Generative models:
With multiple features we need multivariate distributions for the class conditionals

Multivariate data

If $X_1 \ldots, X_p$ are real-valued random variables, we can talk about a p-dimensional random vector $X = (X_1, \ldots, X_p)$.

The usual convention is that random vectors are column vectors, i.e.

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$$

Expectation

The expectation, or mean, of a p-dimensional random vector X is defined as

$$\mathbb{E}\,X = \begin{bmatrix} \mathbb{E}\,X_1 \\ \vdots \\ \mathbb{E}\,X_p \end{bmatrix}$$

The sample mean for vector X is the average

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

which is a vector of coordinate-wise averages (sorry about notation).

Variance

The variance of a p-dimensional random vector X is defined as

$$\begin{aligned}
\operatorname{Var} X &= \mathbb{E}\left[(X - \mathbb{E}\,X)(X - \mathbb{E}\,X)^T\right] \\
&= \begin{bmatrix}
\operatorname{Var} X_1 & \operatorname{Cov}(X_1, X_2) & \dots & \operatorname{Cov}(X_1, X_p) \\
\operatorname{Cov}(X_2, X_1) & \operatorname{Var} X_2 & \dots & \operatorname{Cov}(X_2, X_p) \\
\vdots & \vdots & \ddots & \vdots \\
\operatorname{Cov}(X_p, X_1) & \operatorname{Cov}(X_p, X_2) & \dots & \operatorname{Var} X_p
\end{bmatrix}
\end{aligned}$$

This is called the *covariance matrix* or *variance matrix* for X.

The Multivariate Normal Distribution

The Multivariate Normal Distribution

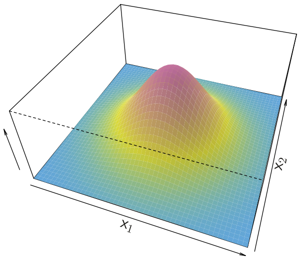


FIGURE 4.5. *Two multivariate Gaussian density functions are shown, with* $p = 2$. Left: *The two predictors are uncorrelated.* Right: *The two variables have a correlation of* 0.7.

The Multivariate Normal Distribution

The p-dimensional multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and variance Σ has probability density function

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}$$

The Multivariate Normal Distribution

The p-dimensional multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and variance Σ has probability density function

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

The exponent contains the (squared) *Mahalanobis distance* based on Σ:

$$(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) = \|\boldsymbol{x}-\boldsymbol{\mu}\|_{\Sigma}^2$$

The Multivariate Normal Distribution

The p-dimensional multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and variance Σ has probability density function

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}$$

The exponent contains the (squared) *Mahalanobis distance* based on Σ:

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = \|\boldsymbol{x} - \boldsymbol{\mu}\|_{\Sigma}^2$$

When $\Sigma = I$,

The Multivariate Normal Distribution

The p-dimensional multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and variance Σ has probability density function

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}$$

The exponent contains the (squared) *Mahalanobis distance* based on Σ:

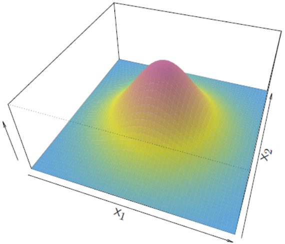
$$(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = \|\boldsymbol{x} - \boldsymbol{\mu}\|_\Sigma^2$$

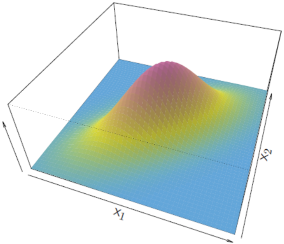
When $\Sigma = I$, all variables are independent and standard normal, and

$$\|\boldsymbol{x} - \boldsymbol{\mu}\|_I^2 = \sum_{i=1}^{p} (x_i - \mu_{ik})^2$$

is the squared geometric (Euclidean) distance from x to $\boldsymbol{\mu}$.

The Multivariate Normal Distribution



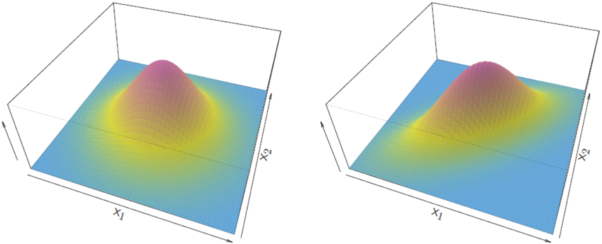


Looking only at factors involving $\boldsymbol{x}$ the pdf is

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}$$

$$\propto e^{-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}\|_{\Sigma}^2}$$

The Multivariate Normal Distribution



Looking only at factors involving $\boldsymbol{x}$ the pdf is

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right\}$$
$$\propto e^{-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}\|_\Sigma^2}$$

A contour curve for the pdf

- shows all points with a given density value.
- is a quadratic form (thus elliptic in 2D).
- shows all points of the same distance to the mean.
- does not change shape when we transform the pdf (e.g. take log)

Any linear transformation or translation of a Gaussian variable is also Gaussian

If X is p-dimensional multivariate normal, $\mathrm{MVN}(\boldsymbol{\mu}, \Sigma)$, then

$$AX + \boldsymbol{b} = \mathrm{MVN}(A\boldsymbol{\mu} + \boldsymbol{b}, A\Sigma A^T)$$

where A is any $q \times p$-matrix.

Consequence: **All of the marginal distributions are Gaussian.**

(Be aware that a transformation can result in a singular covariance matrix.)

LDA and QDA with multiple features

Discriminant analysis

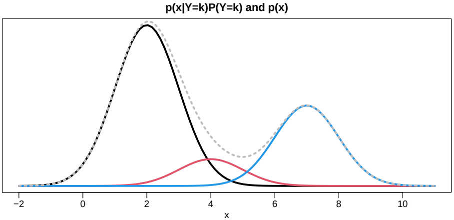
Remember…

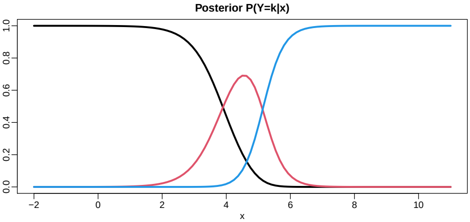
LDA is a generative model where the class conditionals $p(x \mid Y = k)$ are assumed Gaussian with individual class means, but ***equal covariance matrices***.

QDA is also a generative model, but there the class conditionals $p(x \mid Y = k)$ are assumed Gaussian with individual class means, and ***class-specific covariance matrices***.

