Stats 102B - Week 6, Lecture 2

Miles Chen, PhD

Department of Statistics

Week 6 Wednesday

UCLA

Section 1

Classification - Bayes Classifier

Review: The Bayes Classifier

We apply Bayes' rule: $\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$

$$\Pr(T_{new} = c|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}) = \frac{\Pr(\mathbf{x}_{new}|T_{new} = c, \mathbf{X}, \mathbf{t})\Pr(T_{new} = c|\mathbf{X}, \mathbf{t})}{\Pr(\mathbf{x}_{new}|\mathbf{X}, \mathbf{t})}$$

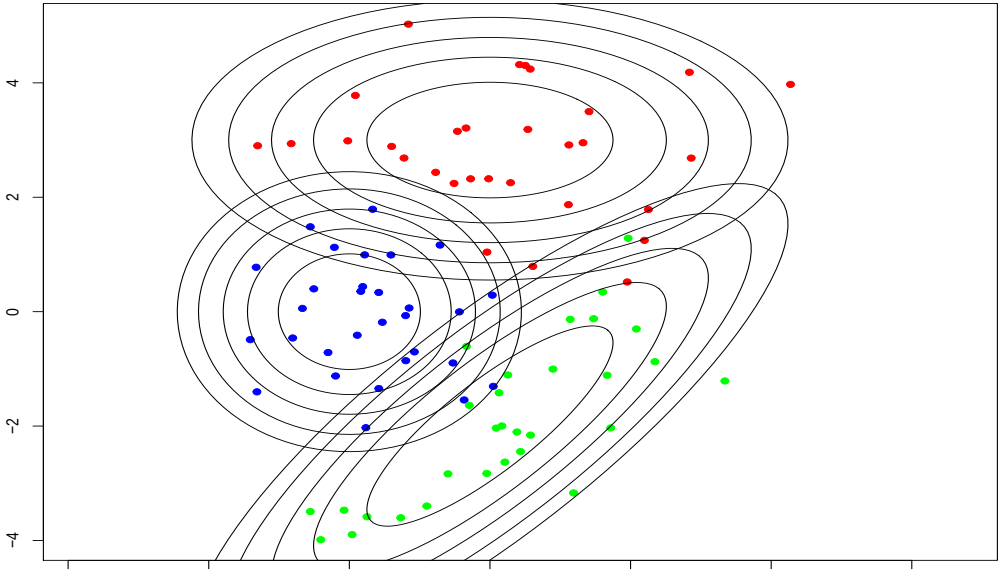
- **Posterior**: $\Pr(T_{new} = c|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t})$ is the probability that we want to find. It is the probability an observation belongs to class c
- **Likelihood**: $\Pr(\mathbf{x}_{new}|T_{new} = c, \mathbf{X}, \mathbf{t})$ is the probability of observing the observed values in $\mathbf{x}_{new}$ if we assumed the new observation belonged to a specific class c.
- **Prior**: $\Pr(T_{new} = c|\mathbf{X}, \mathbf{t})$ is the probability that some new observation belongs to class c before we know anything else about it. It is often equal to the proportion of observations that belong to class c in the observed data.
- **Marginal**: $\Pr(\mathbf{x}_{new}|\mathbf{X}, \mathbf{t})$ is the probability of observing the values in $\mathbf{x}_{new}$ regardless of class label. We find it by summing the numerator values (likelihood $\times$ prior) across all possible classes.

Review: Our Simulated data

We generated data from 3 classes.

```
library(mvtnorm) # to generate multivariate normal data
set.seed(150)

mu_A = c(2, 3)
sigma_A <- matrix(c(3, 0, 0, 1), nrow = 2) # x1 and x2 are independent
Xa <- rmvnorm(30, mu_A, sigma_A)

mu_B = c(0, 0)
sigma_B <- matrix(c(1, 0, 0, 1), nrow = 2) # x1 and x2 are independent
Xb <- rmvnorm(30, mu_B, sigma_B)

mu_C = c(2,-2)
sigma_C <- matrix(c(3, 2.5, 2.5, 3), nrow = 2) # x1 and x2 are not independent
Xc <- rmvnorm(30, mu_C, sigma_C)

X  <- rbind(Xa, Xb, Xc)
```

Code to plot the data with M.V. Normal Contour lines (ellipses)



