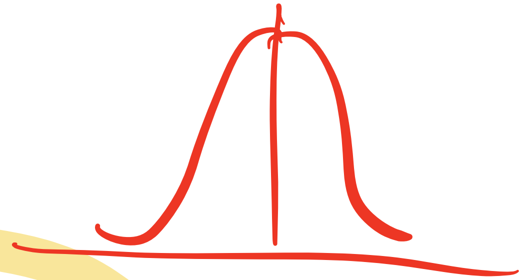
We will use the Bayesian decision criteria applied to normally distributed classes, whose parameters are either known or estimated from the sample.

Parametric Classification

Parametric Classification



- If $p(\mathbf{x} \mid C_i) = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

$$p(\mathbf{x} \mid C_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)\right]$$

- Discriminant functions are:

→ think as score for C_i .

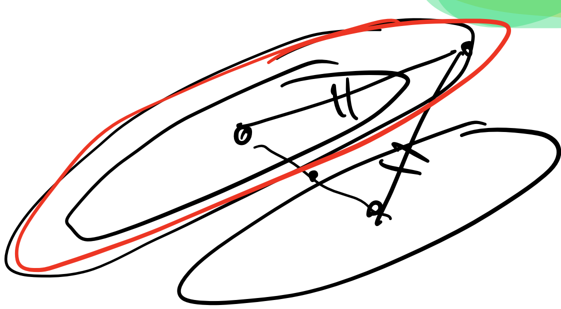
$g_i(x) = P(C_i \mid x)$

↓ $i = 1$ to K classes

$$g_i(\mathbf{x}) = \log P(C_i \mid \mathbf{x}) = \log\left(\underline{P(x \mid C_i)} \cdot \underline{P(C_i)}\right)$$

$$= \log p(\mathbf{x} \mid C_i) + \log P(C_i)$$

$$= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i) + \log P(C_i)$$



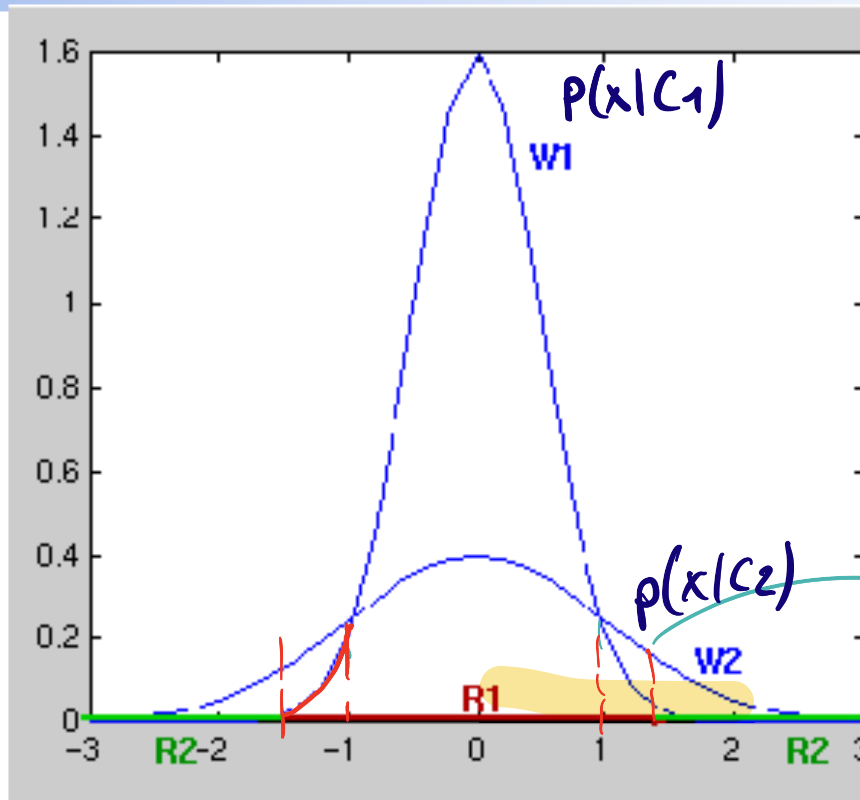
Estimation of Parameters

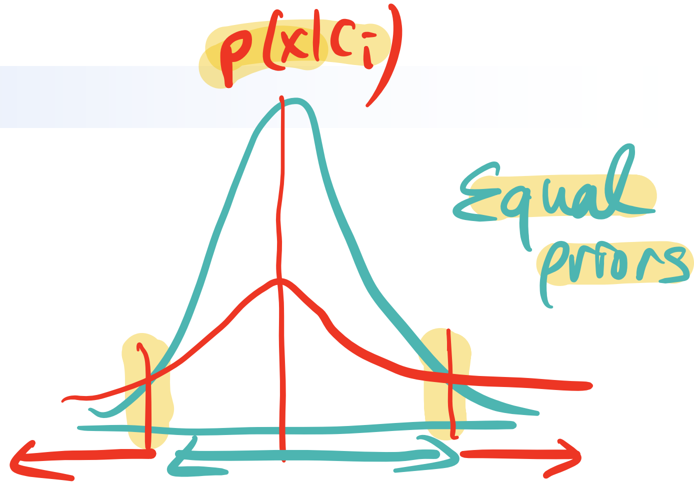
If we estimate the unknown parameters from the sample, the discriminant function becomes:

$$g_i(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \mathbf{S}_i^{-1}(\mathbf{x}-\mathbf{m}_i) + \log \hat{P}(C_i)$$

μ, σ: population parameters

m, S: sample "





$p(C_1) = 0.9$
$p(C_2) = 0.1$

Typical single-variable normal distributions showing a disconnected decision region R_2

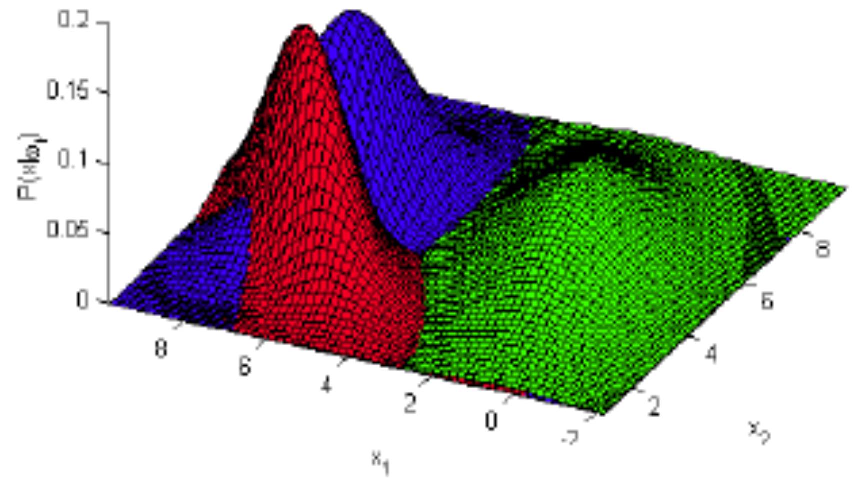
$$\exp\left\{\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}$$