Let us first consider the LDA classifier.

Example: LDA





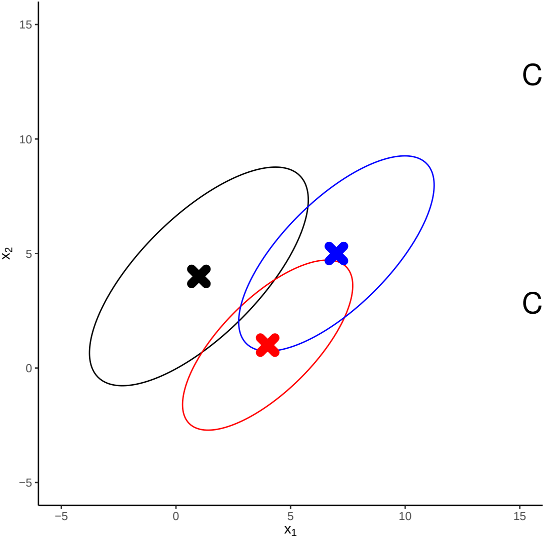
In 1 dimension: Conditionally on the class we assumed the feature x to have a univariate Gaussian distribution as

$$p(x \mid Y = \text{black}) = \mathcal{N}(2, 1)$$
$$p(x \mid Y = \text{red}) = \mathcal{N}(4, 1)$$
$$p(x \mid Y = \text{blue}) = \mathcal{N}(7, 1)$$

The class probabilities we took to be

$$\pi_{\text{black}} = 0.6$$
$$\pi_{\text{red}} = 0.1$$
$$\pi_{\text{blue}} = 0.3$$

Example: LDA



Plot of $p(x, y)$ contours.

Class conditionals, $p(x \mid y)$:

$$
\begin{aligned}
p(x \mid \text{black}) &= \text{MVN}\left(\begin{bmatrix}2\\3\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)\\
p(x \mid \text{red}) &= \text{MVN}\left(\begin{bmatrix}4\\1\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)\\
p(x \mid \text{blue}) &= \text{MVN}\left(\begin{bmatrix}7\\5\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)
\end{aligned}
$$

Class probabilities:

$$
\begin{aligned}
\pi_{\text{black}} &= 0.9\\
\pi_{\text{red}} &= 0.01\\
\pi_{\text{blue}} &= 0.09
\end{aligned}
$$

Discriminants for LDA

Discriminant functions based on the joint distribution are

$$
\begin{aligned}
p(Y = k)p(x \mid Y = k) &= \pi_k \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)} \\
&\propto \pi_k e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)}
\end{aligned}
$$

(suppressing factors that do not depend on k)

Discriminants for LDA

Discriminant functions based on the joint distribution are

$$p(Y = k)p(x \mid Y = k) = \pi_k \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)}$$
$$\propto \pi_k e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)}$$

(suppressing factors that do not depend on k)

Taking logs and multiplying by 2 gives a simpler expression

$$g_k(\boldsymbol{x}) = 2 \log \pi_k - (\boldsymbol{x} - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)$$

Q: What happens to $g_k(\boldsymbol{x})$ when π_k increases? $(\boldsymbol{x} - \underline{\boldsymbol{\mu}})$ increases?

Discriminants for LDA

Discriminant functions based on the joint distribution are

$$p(Y=k)p(x \mid Y=k) = \pi_k \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)}$$
$$\propto \pi_k e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)}$$

(suppressing factors that do not depend on k)

Taking logs and multiplying by 2 gives a simpler expression

$$g_k(\boldsymbol{x}) = 2 \log \pi_k - (\boldsymbol{x} - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)$$

Q: What happens to $g_k(\boldsymbol{x})$ when π_k increases? $(\boldsymbol{x} - \underline{\boldsymbol{\mu}})$ increases?

. . . A point is classified to the closest mean in terms of the Mahalanobis distance, except we also need to account for the class priors.

Constructing the decision boundaries for LDA

Classify $\boldsymbol{x}$ by choosing k with highest discriminant

$$g_k(\boldsymbol{x}) = 2\log \pi_k - (\boldsymbol{x} - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)$$

The decision boundary between class j and k consists of all points where

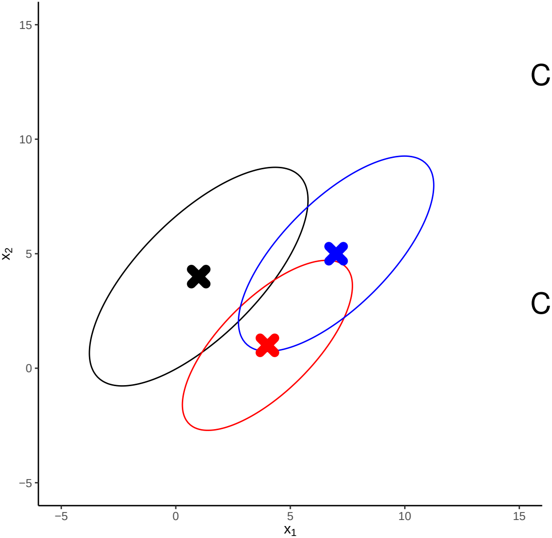
$$g_j(x) = g_k(x)$$

Intersection points between contour curves for a value g are where

$$g_j(x) = g_k(x) = g.$$

The intersection between the curves is a hyperplane: for one feature a point, for two features a line, for three features a plane. (see this by writing out $g_k(x) = g_j(x)$.)

Constructing the decision boundaries for LDA



Class conditionals, $p(x \mid y)$:

$$p(x \mid \text{black}) = \text{MVN}\left(\begin{bmatrix}2\\3\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$

$$p(x \mid \text{red}) = \text{MVN}\left(\begin{bmatrix}4\\1\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$

$$p(x \mid \text{blue}) = \text{MVN}\left(\begin{bmatrix}7\\5\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$