Estimates of distribution parameters from the training data

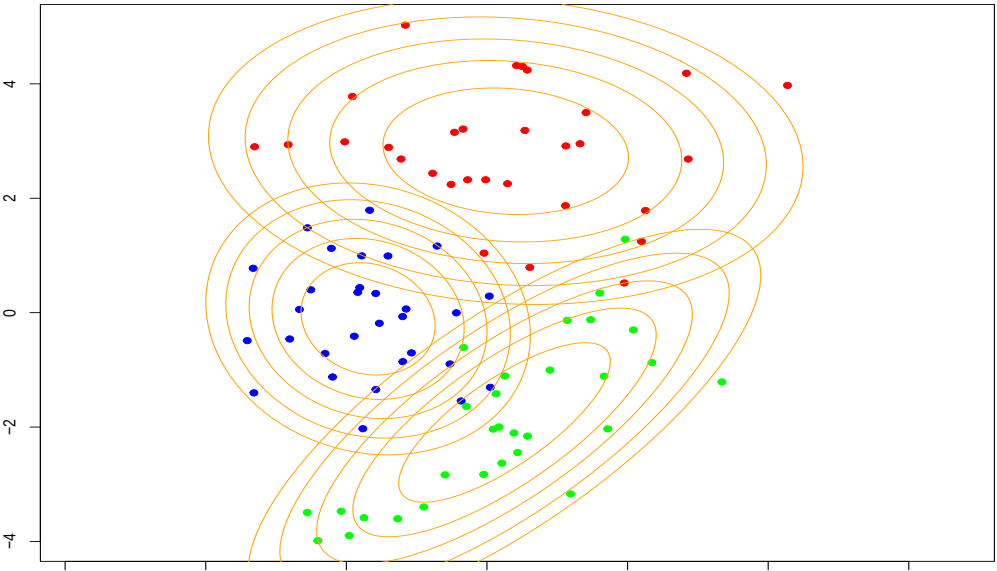
Estimates of the means based on the simulated data

```
xbar_a <- colMeans(Xa)
xbar_b <- colMeans(Xb)
xbar_c <- colMeans(Xc)
```

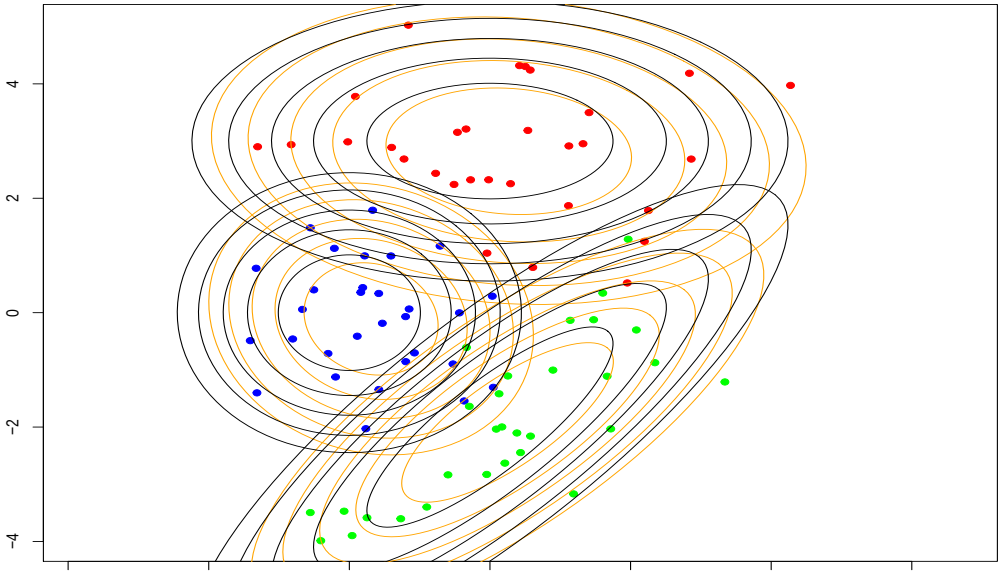
Estimates of the variances based on the simulated data

```
var_a <- var(Xa)
var_b <- var(Xb)
var_c <- var(Xc)
```