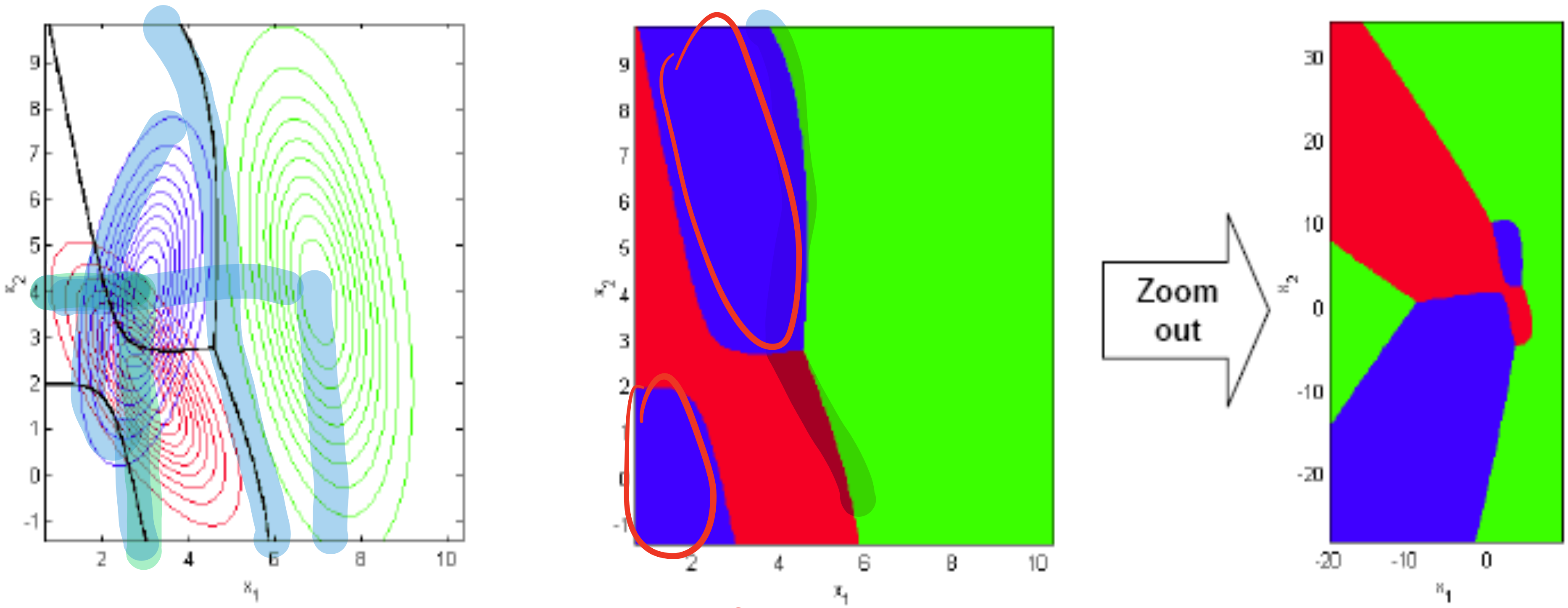
- To illustrate the previous result, we will compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors

$$\mu_1 = [3 \quad 2]^T \qquad \mu_2 = [5 \quad 4]^T \qquad \mu_3 = [2 \quad 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 1 & -1 \\ -1 & 7 \end{bmatrix} \qquad \Sigma_3 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 3 \end{bmatrix}$$

red   green   blue





Boundaries

- If d (dimension) is large with respect to N (number of samples), we may have a problem with this approach:
  - | Σ | may be zero, thus Σ will be singular (inverse does not exist)
  - | Σ | may be non-zero, but very small, instability would result
    - Small changes in Σ would cause large changes in Σ^{-1}

- Solutions:
  - Reduce the dimensionality
    - Feature selection
    - Feature extraction: PCA,…

  - Pool the data and estimate a common covariance matrix for all classes

$$\Sigma = \Sigma_i P(C_i) * \Sigma_i$$

$$\sigma_{12} = \frac{1}{N} \sum_{i=1}^{N} (x_1^{(i)} - \mu_1)(x_2^{(i)} - \mu_2)$$

- In the following slides, **we will make increasing assumptions about the covariance matrix** and see what the corresponding discriminant function and resulting boundaries look like.
  - QDA, LDA, Naïve Bayes, Nearest Mean classifiers

Case 2) Common Covariance Matrix $\mathbf{S}=\mathbf{S}_i$

$\boxed{S_i = S}$

- Shared common sample covariance **S**
  - An arbitrary covariance matrix – **but shared between the classes**

- We had this full discriminant function:

$$g_i(\mathbf{x}) = \cancel{-\frac{1}{2}\log|\mathbf{S}|} - \frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \mathbf{S}_i^{-1}(\mathbf{x}-\mathbf{m}_i) + \log \hat{P}(C_i)$$

which now reduces to:

$$g_i(\mathbf{x}) = \underbrace{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \mathbf{S}_{\cancel{i}}^{-1}(\mathbf{x}-\mathbf{m}_i) + \log \hat{P}(C_i)}$$

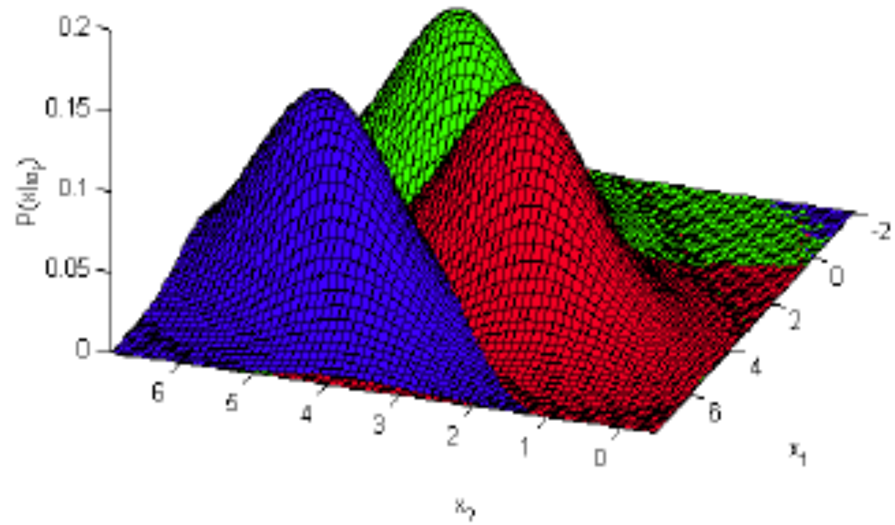
which is a **linear discriminant** (decision boundaries are hyper-planes)

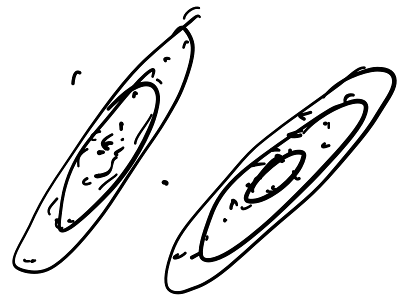
Which class to assign x to? $\max_i g_i(x)$

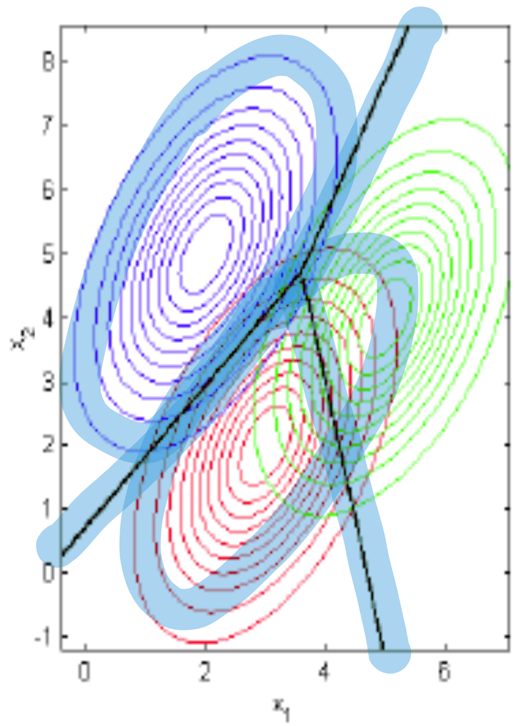
- **To illustrate the previous result, we will compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors**

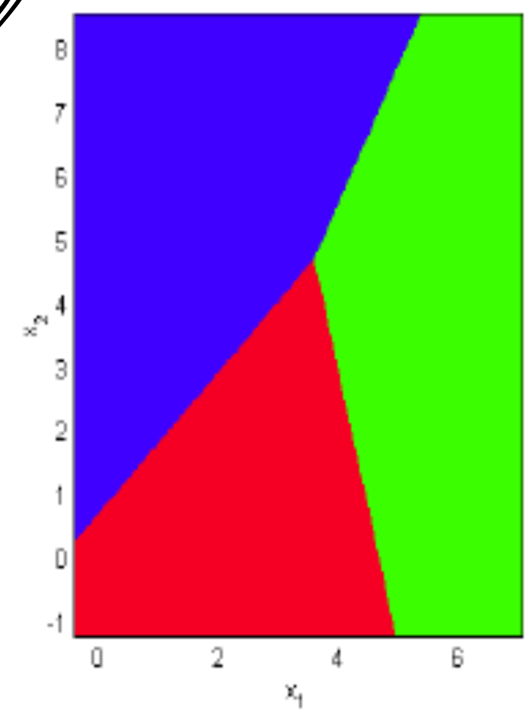
$$\mu_1 = [3 \quad 2]^T \qquad \mu_2 = [5 \quad 4]^T \qquad \mu_3 = [2 \quad 5]^T$$