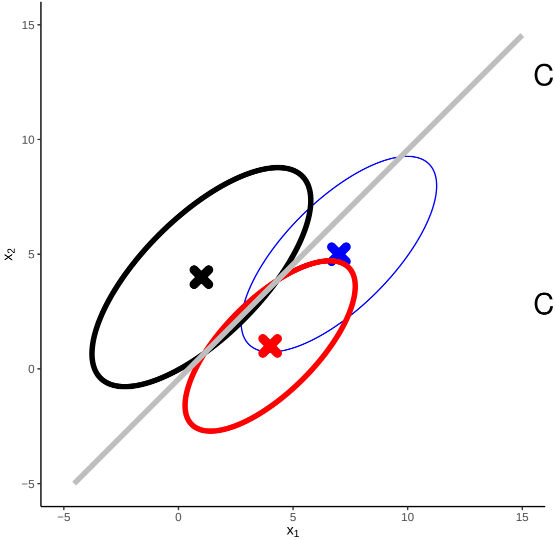
Class probabilities:

$$\pi_{\text{black}} = 0.9$$

$$\pi_{\text{red}} = 0.01$$

$$\pi_{\text{blue}} = 0.09$$

Constructing the decision boundaries for LDA



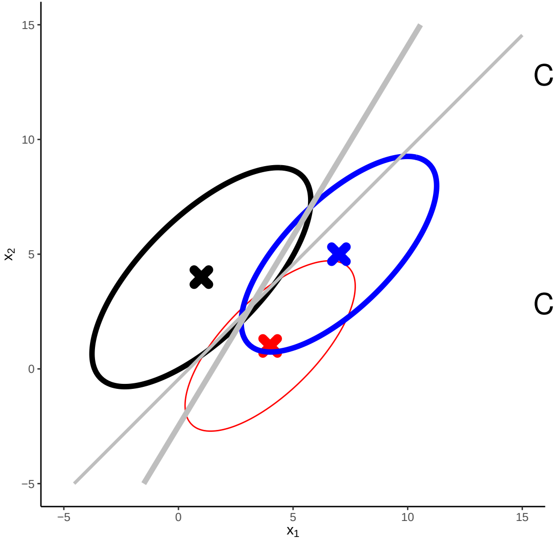
Class conditionals, $p(x \mid y)$:

$$p(x \mid \text{black}) = \text{MVN}\left(\begin{bmatrix}2\\3\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$
$$p(x \mid \text{red}) = \text{MVN}\left(\begin{bmatrix}4\\1\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$
$$p(x \mid \text{blue}) = \text{MVN}\left(\begin{bmatrix}7\\5\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$

Class probabilities:

$$\pi_{\text{black}} = 0.9$$
$$\pi_{\text{red}} = 0.01$$
$$\pi_{\text{blue}} = 0.09$$

Constructing the decision boundaries for LDA



Class conditionals, $p(x \mid y)$:

$$p(x \mid \text{black}) = \text{MVN}\left(\begin{bmatrix}2\\3\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$

$$p(x \mid \text{red}) = \text{MVN}\left(\begin{bmatrix}4\\1\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$

$$p(x \mid \text{blue}) = \text{MVN}\left(\begin{bmatrix}7\\5\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$

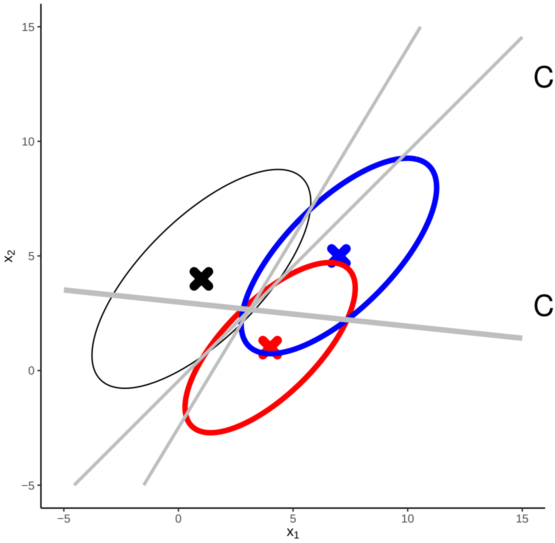
Class probabilities:

$$\pi_{\text{black}} = 0.9$$

$$\pi_{\text{red}} = 0.01$$

$$\pi_{\text{blue}} = 0.09$$

Constructing the decision boundaries for LDA



Class conditionals, $p(x \mid y)$:

$$p(x \mid \text{black}) = \text{MVN}\left(\begin{bmatrix}2\\3\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$

$$p(x \mid \text{red}) = \text{MVN}\left(\begin{bmatrix}4\\1\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$

$$p(x \mid \text{blue}) = \text{MVN}\left(\begin{bmatrix}7\\5\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$

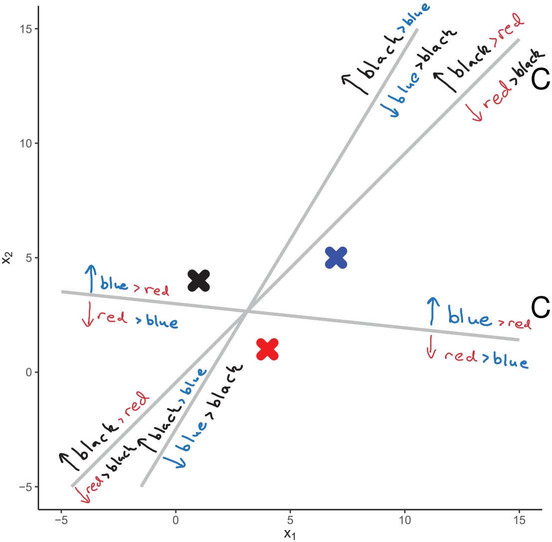
Class probabilities:

$$\pi_{\text{black}} = 0.9$$

$$\pi_{\text{red}} = 0.01$$

$$\pi_{\text{blue}} = 0.09$$

Constructing the decision boundaries for LDA



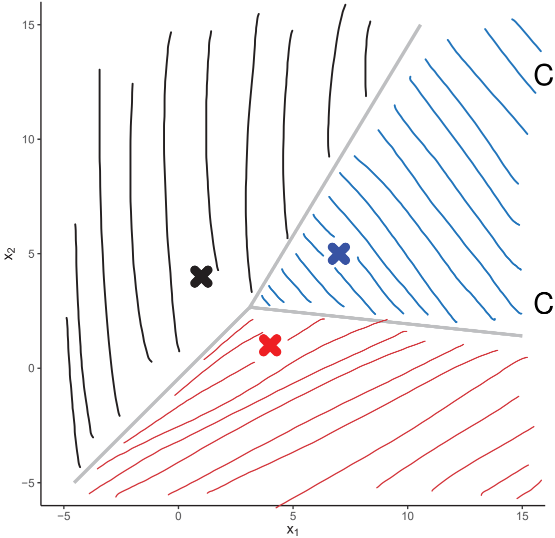
Class conditionals, $p(x \mid y)$:

$$p(x \mid \text{black}) = \text{MVN}\left(\begin{bmatrix}2\\3\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$
$$p(x \mid \text{red}) = \text{MVN}\left(\begin{bmatrix}4\\1\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$
$$p(x \mid \text{blue}) = \text{MVN}\left(\begin{bmatrix}7\\5\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$

