Lecture 8 - 22/09/2024

- With multiple features we need multivariate distributions for the class conditionals.

Multivariate Data

- If $X_1, \dots, X_p$ are real-valued random variables:
  - We can talk about a p-dimensional random vector $X = (X_1, \dots, X_p)$
- Usual convention → random vectors = column vectors:
  - $X = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$
- The (expectation / mean) of a p-dimensional random vector X is:
  - $EX = \begin{bmatrix} EX_1 \\ \vdots \\ EX_p \end{bmatrix}$
- The sample mean for vector X is the avg:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$ → vector of coordinate-wise avgs.

- The variance of a p-dimensional random vector X is:

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

Multivariate Normal Distribution

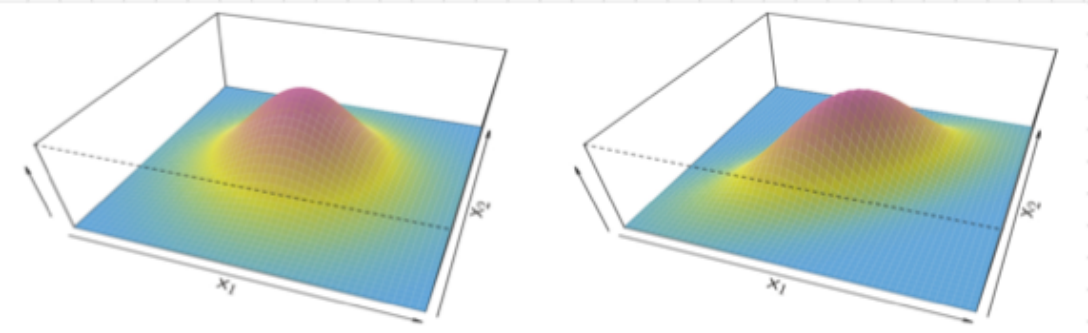


FIGURE 4.5. *Two multivariate Gaussian density functions are shown, with* $p = 2$. Left: *The two predictors are uncorrelated.* Right: *The two variables have a correlation of* 0.7.

- The p-dimensional multivariate Gaussian dist has prob density function:

variance

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

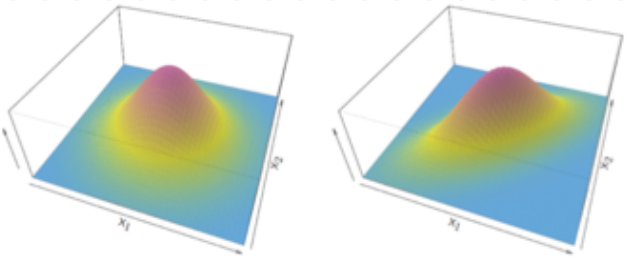
The exponent contains the (squared) *Mahalanobis distance* based on Σ:

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

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

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

is the squared geometric (Euclidean) distance from x to μ.



Looking only at factors involving x the pdf is

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

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

(1x1) (nxn) (nx1)

1x1 → SCALAR

n → # OF RANDOM VARS

A CONTOUR CURVE FOR THE PDF:

* SHOWS ALL POINTS WITH A GIVEN DENSITY VAL.
* IS A QUADRATIC FORM
* SHOWS ALL POINTS WITH DISTANCE FROM μ
* NO CHANGING SHAPE WHEN WE TAKE LOG.

IF X IS P-DIMENSIONAL MULTIVARIATE NORMAL, $MVN(\mu,\Sigma)$, THEN:

$$Ax+b = MVN(A\mu+b, A\Sigma A^T)$$

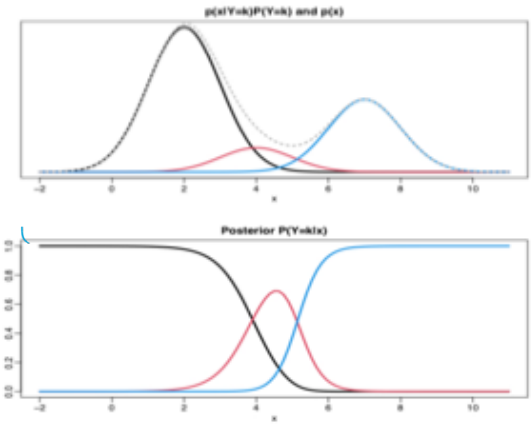
WHERE A IS ANY $q \times p$-MATRIX

LDA & QDA WITH MULTIPLE FEATURES!