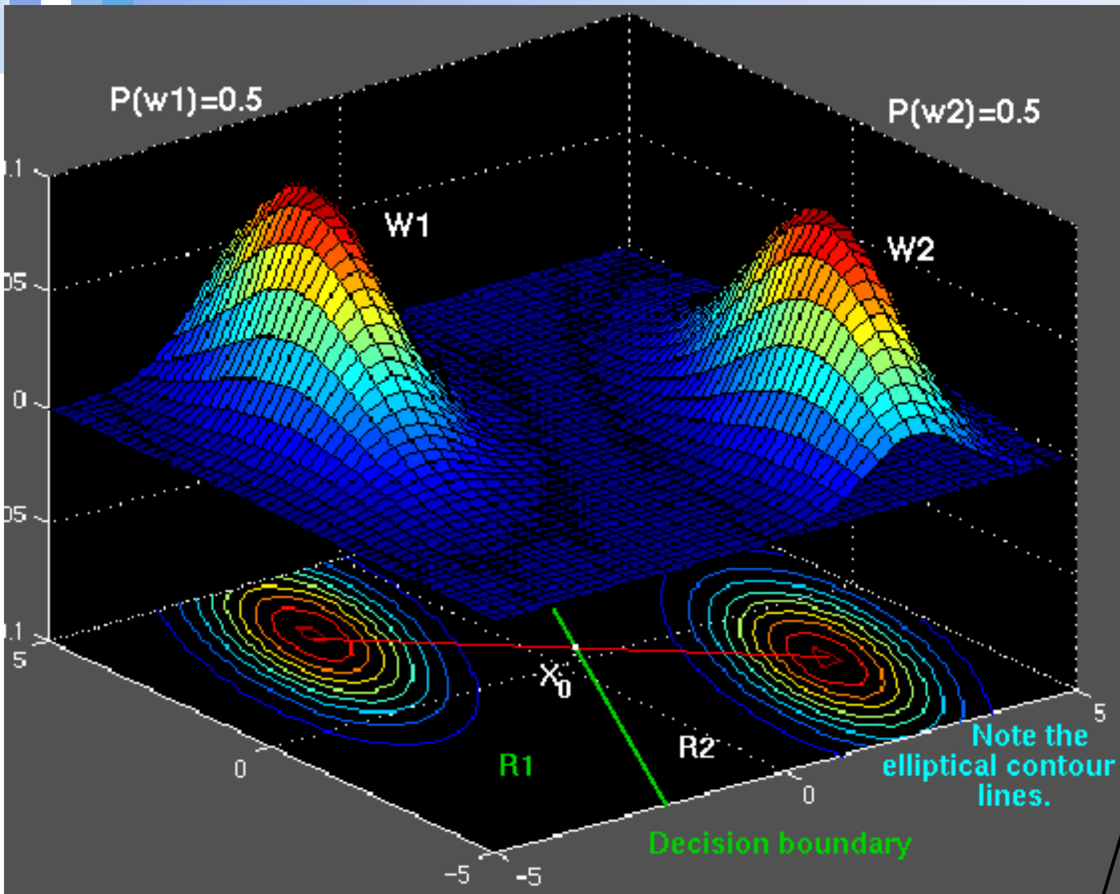
$$\Sigma_1 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix} \qquad \Sigma_3 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix}$$