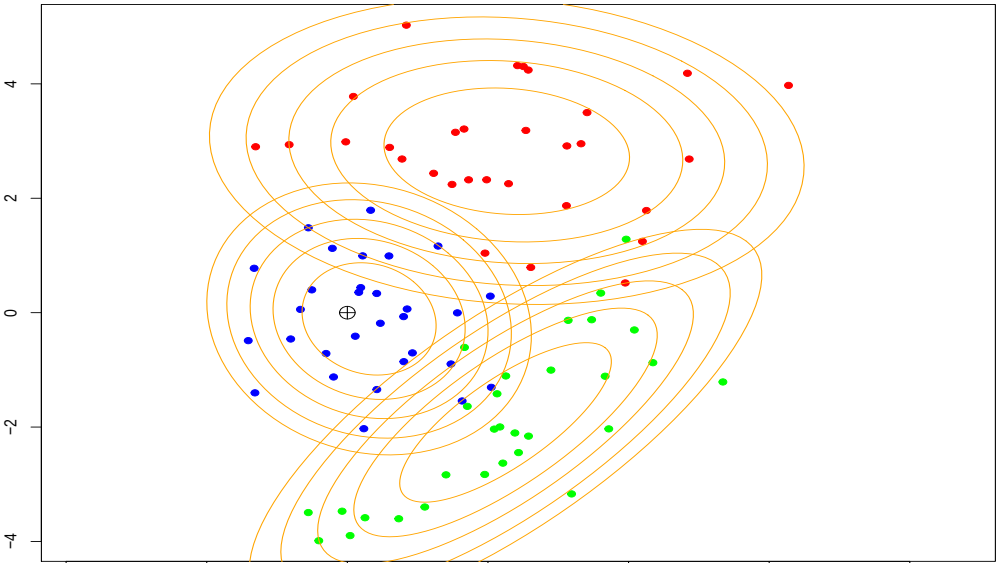
Plot of data with estimated contour lines



Plot with both “known” contours and estimated contours



A Test Case at (0, 0) - Which class does it belong to?



Bayes Classifier:

We will calculate our classification probability for each of the classes.

$$\Pr(T_{new} = c|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}) = \frac{\Pr(\mathbf{x}_{new}|T_{new} = c, \mathbf{X}, \mathbf{t}) \Pr(T_{new} = c|\mathbf{X}, \mathbf{t})}{\Pr(\mathbf{x}_{new}|\mathbf{X}, \mathbf{t})}$$

$$\text{Posterior Probability} = \frac{\text{Likelihood} \times \text{Prior Probability}}{\text{Marginal Probability}}$$

- **Posterior**: The probability our test case belongs to class c
- **Likelihood**: The probability of observing the test-case values $(0, 0)$ if we assumed the new observation belonged to a specific class c. This will be calculated using the multivariate normal density.
- **Prior**: The probability that our test case belongs to class c before we know anything else about it. We will use the proportion of observations that belong to class c in the observed data.
- **Marginal**: The probability of observing the test-case values $(0, 0)$ regardless of class label. We find it by summing the numerator values (likelihood $\times$ prior) across all possible classes.

The Likelihood

The likelihood is the probability of observing the test-case values $(0, 0)$ if we assumed the new observation belonged to a specific class c.

The likelihood will be calculated using the multivariate Normal PDF with the estimates of the mean and variance for each class.

$$\frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

For class A, we use the estimates `xbar_a` for $\boldsymbol{\mu}$ and `var_a` for $\mathbf{\Sigma}$ to calculate our likelihood.

```
X_new <- c(0, 0) # Our new point is at (0,0)
1 / ( (2 * pi) ^ (2 / 2) * det(var_a) ^ 0.5) *
  exp( -0.5 * t(X_new - xbar_a) %*% solve(var_a) %*% (X_new - xbar_a))
```

```
##                [,1]
## [1,] 0.0008795399
```

The Likelihood

The calculations are made simple with with multi-variate normal density function `dmvnorm()`

```
like_a <- dmvnorm(X_new, mean = xbar_a, sigma = var_a); like_a # class A
```

```
## [1] 0.0008795399
```

```
# above result matches previous slide
like_b <- dmvnorm(X_new, mean = xbar_b, sigma = var_b); like_b # class B
```

```
## [1] 0.1654324
```

```
like_c <- dmvnorm(X_new, mean = xbar_c, sigma = var_c); like_c # class C
```

```
## [1] 1.643014e-05
```

The Priors

The Prior probability is the probability that our test case belongs to class c before we know anything else about it. We will use the proportion of observations that belong to class c in the observed data.

In our case, all three classes have the same proportion of observations. (There are 90 points total and each class has 30 points.) Thus all three classes have the same prior probability of $1/3$.