Class probabilities:

$$\pi_{\text{black}} = 0.9$$
$$\pi_{\text{red}} = 0.01$$
$$\pi_{\text{blue}} = 0.09$$

Constructing the decision boundaries for LDA



Class conditionals, $p(x \mid y)$:

$$p(x \mid \text{black}) = \text{MVN}\left(\begin{bmatrix}2\\3\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$

$$p(x \mid \text{red}) = \text{MVN}\left(\begin{bmatrix}4\\1\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$

$$p(x \mid \text{blue}) = \text{MVN}\left(\begin{bmatrix}7\\5\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$$

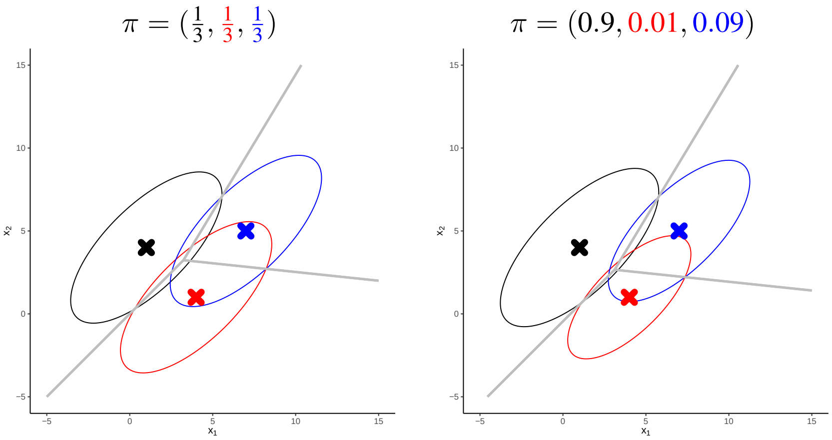
Class probabilities:

$$\pi_{\text{black}} = 0.9$$

$$\pi_{\text{red}} = 0.01$$

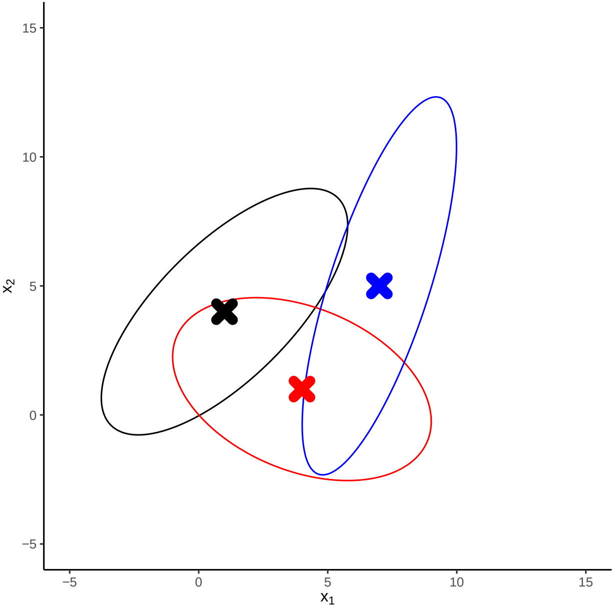
$$\pi_{\text{blue}} = 0.09$$

Decision boundaries: effect of changing priors



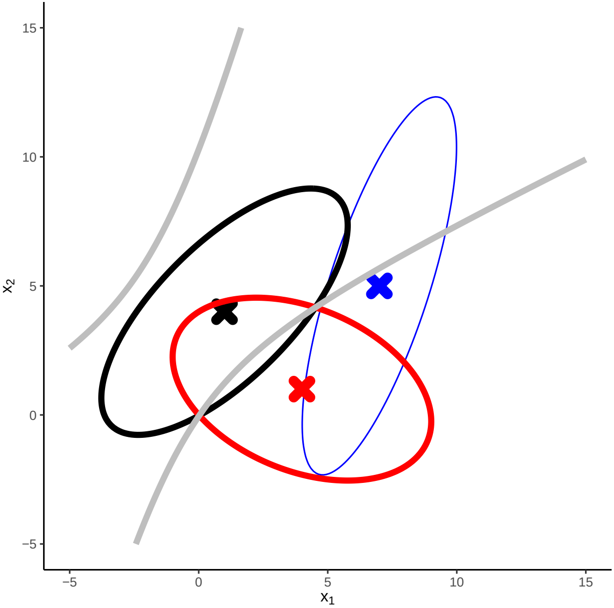
The decision boundary moves away from the mean of the class with highest prior.

Decision boundaries for QDA? More complex!



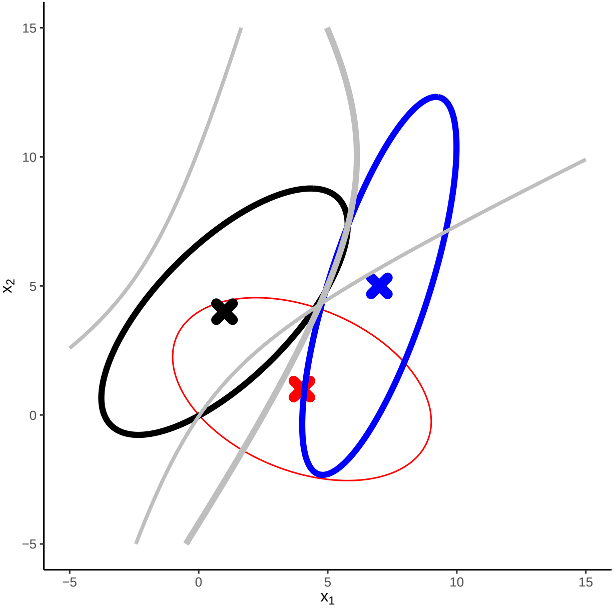
Q: Will boundaries still cross through contour intersections?

Decision boundaries for QDA? More complex!