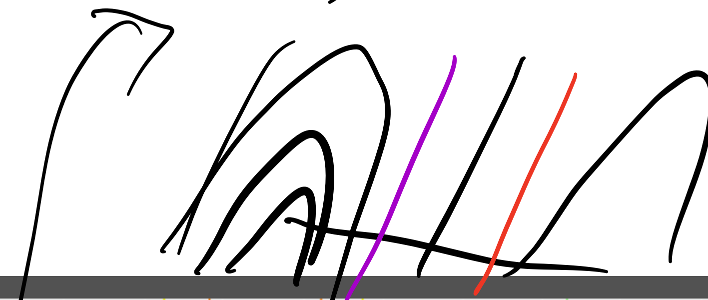


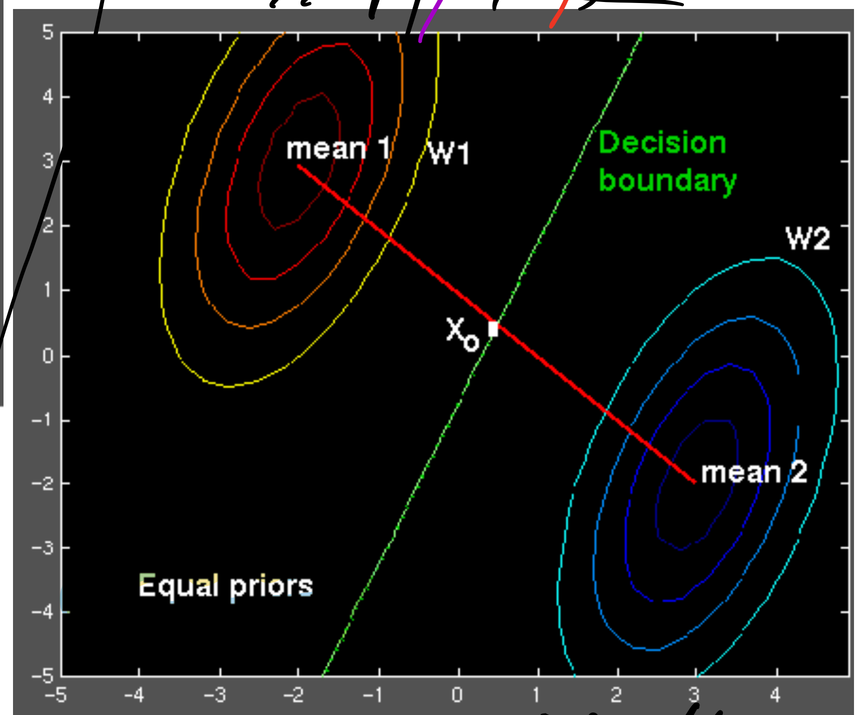




Bound. wd move tw. w_2

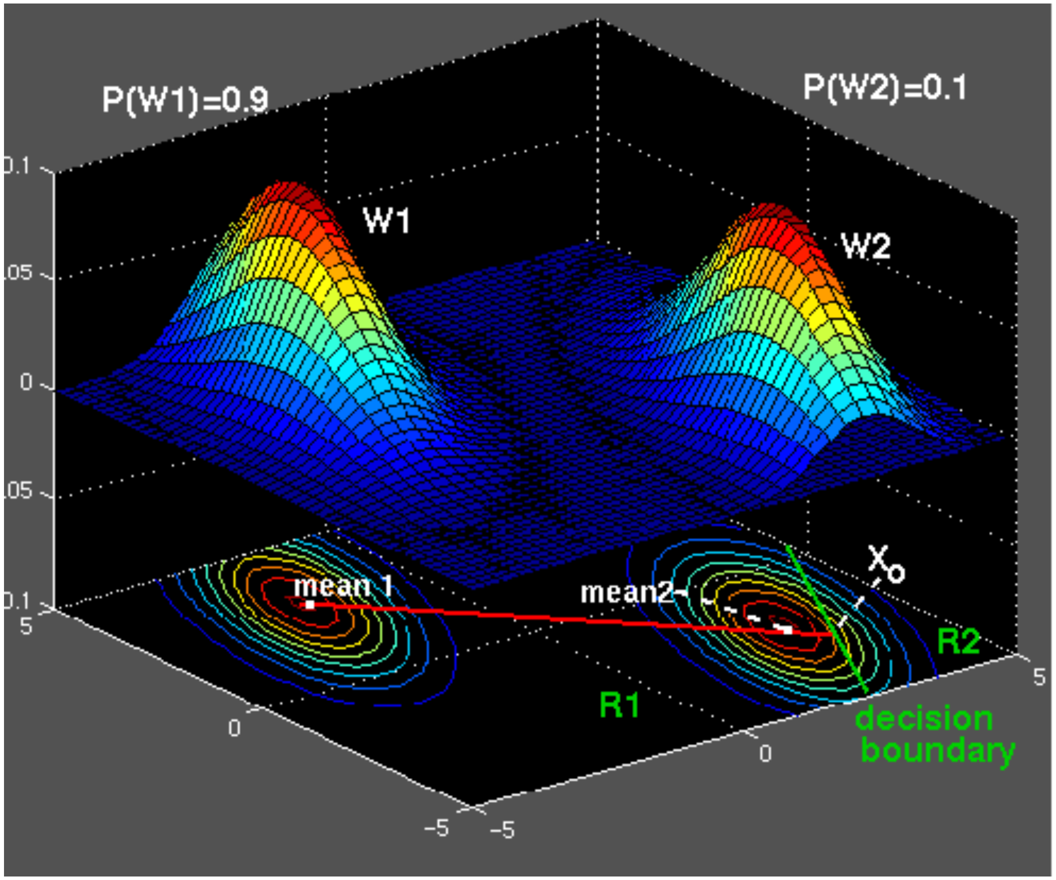
$P(C_1) >> P(C_2)$





If we assume equal class priors, the classifier becomes a minimum Mahalanobis classifier

Not equal priors!
what happens -



Unequal priors shift the decision boundary towards the less likely class.

Case 3) Common Covariance Matrix S which is Diagonal

$S_i = S +$ Diagonal

- In the previous case, we had a common, general covariance matrix, resulting in these discriminant functions:

$$S = \begin{bmatrix} \sigma_1^2 & & \phi \\ & \sigma_2^2 & \\ \phi & & \sigma_3^2 \end{bmatrix}$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \mathbf{S}^{-1}(\mathbf{x}-\mathbf{m}_i) + \log \hat{P}(C_i)$$

- **When x_j (j = 1,..d) are independent** (or assumed to be independent for simplicity)**, then Σ is diagonal:**

Case 3) Common Covariance Matrix S which is Diagonal

- In the previous case, we had a common, general covariance matrix, resulting in these discriminant functions:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \mathbf{S}^{-1}(\mathbf{x}-\mathbf{m}_i) + \log \hat{P}(C_i)$$

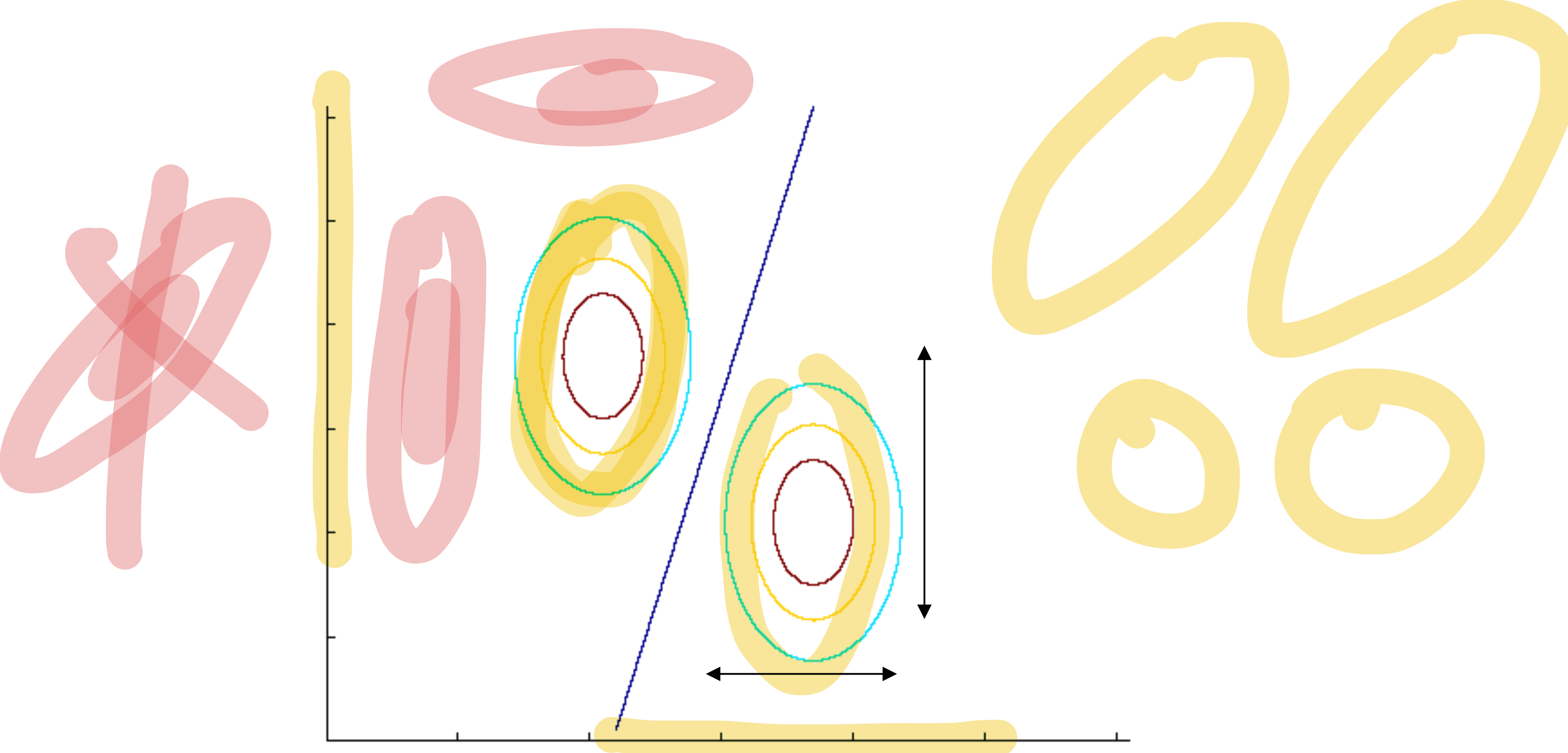
- When x_j ($j = 1,..d$) are independent (or assumed to be independent for simplicity), then Σ is diagonal.

This is the **Naive Bayes classifier** where $p(x_j|C_i)$ are univariate Gaussian.

Case 3) Common Covariance Matrix S which is Diagonal

Diagonal covariance matrices means no correlation among attributes – hence contours are axis-aligned

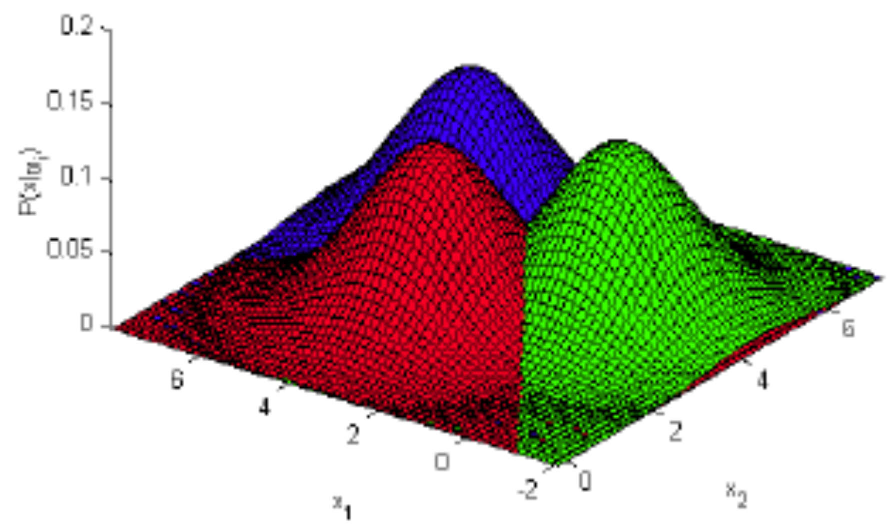
But note that variances in individual dimensions can be different.