* REMEMBER...
  - LDA IS A GENERATIVE MODEL WHERE CLASS CONDITIONALS $p(x|y=k)$ ARE ASSUMED GAUSSIAN WITH INDIVIDUAL CLASS MEANS, BUT EQUAL COVARIANCE MATRICES.
  - QDA IS THE SAME APART FROM THE FACT THAT THEY HAVE CLASS-SPECIFIC COVARIANCE MATRIX.

LDA EXAMPLE

- IN 1 DIMENSION WE HAVE WHAT WE'VE SEEN TILL NOW:



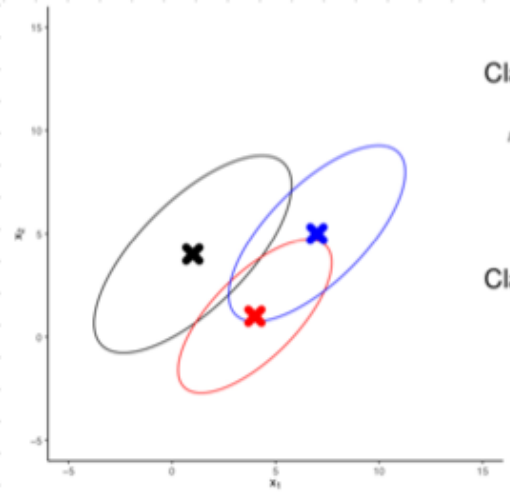
In 1 dimension: Conditionally on the class we assumed the feature x to have a univariate Gaussian distribution as

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

The class probabilities we took to be

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

- NOW BY USING MULTIPLE FEATURES:



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:

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

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

REMEMBER:

- $\text{MVN}(\mu, \Sigma)$

μ = MEAN

$\Sigma = \begin{bmatrix} \text{VAR} & \cdots & \text{COVAR} \\ \vdots & \ddots & \vdots \\ \text{COVAR} & \cdots & \text{VAR} \end{bmatrix}$

- THIS FUNCTION IS A BIT UGLY SO WE CAN GET RID OF SOME PARAMS:
  - TO DO THIS WE CAN TAKE LOGS & MULTIPLY BY 2

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

- A POINT IS CLASSIFIED TO THE CLOSEST MEAN IN TERMS OF MAHALANOBIS DISTANCE.

CONSTRUCTING DECISION BOUNDARY

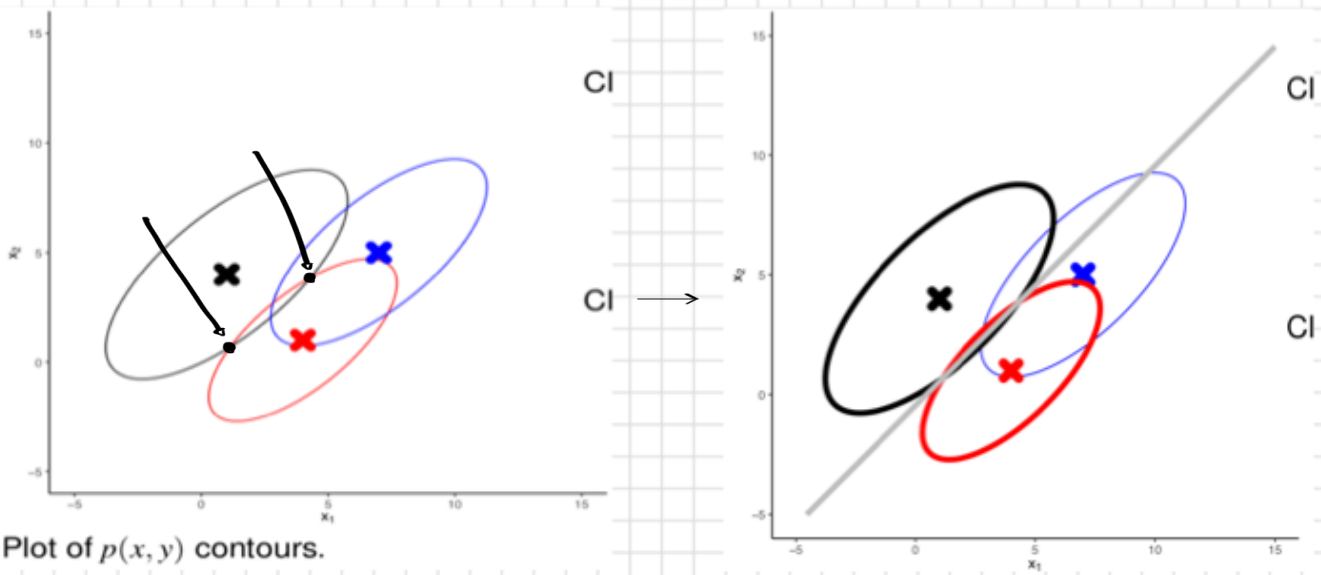
- WE WANNA CLASSIFY X BY CHOOSING K WITH HIGHEST DISCRIMINANT
- THE DECISION BOUNDARY BETWEEN j & k CONSISTS OF ALL POINTS WHERE:

$$g_j(x) = g_k(x) \longrightarrow \text{INTERSECTION}$$

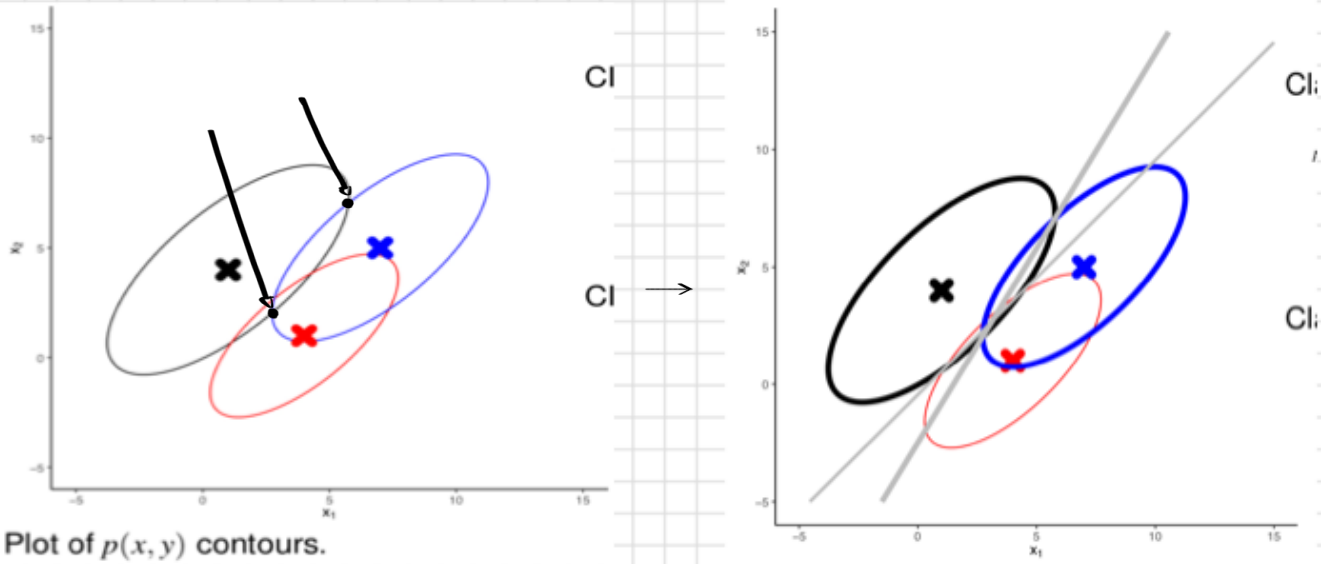
- THE INTERSECTION POINTS BETWEEN CONTOUR CURVES FOR A VALUE g ARE WHERE:

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

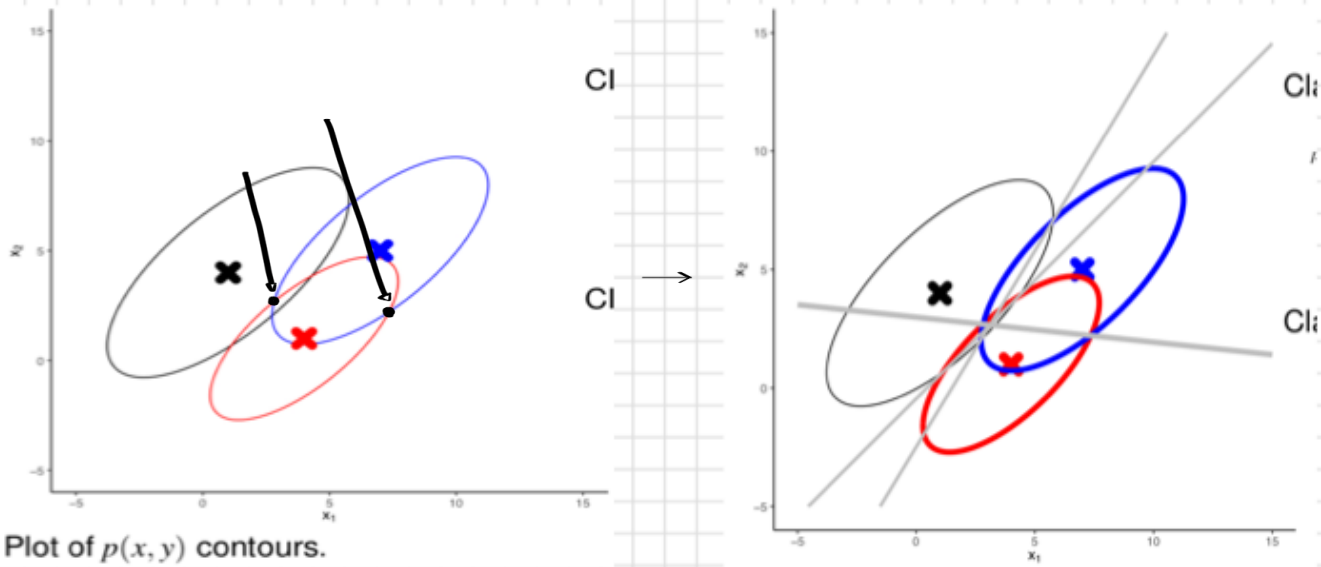
- WE CAN THE INTERSECTION BETWEEN THE CURVES A HYPERPLANE:
  - FOR ONE FEATURE A POINT
  - FOR TWO FEATURES A LINE
  - FOR THREE FEATURES A PLANE.