Technically, it's not necessary to include them in the calculations as they will all cancel out, but I'll go ahead and include them to stay consistent with the Bayes' rule formula.

```
prior_a = prior_b = prior_c = 1/3
```

The Marginal Probability

The marginal is found by summing up the numerators (product of likelihood and prior) across all possible classes:

```
marginal <- like_a * prior_a + like_b * prior_b + like_c * prior_c
```

The Posterior Class Probabilities

The posterior probability is then found using Bayes' Rule:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal}}$$

```
like_a * prior_a / marginal
```

```
## [1] 0.005287974
```

```
like_b * prior_b / marginal
```

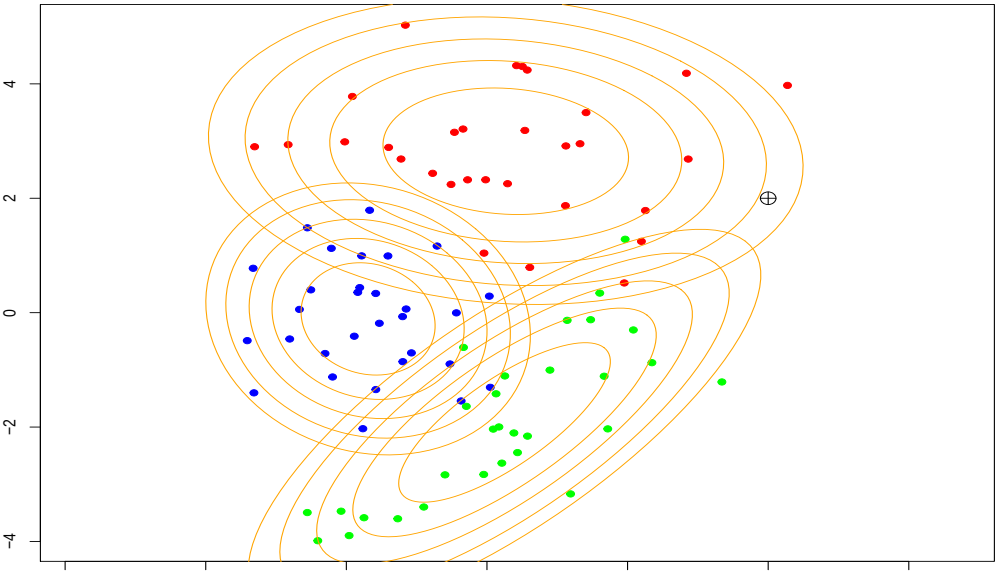
```
## [1] 0.9946132
```

```
like_c * prior_c / marginal
```

```
## [1] 9.878139e-05
```

As expected, the probability that (0,0) belongs to class B (blue dots) is highest.

Another Test Case at (6, 2)



```r
X_new = c(6,2)
like_a <- dmvnorm(X_new, mean = xbar_a, sigma = var_a) # likelihood of class A
like_b <- dmvnorm(X_new, mean = xbar_b, sigma = var_b) # likelihood of class B
like_c <- dmvnorm(X_new, mean = xbar_c, sigma = var_c) # likelihood of class C
prior_a = prior_b = prior_c = 1/3 # all three classes have the same prior
marginal <- like_a * prior_a + like_b * prior_b + like_c * prior_c
like_a * prior_a / marginal # Prob (6,2) belongs to Class A
```

```
## [1] 0.7948236
```

```r
like_b * prior_b / marginal # Prob (6,2) belongs to Class B
```

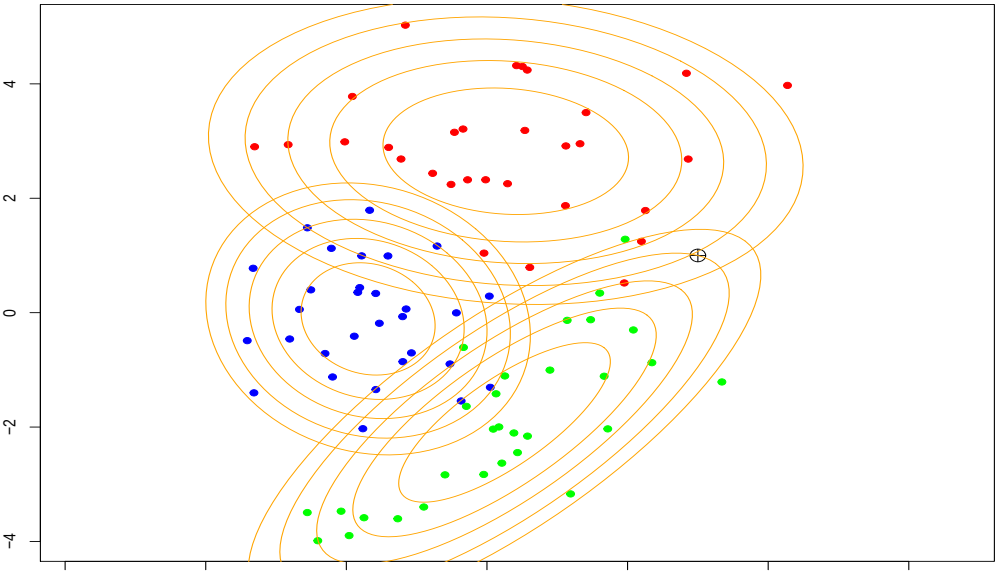
```
## [1] 3.317476e-09
```

```r
like_c * prior_c / marginal # Prob (6,2) belongs to Class C
```

```
## [1] 0.2051764
```

Class A has highest probability.