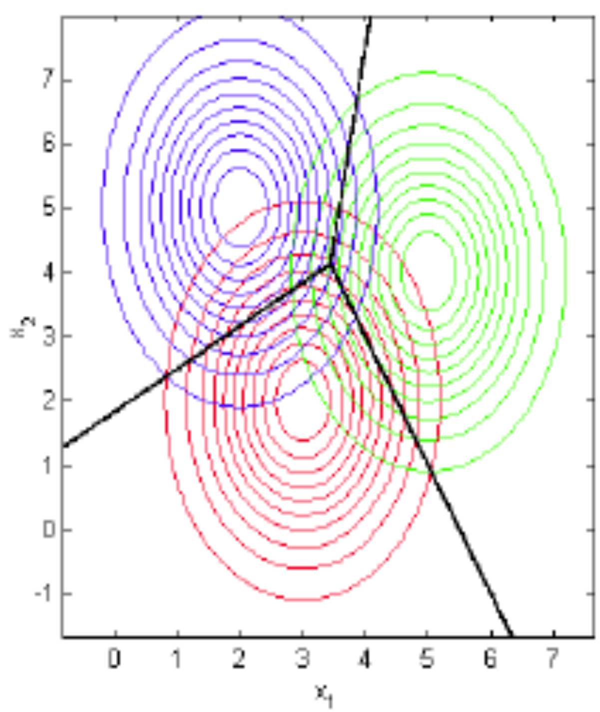


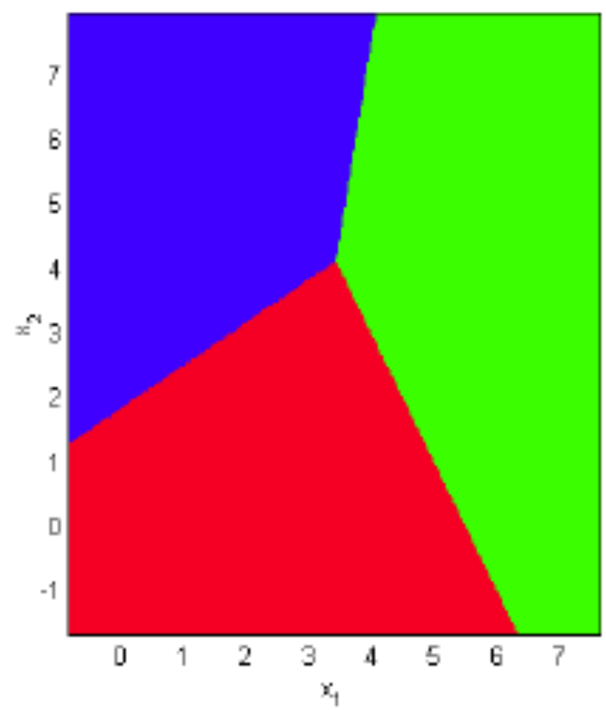
- **To illustrate the previous result, we will compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors**

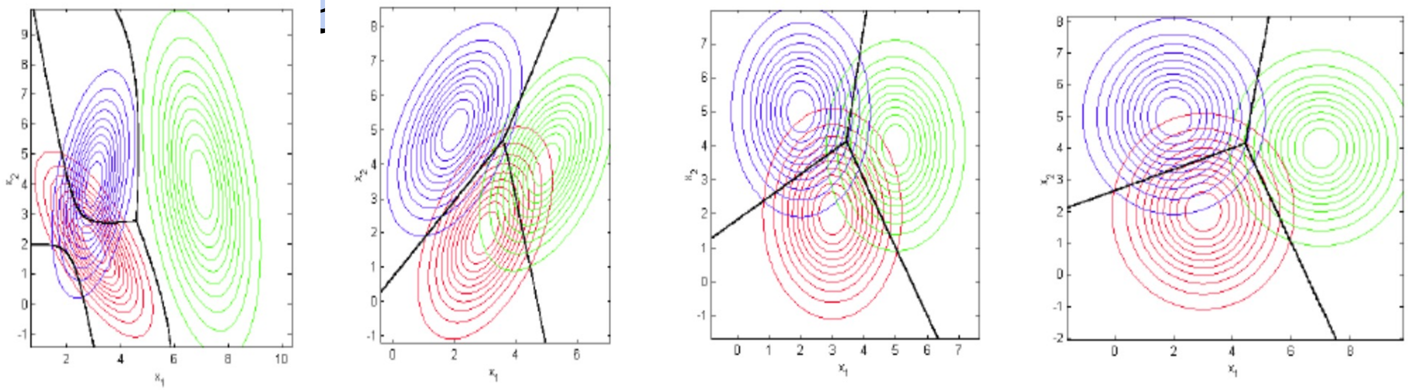
$$\mu_1 = [3 \quad 2]^T \quad \mu_2 = [5 \quad 4]^T \quad \mu_3 = [2 \quad 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$





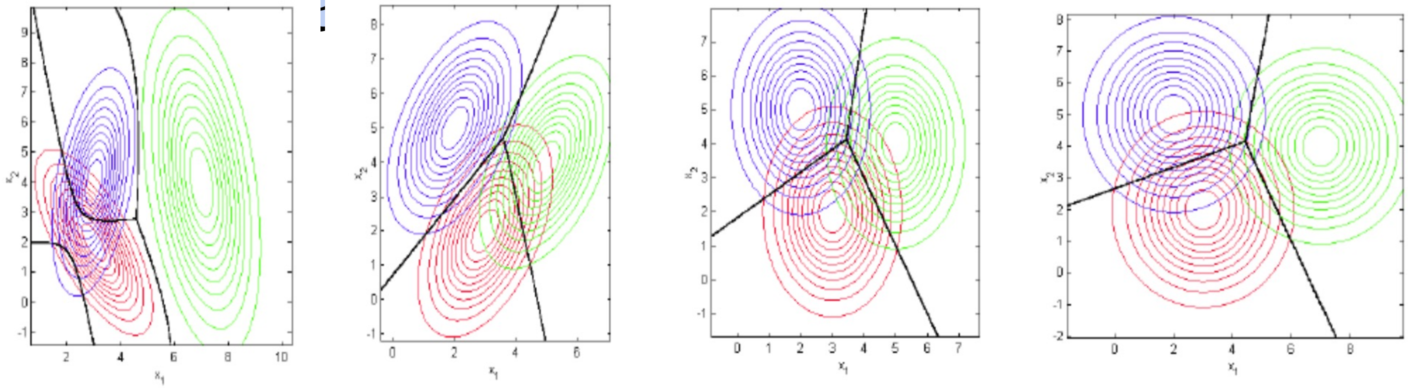




Assumption	*Covariance matrix*	*Number of parameters*
Case 1) None. All different, Hyperellipsoidal	$\mathbf{S}_i$	$K\ d(d+1)/2$
Case 2) Shared, Hyperellipsoidal	$\mathbf{S}_i=\mathbf{S}$	$\cancel{K}.\ d(d+1)/2$
Case 3) Shared & Axis-aligned	$\mathbf{S}_i=\mathbf{S}$, with $s_{ij}=0$	d
Case 4) Shared & Hyperspheric	$\mathbf{S}_i=\mathbf{S}=s^2\mathbf{I}$	1

- As we increase complexity (less restricted **S**), it is more important to have sufficient data to properly estimate the parameters.
- Simpler models may not model the underlying distributions fully correctly, but may even work better.