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

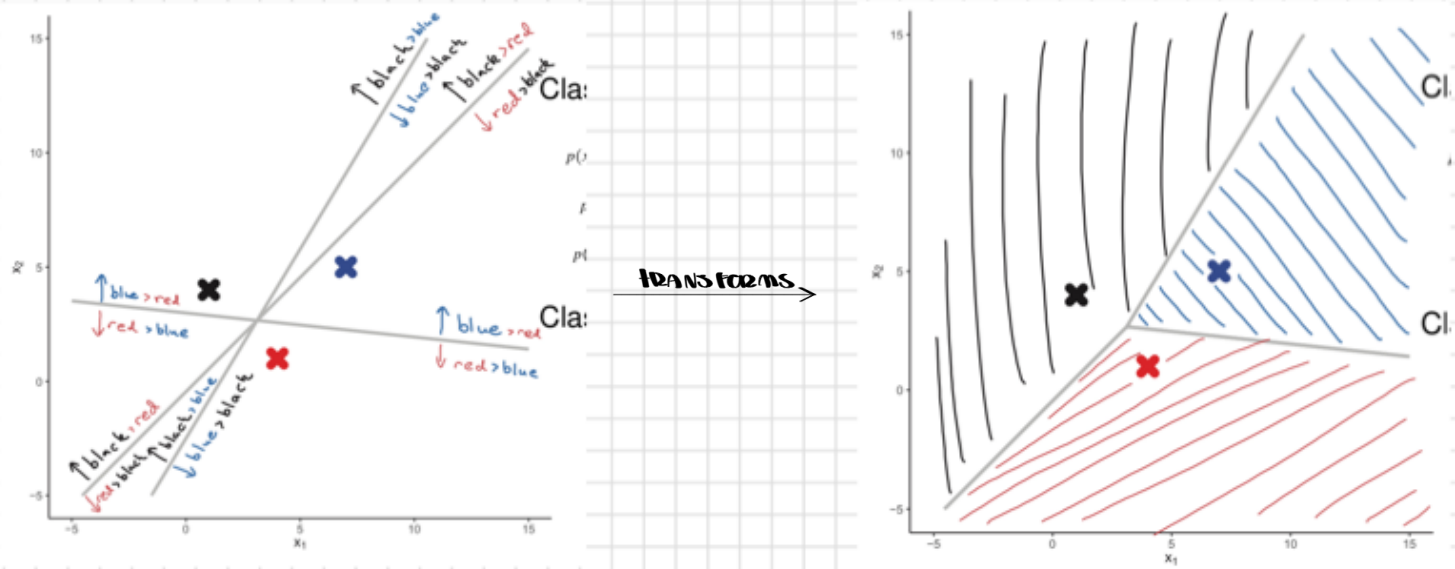


Plot of $p(x, y)$ contours.



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

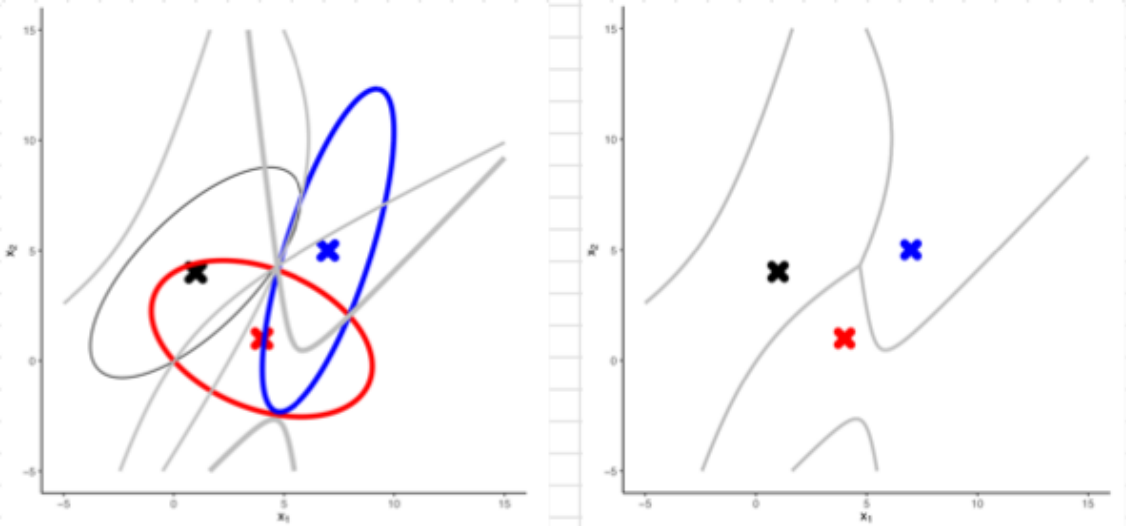
- NOW THAT WE WEREBE ABLE TO CONSTRUCT THE DECISION BOUNDARIES
  - WE NEED TO ACTUALLY COVER THE DIVIDED AREA WITH THE CLASSES



- By CHANGING PRIORS THE DECISION BOUNDARY MOVES AWAY:
  - IT MOVES AWAY FROM THE MEAN OF K WITH HIGHEST PRIOR.

Constructing Decision Boundary for QDA!

- It is definitely more complex to construct for QDA:
  - This is due to its flexibility (curve shaped)



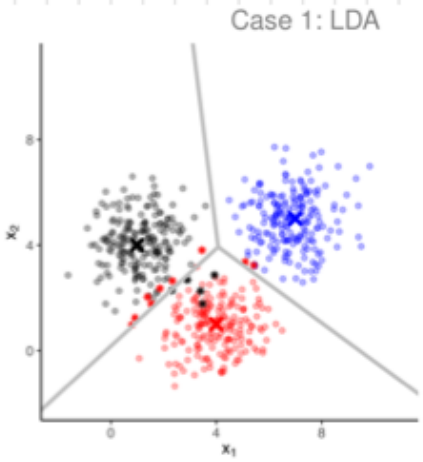
Issues:

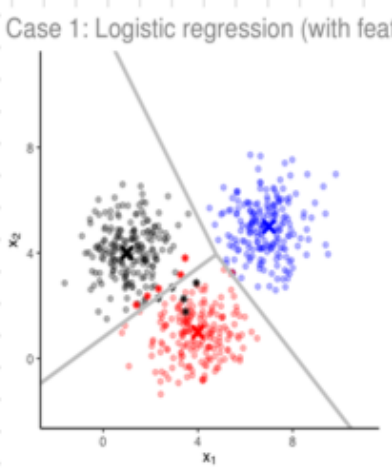
- The model that best captures variability in data may have too many params
  - A covariance matrix need $p(p+1)/2$ params.
  - Luckily ∃ a practical reduction
    - QDA → ≠ covariance matrix ∀ k
    - LDA → same covariance matrix for all classes
- Gaussian Naive Bayes:
  - Diagonal covariance matrix (≠ or !≠ ∀ k)

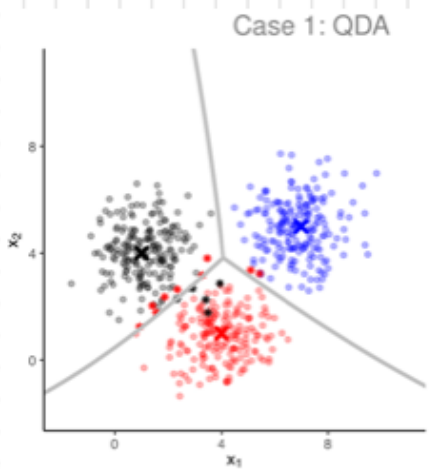
Note!

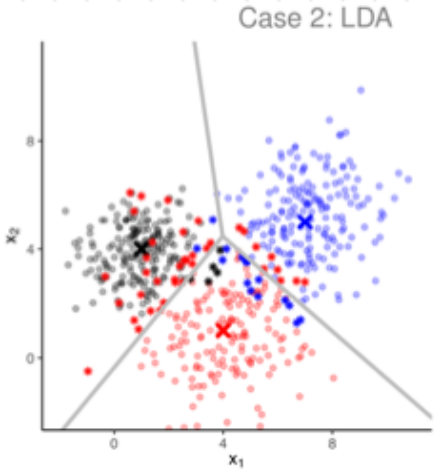
> Naive Bayes classifiers are generative models with a simplifying assumption that all features are independent, when specifying class conditionals.

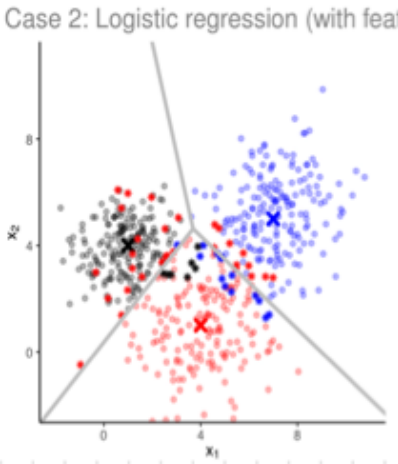
Case studies for LDA, QDA, Logistic Regression!