Another Test Case - (5, 1)



```
X_new <- c(5,1)
like_a <- dmvnorm(X_new, mean = xbar_a, sigma = var_a) # likelihood of class A
like_b <- dmvnorm(X_new, mean = xbar_b, sigma = var_b) # likelihood of class B
like_c <- dmvnorm(X_new, mean = xbar_c, sigma = var_c) # likelihood of class C
prior_a = prior_b = prior_c = 1/3 # all three classes have the same prior
marginal <- like_a * prior_a + like_b * prior_b + like_c * prior_c
like_a * prior_a / marginal # Prob (5,1) belongs to Class A
```

```
## [1] 0.3908285
```

```
like_b * prior_b / marginal # Prob (5,1) belongs to Class B
```

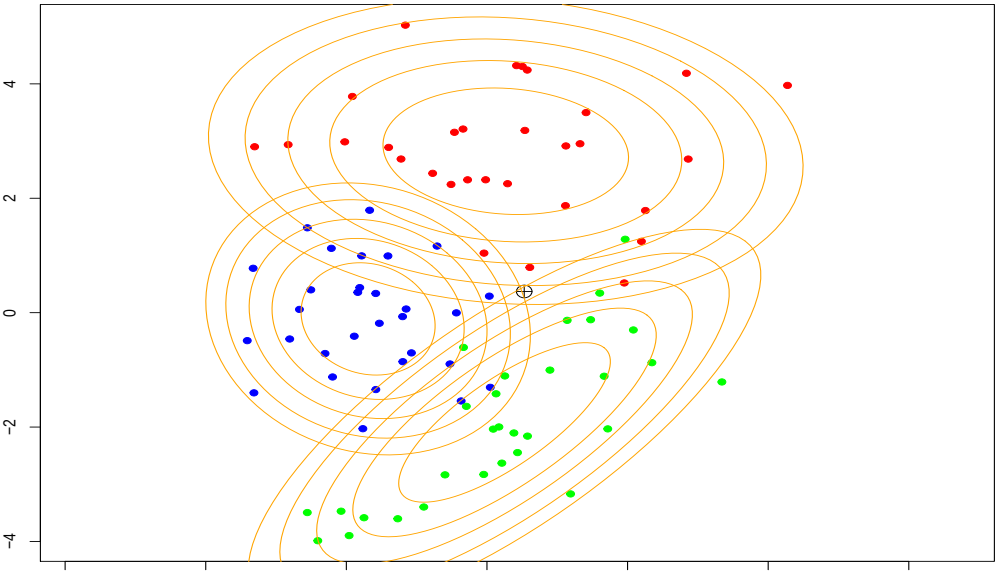
```
## [1] 7.987098e-06
```

```
like_c * prior_c / marginal # Prob (5,1) belongs to Class C
```

```
## [1] 0.6091635
```

Class C has highest probability.

Another Test Case



```r
X_new <- c(2.53, 0.37)
like_a <- dmvnorm(X_new, mean = xbar_a, sigma = var_a) # likelihood of class A
like_b <- dmvnorm(X_new, mean = xbar_b, sigma = var_b) # likelihood of class B
like_c <- dmvnorm(X_new, mean = xbar_c, sigma = var_c) # likelihood of class C
prior_a = prior_b = prior_c = 1/3 # all three classes have the same prior
marginal <- like_a * prior_a + like_b * prior_b + like_c * prior_c
like_a * prior_a / marginal
```

```
## [1] 0.3063477
```

```r
like_b * prior_b / marginal
```

```
## [1] 0.3584183
```

```r
like_c * prior_c / marginal
```

```
## [1] 0.335234
```

Class B has highest probability, but all three classes have similar values.

Summary

The probability that an observation belongs to a particular class can be found using Bayes rule.

$$\text{posterior} = \text{likelihood} \times \text{prior} \div \text{marginal}$$

- **Posterior** is the probability that we want to find. It is the probability an observation belongs to class c
- **Likelihood** is the probability of observing the observed values in $\mathbf{x}_{new}$ if we assumed the new observation belonged to a specific class c.
- **Prior** is the probability that some new observation belongs to class c before we know anything else about it. It is often equal to the proportion of observations that belong to class c in the observed data.
- **Marginal** is the probability of observing the values in $\mathbf{x}_{new}$ regardless of class label. We find it by summing the numerator values (likelihood $\times$ prior) across all possible classes.