Mo

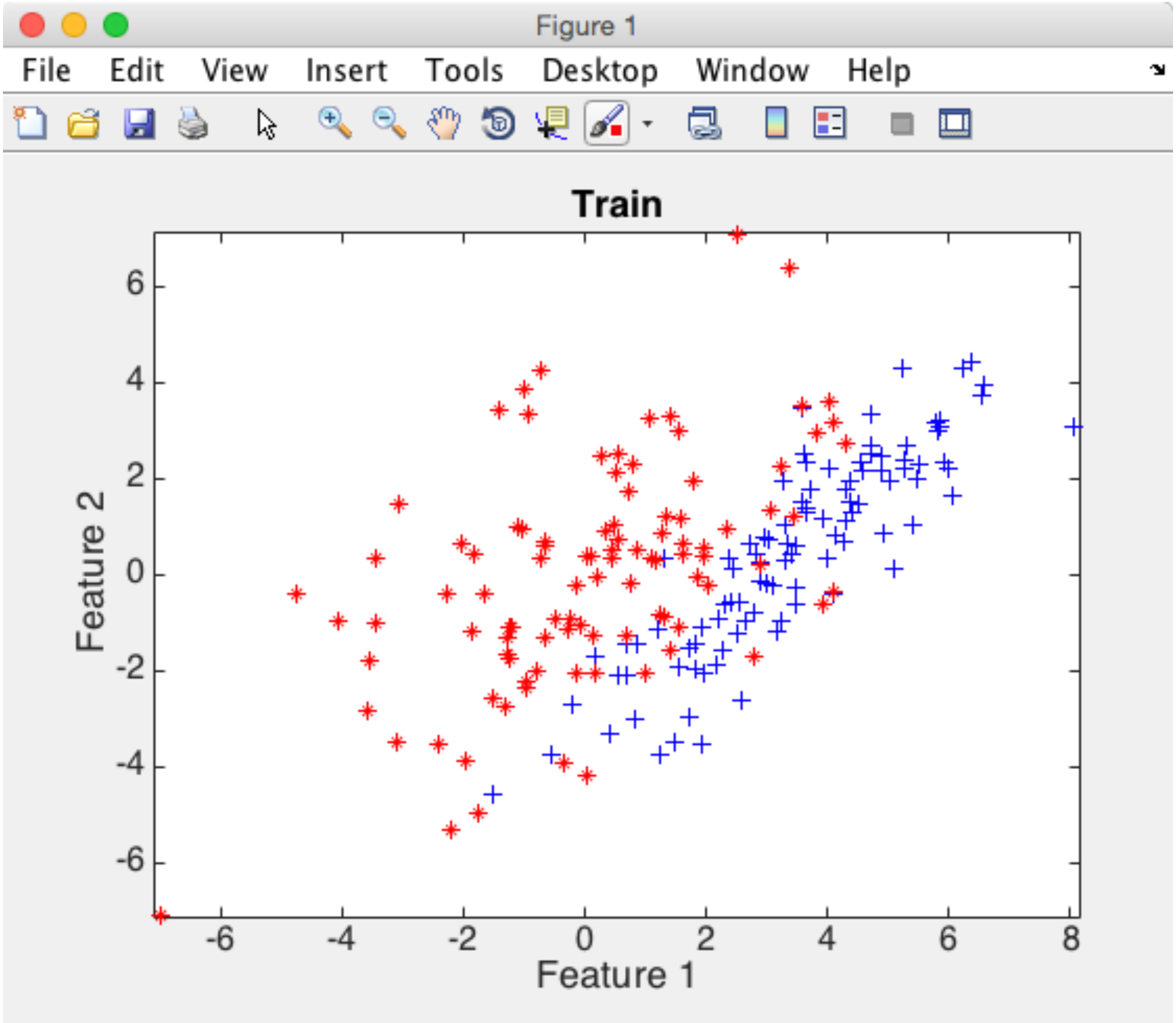


Assumption	*Covariance matrix*	*Number of parameters*
Case 1) None. All different, Hyperellipsoidal	$\mathbf{S}_i$	$K\ d(d+1)/2$
Case 2) Shared, Hyperellipsoidal	$\mathbf{S}_i=\mathbf{S}$	$d(d+1)/2$
Case 3) Shared & Axis-aligned	$\mathbf{S}_i=\mathbf{S}$, with $s_{ij}=0$	d
Case 4) Shared & Hyperspheric	$\mathbf{S}_i=\mathbf{S}=s^2\mathbf{I}$	1

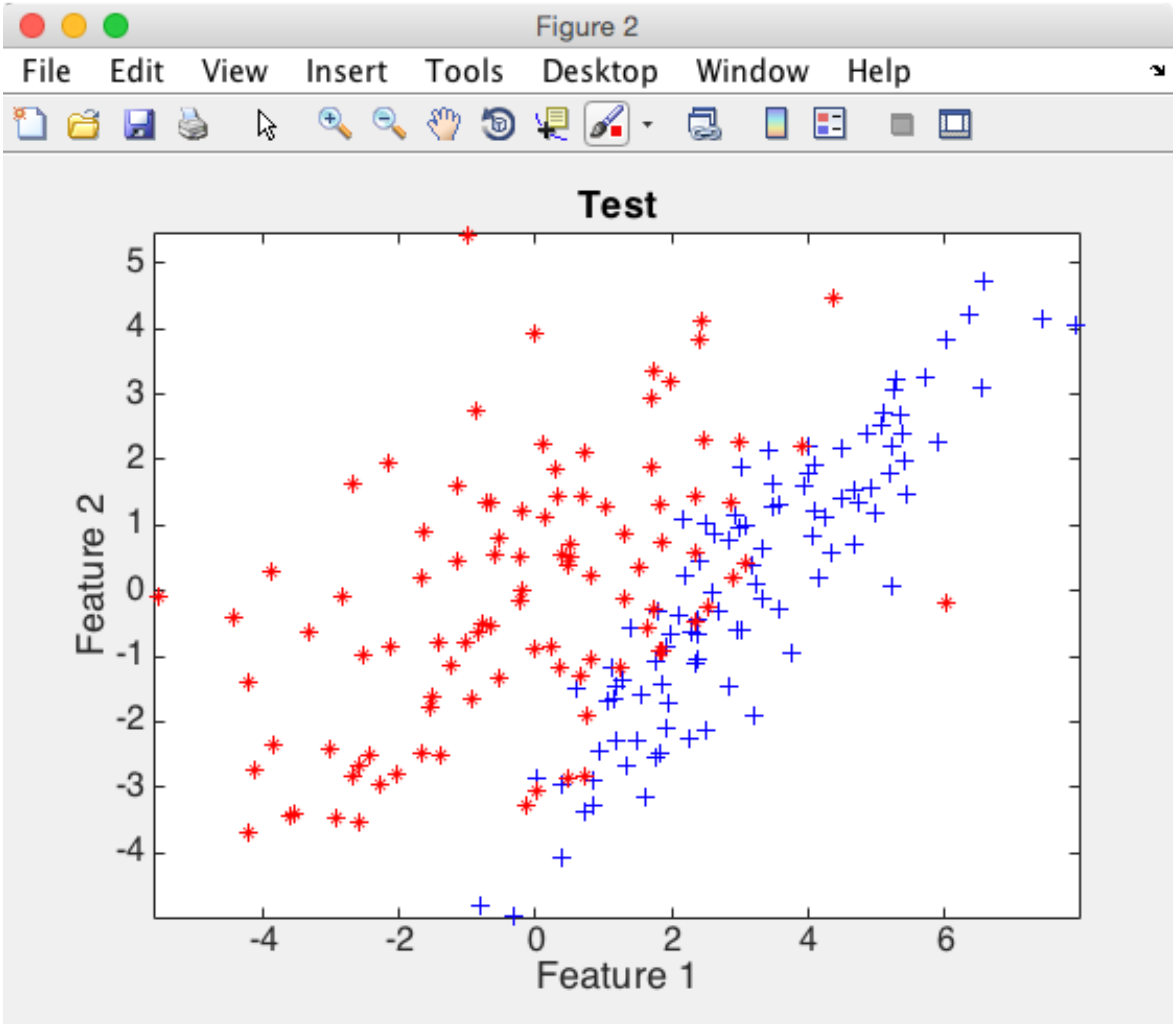
- As we increase complexity (less restricted **S**), bias decreases and variance increases
- Assume simple models (allow some bias) to control variance (regularization)

- QDA (short for Quadratic Bayes classifier or Quadratic Discriminant Analysis) and
- LDA (short for Linear Bayes classifier or Linear Discriminant Analysis) are the two Gaussian Bayes classifiers;….
    - First one corresponds to the general covariance matrix case and second one to the shared covariance matrix case
    - See http://scikit-learn.org/stable/modules/lda_qda.html Section 1.2.2.