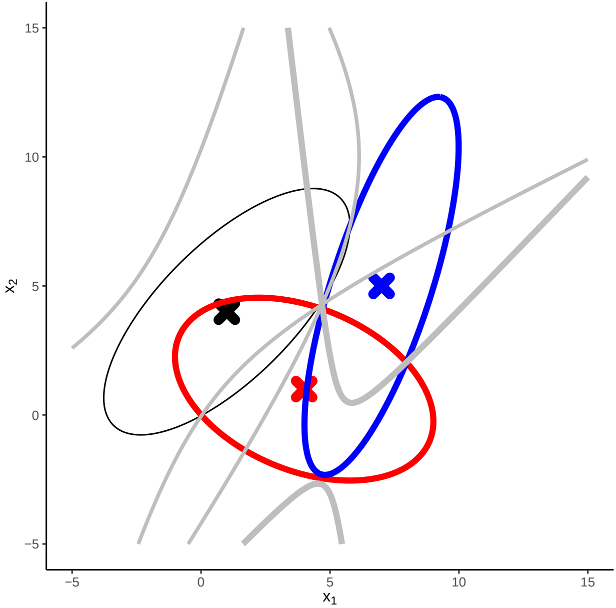
Q: Will boundaries still cross through contour intersections?

Decision boundaries for QDA? More complex!



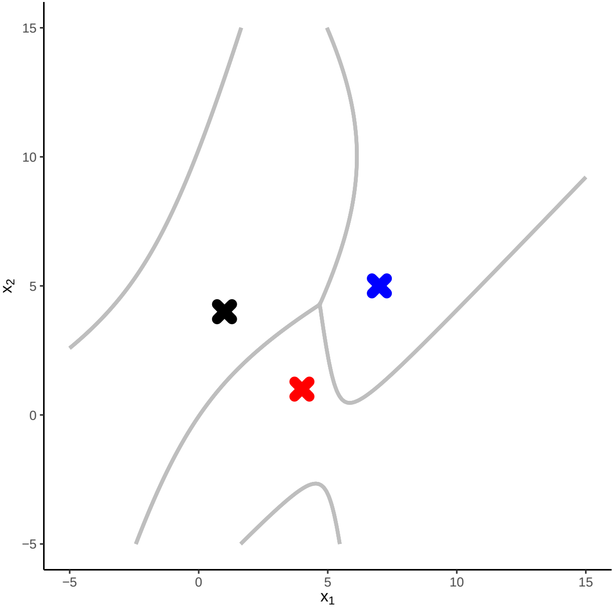
Q: Will boundaries still cross through contour intersections?

Decision boundaries for QDA? More complex!



Q: Will boundaries still cross through contour intersections?

Decision boundaries for QDA? More complex!



Q: Will boundaries still cross through contour intersections?

Issues with LDA and QDA

The model that best captures variability in data may have too many parameters to estimate.

A covariance matrix needs $p(p+1)/2$ parameters.
(Luckily there is a practical reduction in complexity, since only a *function* of them is needed for the decision boundaries.)

QDA:	different covariance matrix for each class.
LDA:	same covariance matrix for all classes.
Gaussian Naive Bayes:	diagonal covariance matrix (different or same for all classes)

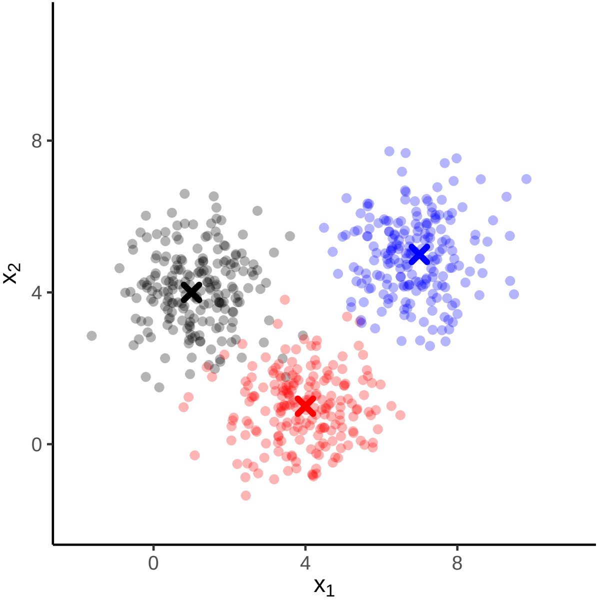
Naive Bayes classifiers are generative models with a simplifying assumption that all features are independent, when specifying the class conditionals.

$$f_k(x) = f_{k1}(x_1) f_{k2}(x_2) \ldots f_{kp}(x_p)$$

[More on these later]

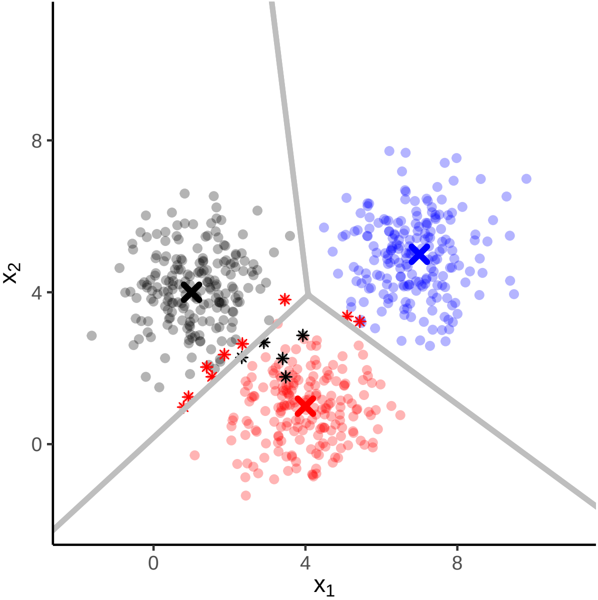
Case studies:
LDA, QDA, (Gaussian) Naive Bayes,
and Logistic Regression

Case 1

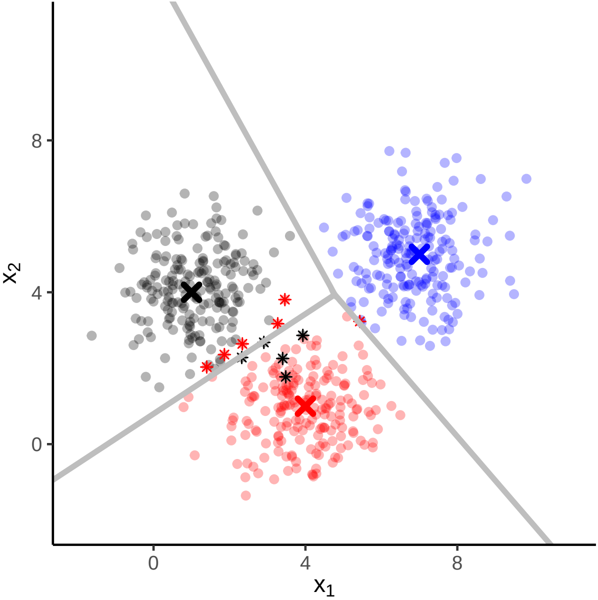


(Equal covariance matrix, no correlations)