Section 2

Naive Bayes Classifier

Review: The Multivariate Gaussian (Normal) Distribution

If $\mathbf{x} = [x_1, x_2, \ldots, x_D]^T$ comes from a multivariate Gaussian distribution in D dimensions, the PDF is:

$$\frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

where $\boldsymbol{\mu}$ is a vector of means of the population (the same size as $\mathbf{x}$), the dth element tells us the mean of x_d. The variance $\mathbf{\Sigma}$ is a $D \times D$ variance-covariance matrix of the x-variables.

If the off-diagonal elements in $\mathbf{\Sigma}$ are 0, then the x-variables are independent and the multivariate PDF can be factored into a product of univariate Gaussian PDFs.

The Naive Assumption

The Naive Assumption in the Naive Bayes classifier is that the x-variables are independent of each other.

This assumption is generally wrong, but can produce a model that is more useful.

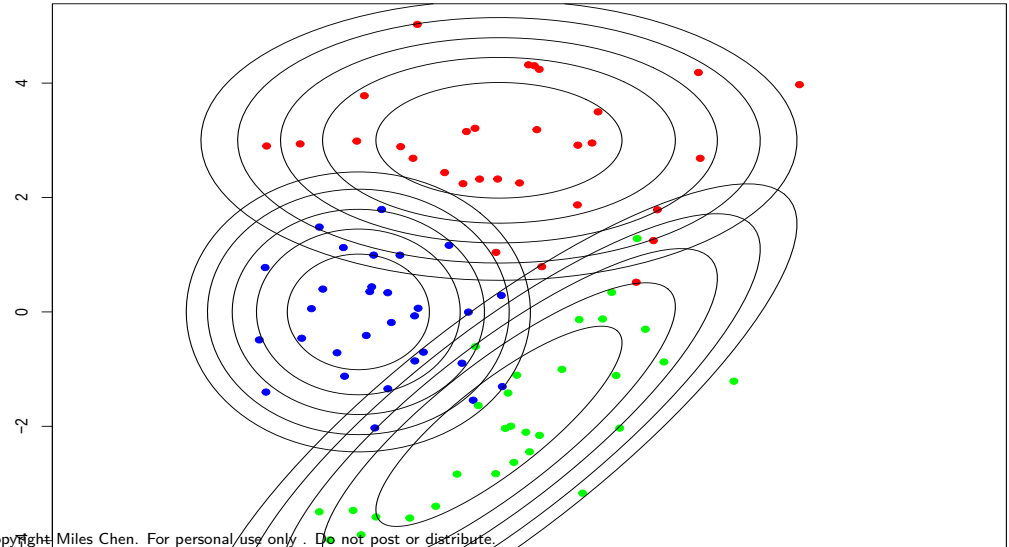
To get trustworthy estimates of the covariances between x-variables we need many observations, especially if there are very many x-variables. On the other hand, it is much easier to estimate the individual variances while we assume that all the covariances are 0.

Thus, for the multivariate normal distribution, we will replace the multivariate density function with a product of univariate normal densities:

$$\frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} = \prod_{d=1}^{D} \frac{1}{\sigma_d\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x_d-\mu_d}{\sigma_d}\right)^2\right\}$$

Our Simulated data

We can use the same simulated data from before.