Matlab Exercise: Generate some training data

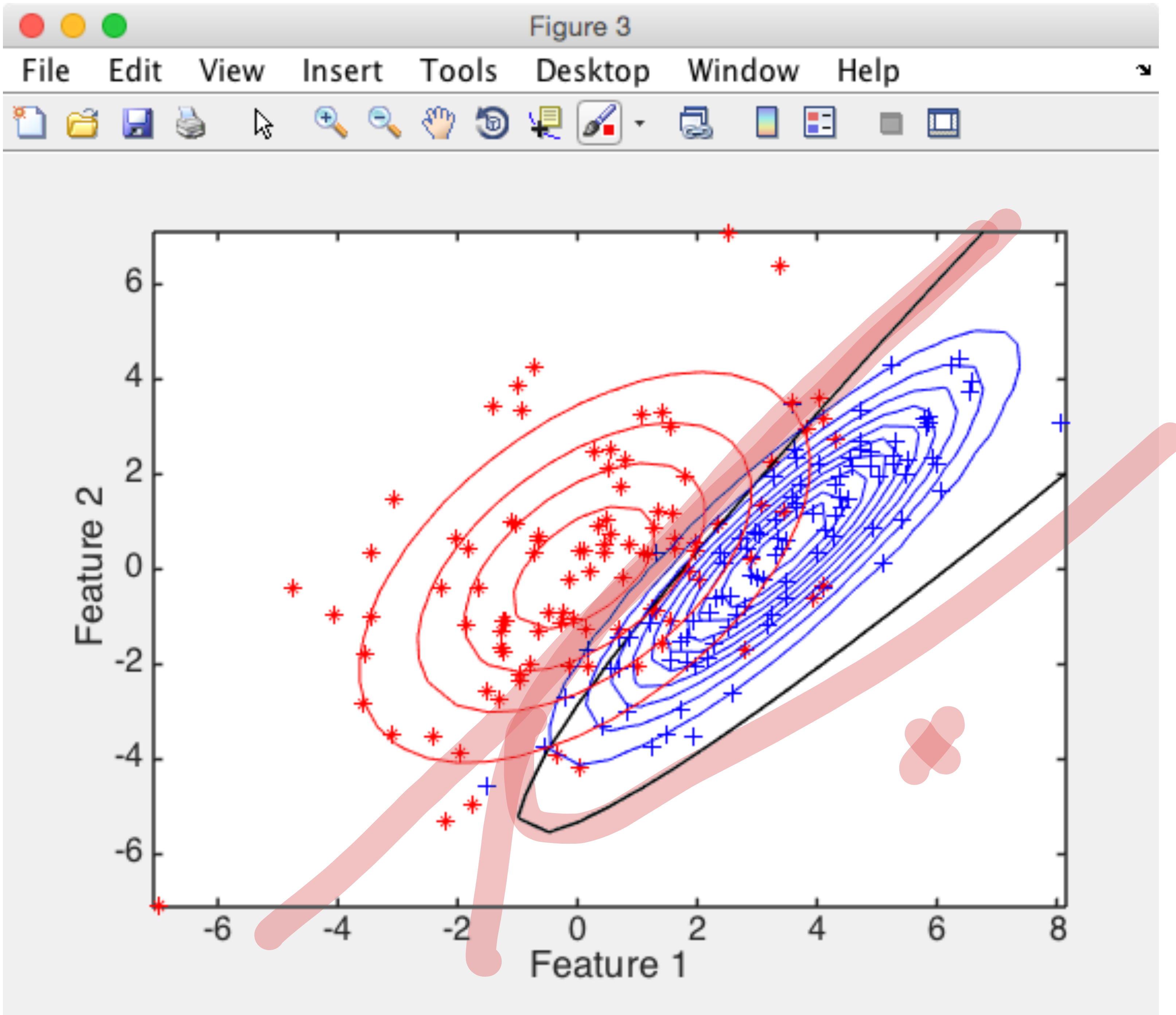


Matlab Exercise: Generate some test data



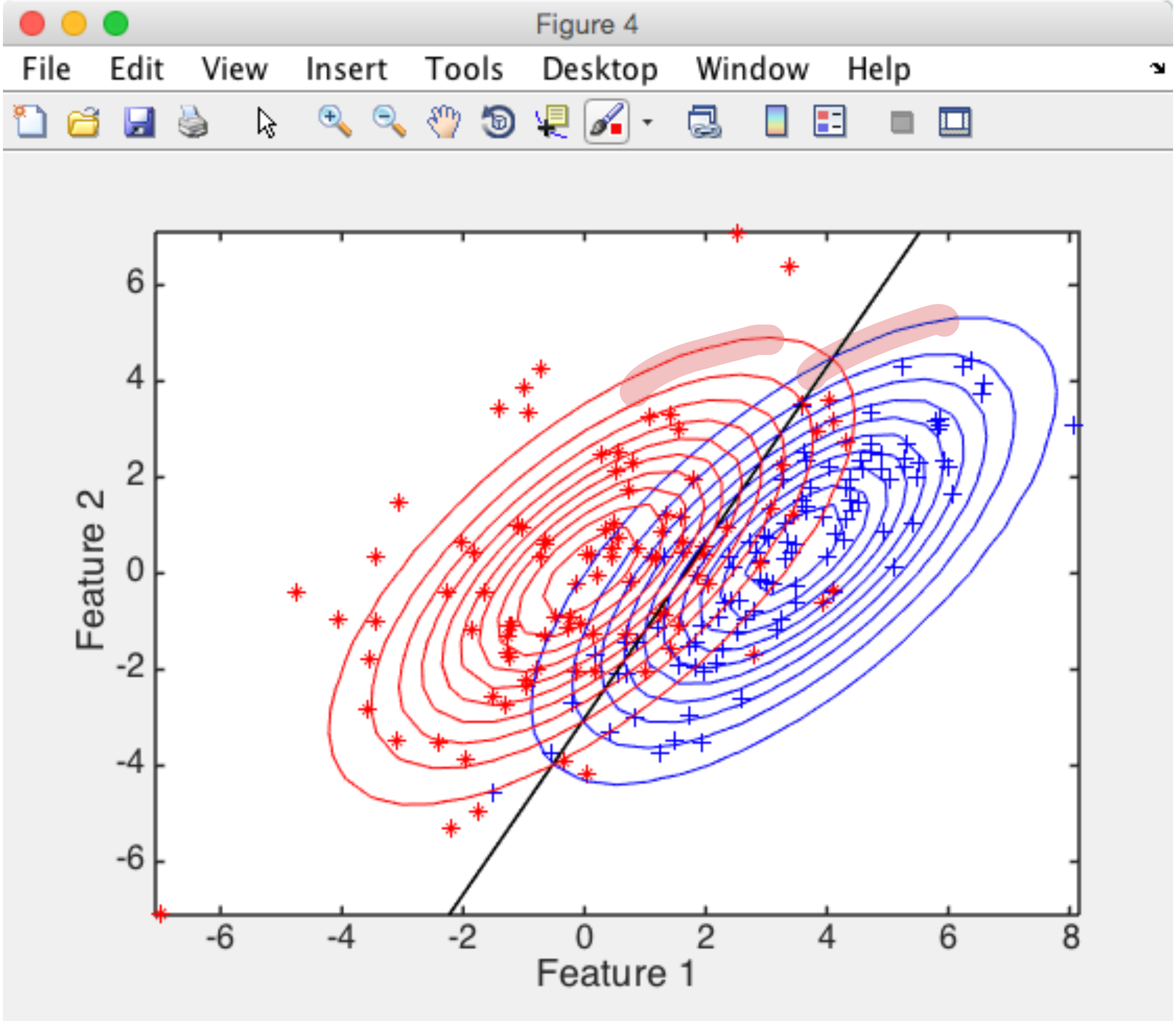
Quadratic classifier

error on test: 9.5%

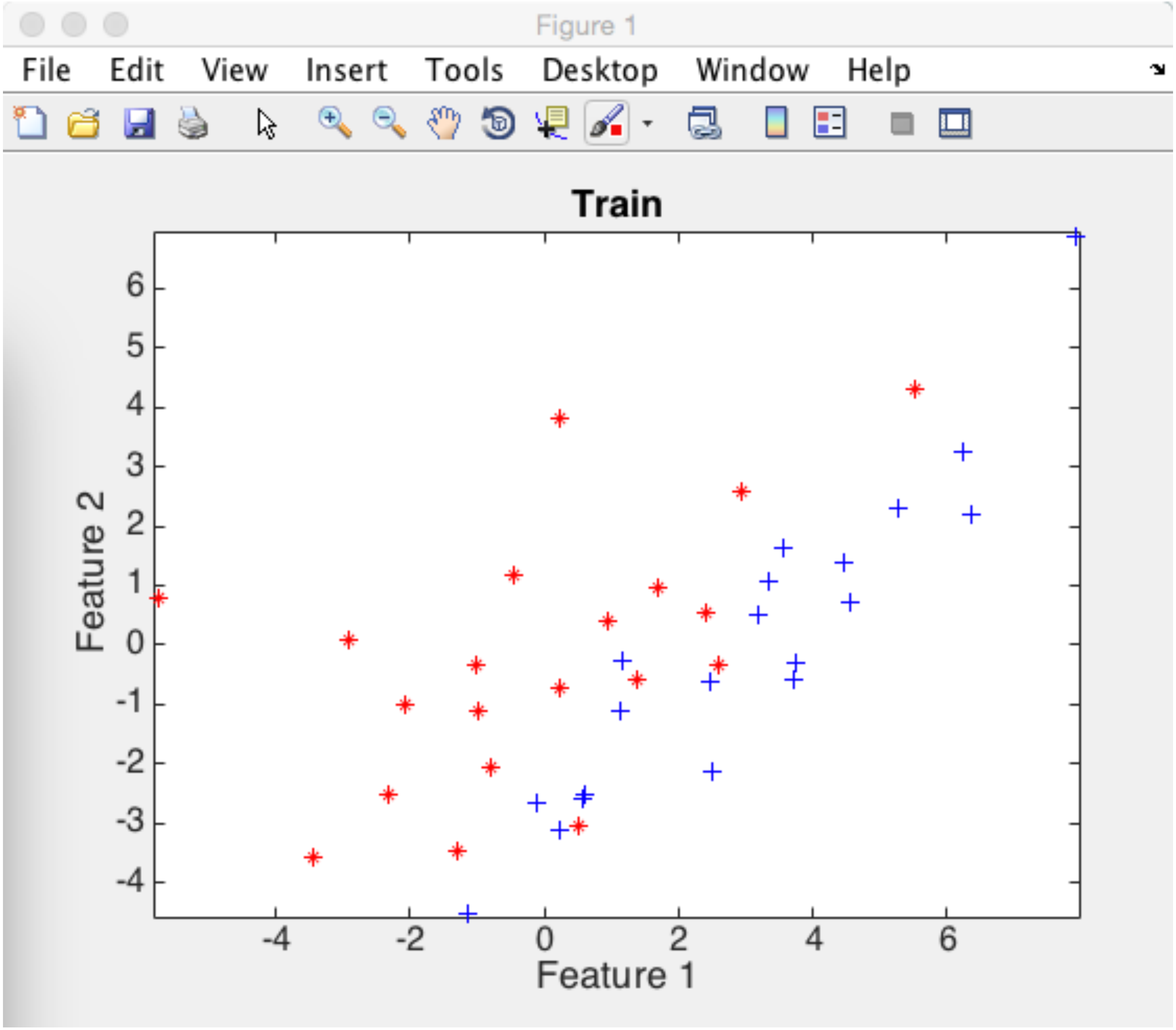


