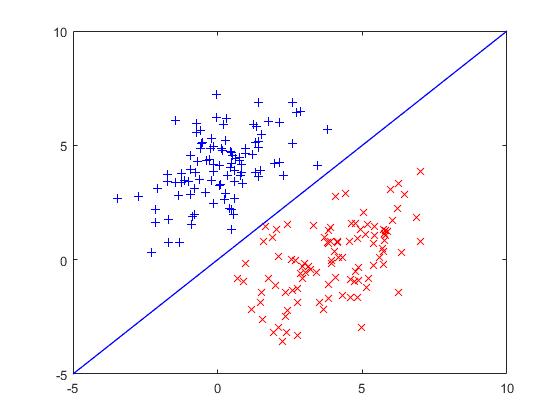
As described in lab one, two bivariate Gaussian distributions were created with means and equal covariance. These distributions are not isotropic, have some overlap, equal covariance and distinct means.

# Question 3.

Given this training data we want to be able to classify any new observations as belonging to distribution 1 or 2. We want to calculate given the features (x1 and x2) of a new observation the probability of that observation being described by each distribution separately. . This is given by Bayes’ theorem . To calculate this we need to know the prevalence [distribution] and the likelihood. The prevalence also known as the prior probability for this case is 0.5. Neither distribution is inherently more likely in the population (same number of both cases). The likelihood is the probability density of each distribution at the point, **.** This can either be calculated explicitly (because we know the population distribution) or estimated using the training set and some assumptions about the population distribution, such as that it is normally distributed.

When comparing two posterior probabilities, in this case we are trying to determine the conditions for one to be more likely than the other, assigning equal weight to either type I and type II errors.

In this case we are trying to evaluate using Bayes’ theorem. The bottom term of the fraction is the same for each distribution so it can be discarded leaving: . Then substituting in our known probability distributions and prevalence and simplifying we get. This is vector notation of a linear classifier. The log term can be ignored because the prevalence is equal (term = 0). Drawing this onto our distribution plot gives:

The classifier seems to classify the generated distributions well with only a handful of points on the wrong side of the line. Of course for these distributions it is impossible to linearly classify them accurately because the distributions overlap. If we separate the means of the distributions the classifier can distinguish between every observation.

This naïve Bayes’ classifier had to make a number of assumptions about the population distribution to accurately work. We assumed the data was normally distributed with equal covariance. This while accurate for our generated data is not often the case for real world data, however approximating the data in this way can often be useful. In practice it is important to inspect your training data to ensure a normal distribution is at least an approximate match and to transform the features of the data if this isn’t the case.

# Question 4.

Finally I implemented a perceptron algorithm to try and algorithmically produce a linear classifier.