Estimates of distribution parameters from the training data

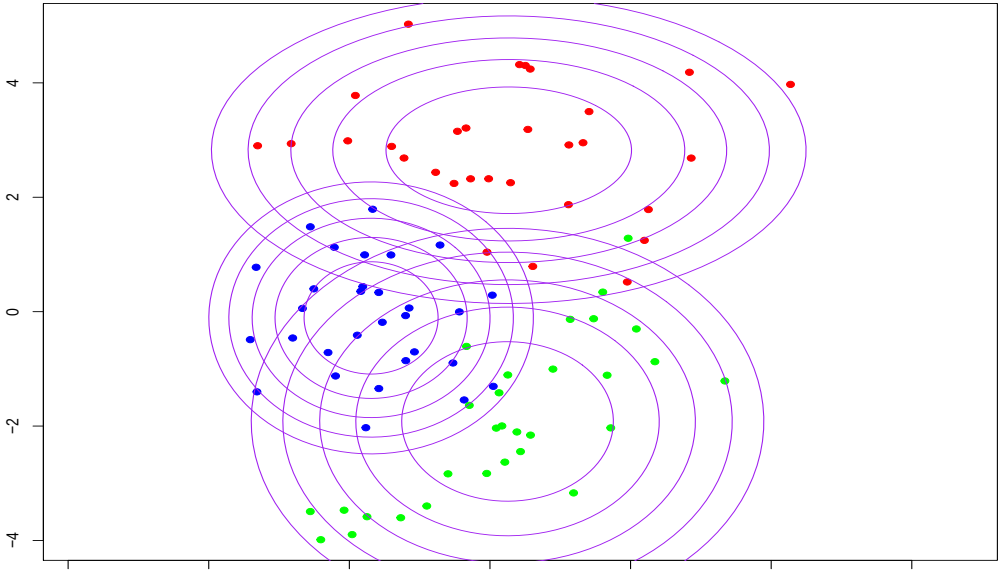
We estimate the means one variable at a time. (Technically, `colMeans` could have worked too.)

```
x1bar_a <- mean(Xa[,1])
x1bar_b <- mean(Xb[,1])
x1bar_c <- mean(Xc[,1])
x2bar_a <- mean(Xa[,2])
x2bar_b <- mean(Xb[,2])
x2bar_c <- mean(Xc[,2])
```

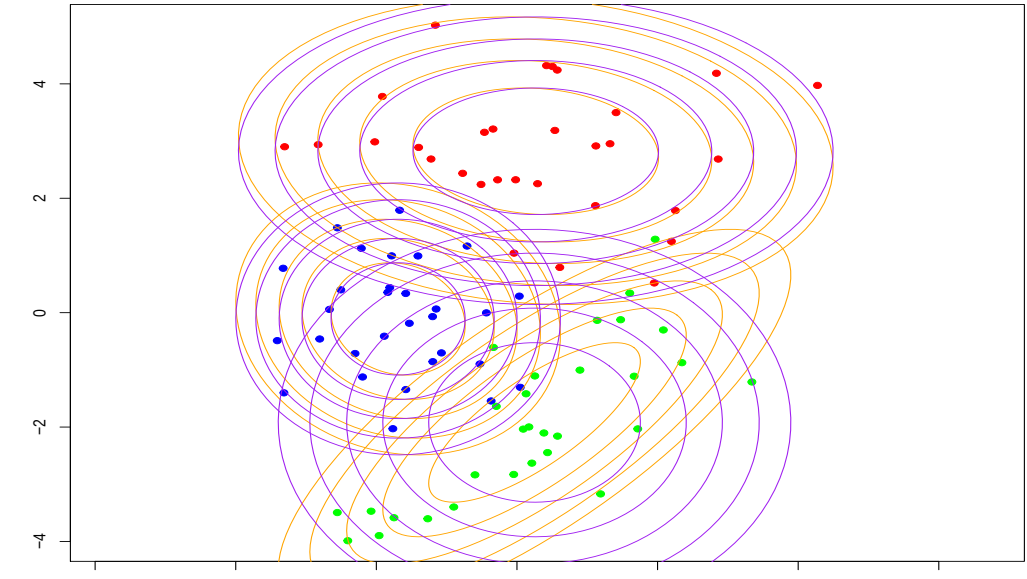
Similarly, we estimate the variances (and sd) one variable at a time.

```
var1_a <- var(Xa[,1]); sd1_a <- sd(Xa[,1])
var1_b <- var(Xb[,1]); sd1_b <- sd(Xb[,1])
var1_c <- var(Xc[,1]); sd1_c <- sd(Xc[,1])
var2_a <- var(Xa[,2]); sd2_a <- sd(Xa[,2])
var2_b <- var(Xb[,2]); sd2_b <- sd(Xb[,2])
var2_c <- var(Xc[,2]); sd2_c <- sd(Xc[,2])
```

Plot of data with estimated contour lines under naive assumption



Estimated contour lines: naive (purple), multivariate (orange)