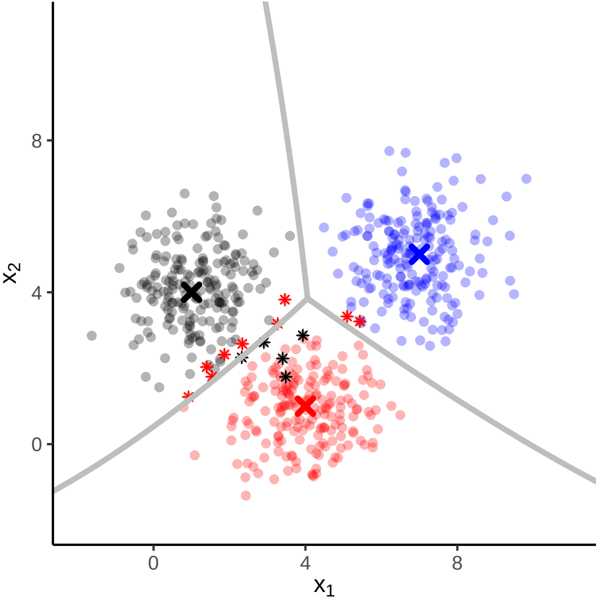
Case 1: LDA



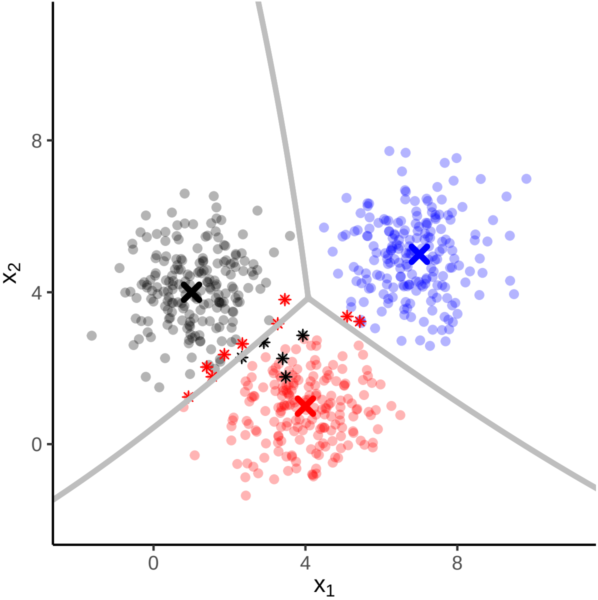
Case 1: Logistic regression (with features x_1 and x_2)



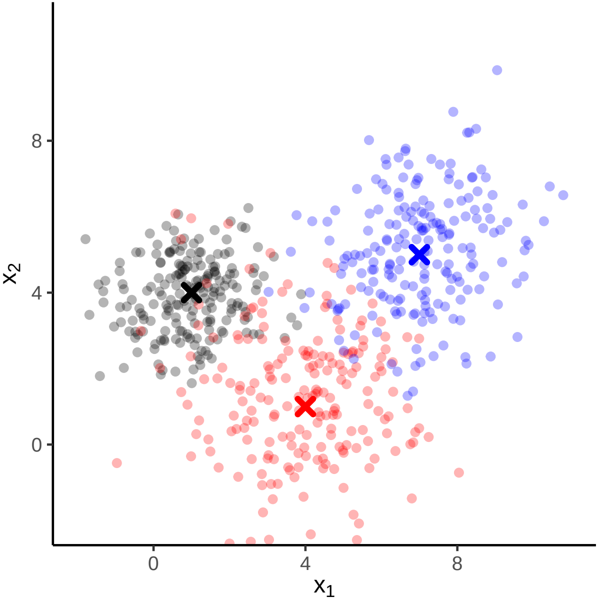
Case 1: QDA



Case 1: Naive Bayes (different, but diagonal Cov)

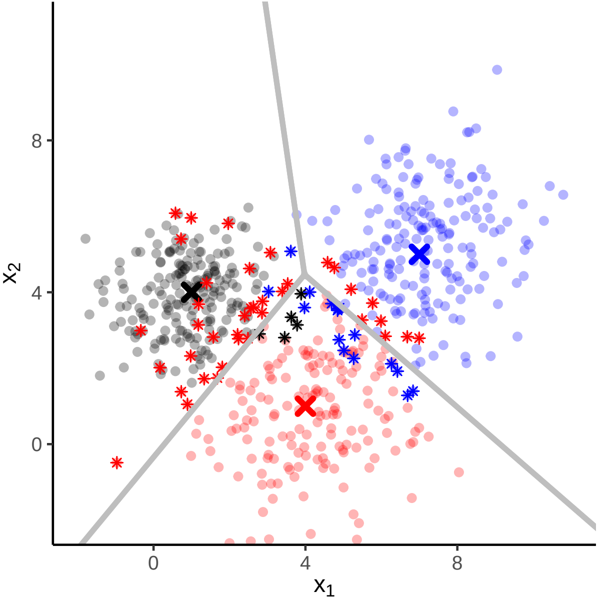


Case 2

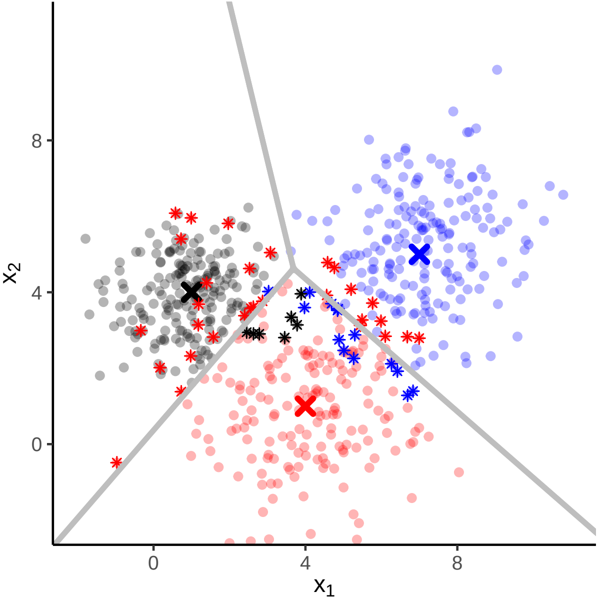


(Unequal covariance matrix, no correlations)