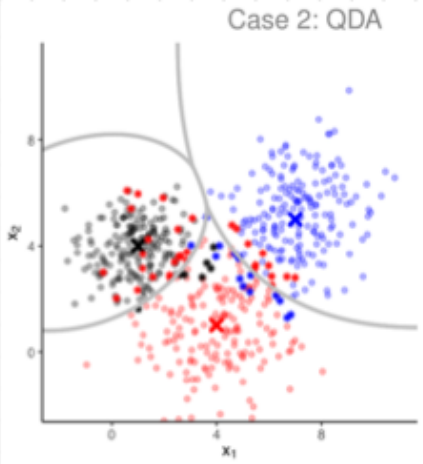


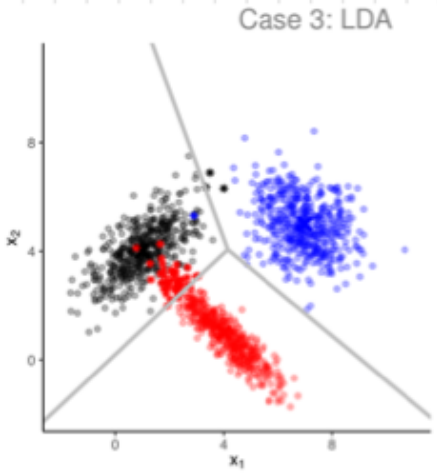


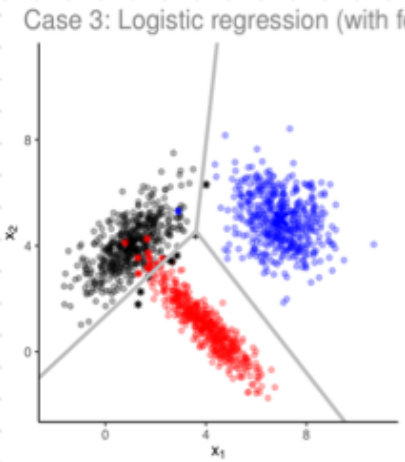


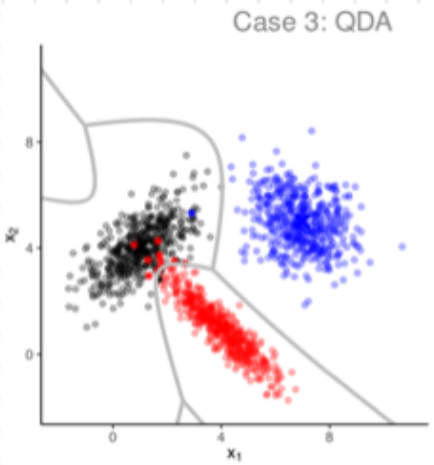




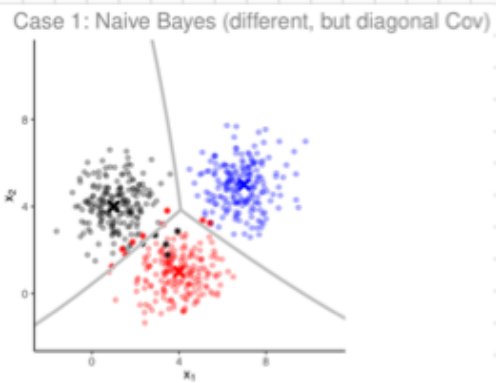


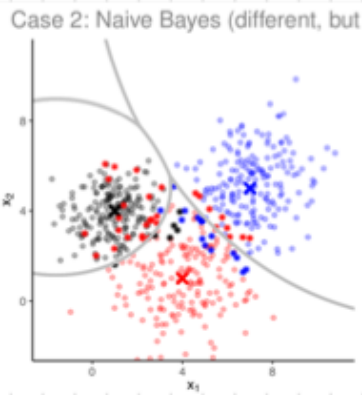


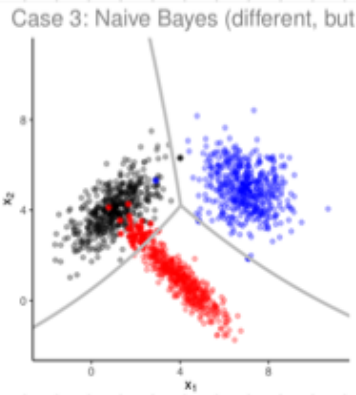




Case Study for Naive Bayes







Test Error for the three cases

- Case 1 → Equal covariance matrices
- Case 2 → Unequal covariance matrices, no 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

small K

Big K

- What happens is:
  - We classify to k over K whenever:

$$P(y=k|x) > P(y=K|x) \xrightarrow{\text{that is}} \log\left(\frac{P(y=k|x)}{P(y=K|x)}\right) > 0$$

- What **Logistic regression** does:
  - Directly models **log** of posterior odds between k & K:
    - It models it as a linear comb → $\log\left(\frac{P(y=k|x)}{P(y=K|x)}\right) = \boxed{a_k + b_k^T x}$

- What **LDA** does:
  - LDA assumptions imply a linear model for the log of posterior odds
  - QDA implies a quadratic model.
  - LDA assumes that class conditionals are Gaussian
    - If true, then LDA is a more efficient classifier than LR

LR offers more flexibility in creating decision boundaries (X_1X_2, + predictors, etc.-)

Both model GWE linear decision boundaries!!!