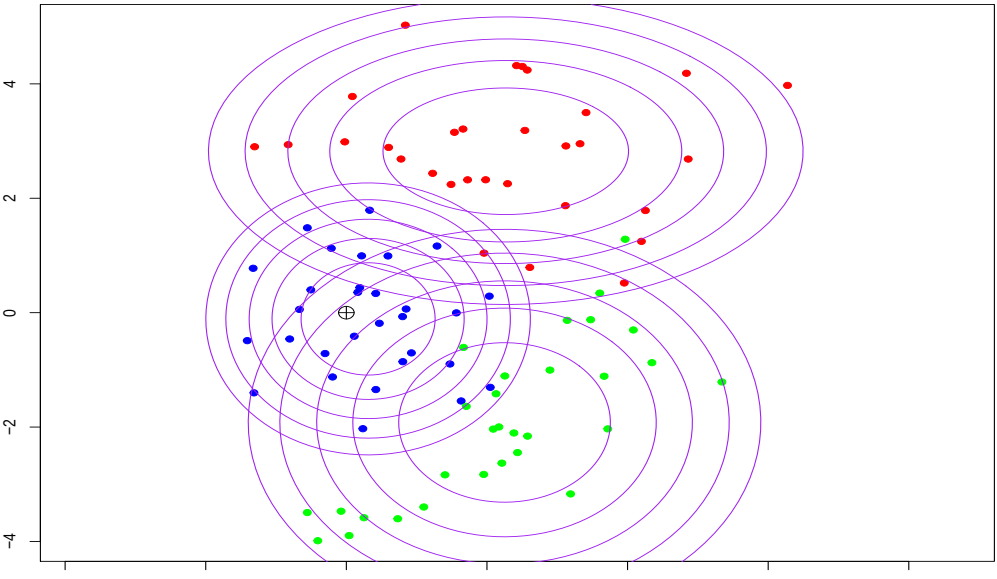


Comparison

The difference is most pronounced for the green dots, which were generated from a distribution with high covariance.

With the naive assumption, there is no covariance, so the major and minor axes of the distribution ellipses are parallel to the plot axes.

Test Case at (0,0)



Naive Bayes Classifier:

The calculations for the naive Bayes Classifier is pretty much the same as the Bayes Classifier. The only difference is for the likelihood we replace a Multivariate Distribution with the product of univariate distributions.

$$\Pr(T_{new} = c|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}) = \frac{\Pr(\mathbf{x}_{new}|T_{new} = c, \mathbf{X}, \mathbf{t}) \Pr(T_{new} = c|\mathbf{X}, \mathbf{t})}{\Pr(\mathbf{x}_{new}|\mathbf{X}, \mathbf{t})}$$

$$\text{Posterior Probability} = \frac{\text{Likelihood} \times \text{Prior Probability}}{\text{Marginal Probability}}$$

- Posterior: The probability our test case belongs to class c
- **Likelihood**: The probability of observing the test-case values $(0, 0)$ if we assumed the new observation belonged to a specific class c. **This will be calculated using the a product of independent normal density functions.**
- Prior: The probability that our test case belongs to class c before we know anything else about it. We will use the proportion of observations that belong to class c in the observed data.
- Marginal: The probability of observing the test-case values $(0, 0)$ regardless of class label. We find it by summing the numerator values (likelihood $\times$ prior) across all possible classes.

The Likelihood

The likelihood is the probability of observing the test-case values $(0, 0)$ if we assumed the new observation belonged to a specific class c.

The likelihood will be calculated using **a product of Normal PDFs** with the estimates of the mean and variance for each variable and each class.

```
like_a <- dnorm(X_new[1], mean = x1bar_a, sd = sd1_a) *
  dnorm(X_new[2], mean = x2bar_a, sd = sd2_a); like_a # class A
```

```
## [1] 0.00128171
```

```
like_b <- dnorm(X_new[1], mean = x1bar_b, sd = sd1_b) *
  dnorm(X_new[2], mean = x2bar_b, sd = sd2_b); like_b # class B
```

```
## [1] 0.1635916
```

```
like_c <- dnorm(X_new[1], mean = x1bar_b, sd = sd1_c) *
  dnorm(X_new[2], mean = x2bar_c, sd = sd2_c); like_c # class C
```

```
## [1] 0.02879979
```

The Prior and Marginal Probabilities

The Prior probability is the probability that our test case belongs to class c before we know anything else about it. We will use the proportion of observations that belong to class c in the observed data. All three classes have the same prior probability of $1/3$.

```
prior_a = prior_b = prior_c = 1/3
```

The marginal is found by summing up the numerators (product of likelihood and prior) across all possible classes:

```
marginal <- like_a * prior_a + like_b * prior_b + like_c * prior_c
```

The Posterior Class Probabilities

The posterior probability is then found using Bayes' Rule:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal}}$$

```
like_a * prior_a / marginal
```

```
## [1] 0.006617905
```

```
like_b * prior_b / marginal
```

```
## [1] 0.844679
```

```
like_c * prior_c / marginal
```

```
## [1] 0.1487031
```

As expected, the probability that (0,0) belongs to class B (blue dots) is highest, but the difference isn't as extreme as it was when we used the multivariate normal distribution.