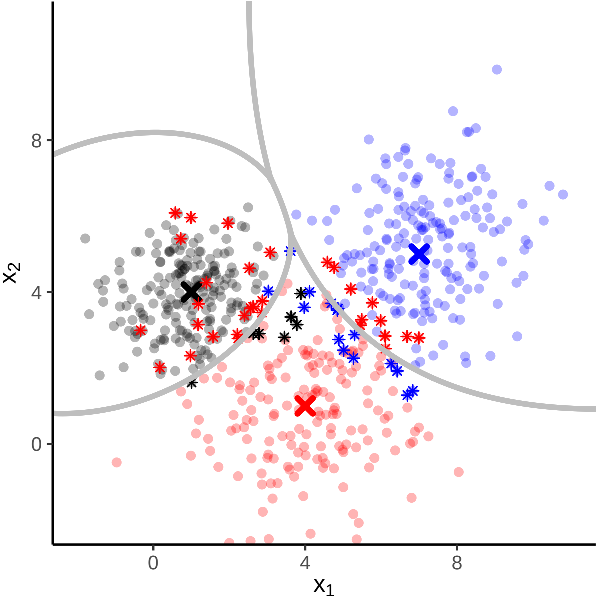
Case 2: LDA



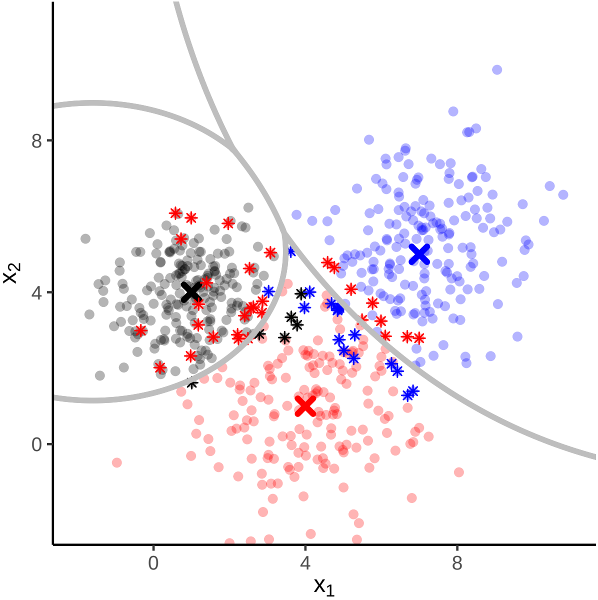
Case 2: Logistic regression (with features x_1 and x_2)



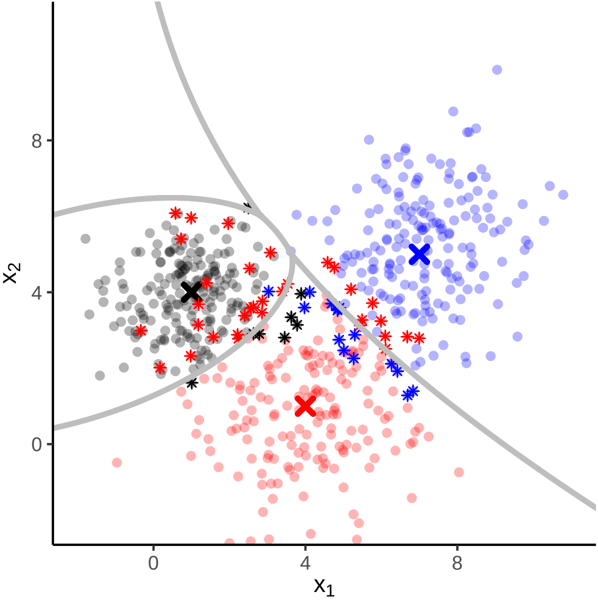
Case 2: QDA



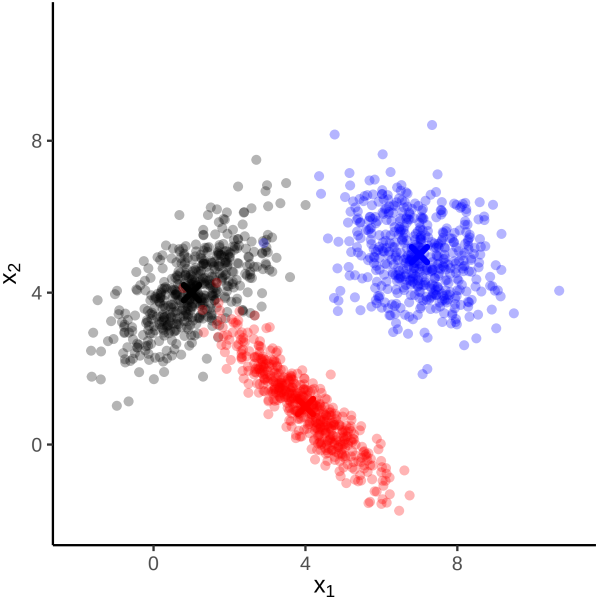
Case 2: Naive Bayes (different, but diagonal Cov)



Case 2: Logistic regression (with features x_1, x_2, x_1x_2)

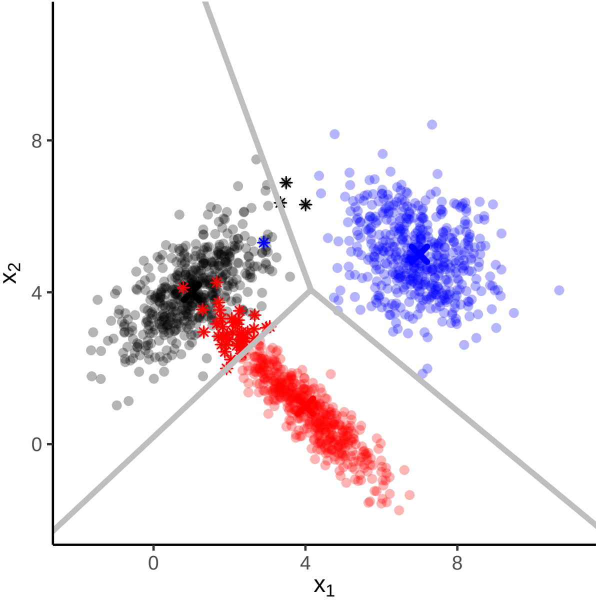


Case 3

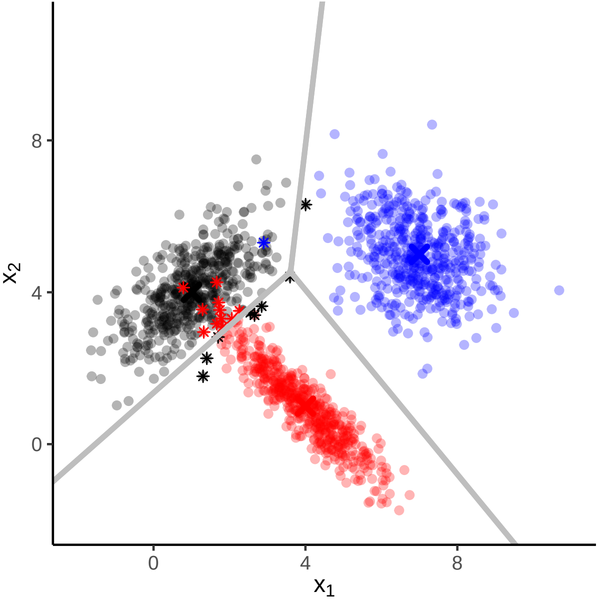


(Unequal covariance matrix, correlations)

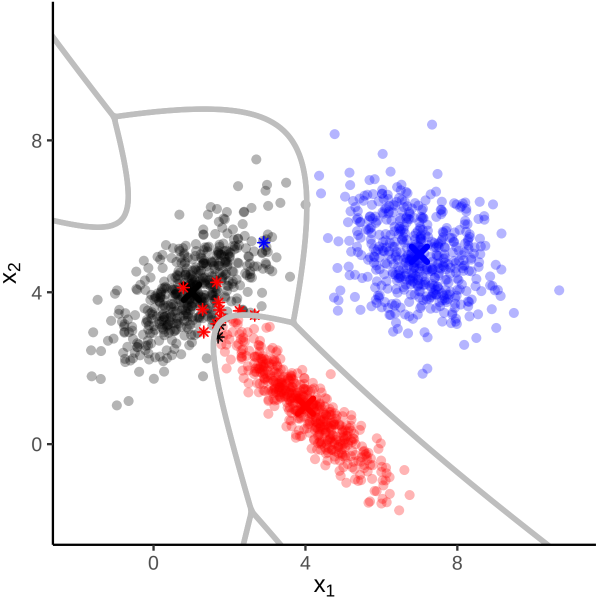
Case 3: LDA



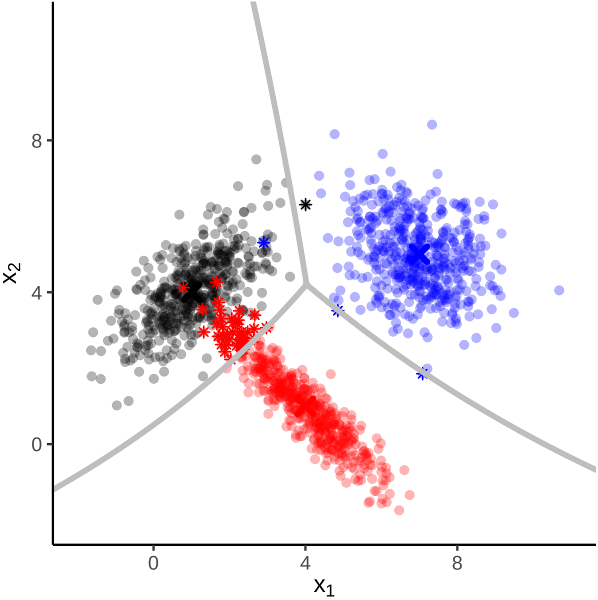
Case 3: Logistic regression (with features x_1, x_2)



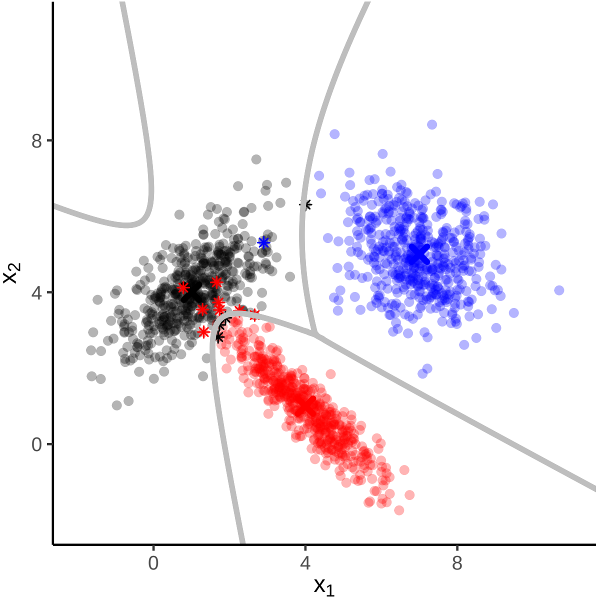
Case 3: QDA



Case 3: Naive Bayes (different, but diagonal Cov)



Case 3: Logistic regression (with features x_1, x_2, x_1x_2)



Test error (misclassification in percent)

Case 1: Equal covariance matrices

Case 2: Unequal covariance matrices without correlation

Case 3: Unequal covariance matrices with correlation

	LDA	QDA	NB	LR	LR(sq.)
Case 1	1.20	1.30	1.33	1.40	1.33
Case 2	8.00	7.27	7.43	7.60	7.57
Case 3	2.27	0.63	2.20	0.87	0.60

LDA vs Logistic regression

We classify to class k over K, whenever $\mathrm{P}(Y = k \,|\, x) > \mathrm{P}(Y = K \,|\, x)$. That is, whenever

$$\log\left(\frac{\mathrm{P}(Y = k \,|\, x)}{\mathrm{P}(Y = K \,|\, x)}\right) > 0.$$

Logistic regression directly models log of posterior odds between class k and K as a linear combination of the features:

$$\log\left(\frac{\mathrm{P}(Y = k \,|\, x)}{\mathrm{P}(Y = K \,|\, x)}\right) = \boldsymbol{a}_k + \boldsymbol{b}_k^T x.$$

LDA assumptions imply a *linear* model for the log of posterior odds!

QDA assumptions imply a *quadratic* model for log of posterior odds.

Model quadratic decision boundaries with LR through quadratic functions of features.

LDA vs logistic regression

Both models give linear decision boundaries.

LDA assumes that class conditionals are Gaussian.

If the normality assumptions behind LDA are true, then it is a more efficient classifier than one based on logistic regression.

Logistic regression offers more flexibility in creating the decision boundaries:

- add more predictors, transformations of predictors, or interaction terms.
- straightforward to include non-continuous features