Linear classifier

error on test: 11.5%



Another example with N=20 points



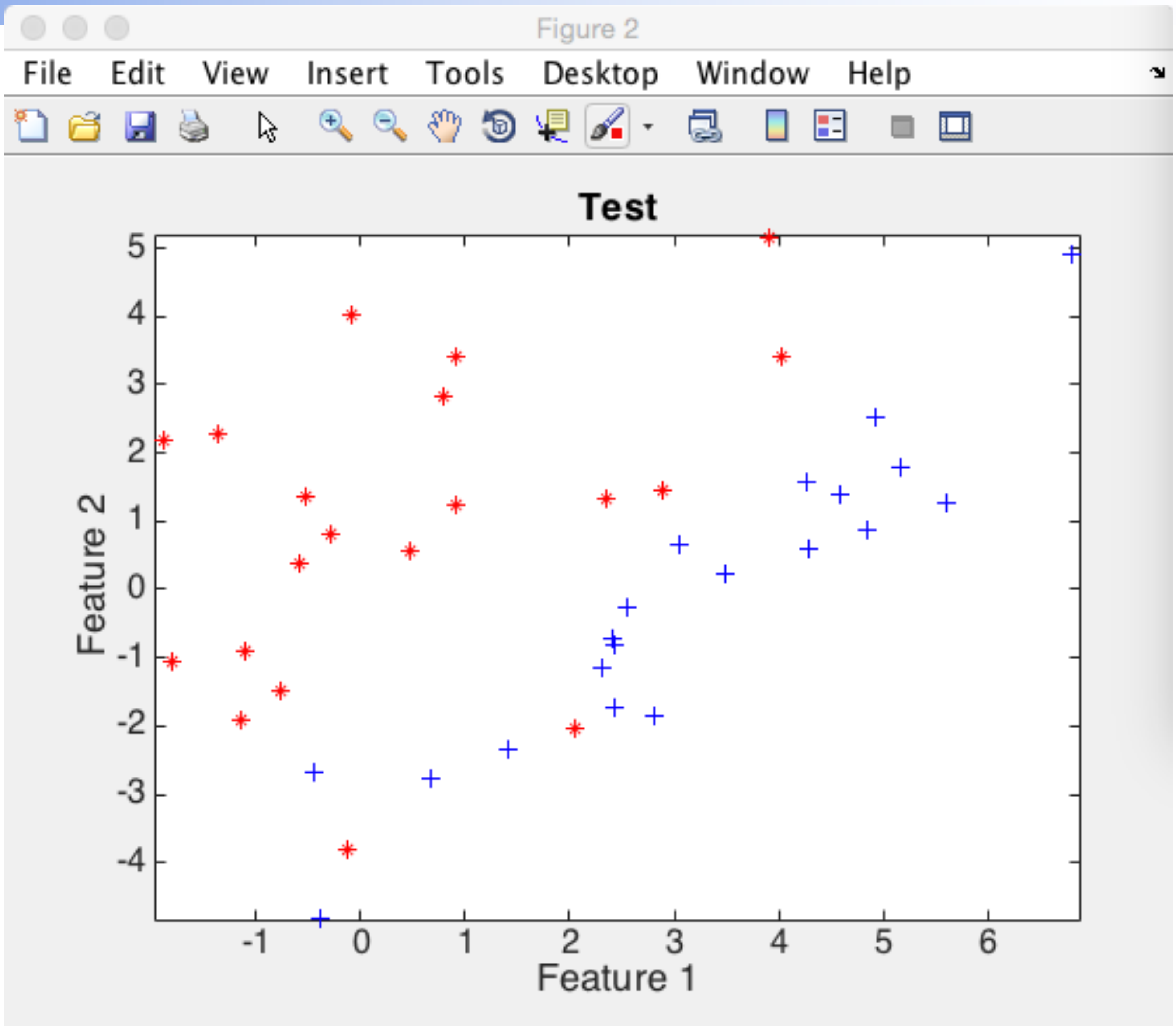


Figure 2
File Edit View Insert Tools Desktop Window Help
Test
Feature 2
Feature 1
5
4
3
2
1
0
-1
-2
-3
-4
-1
0
1
2
3
4
5
6

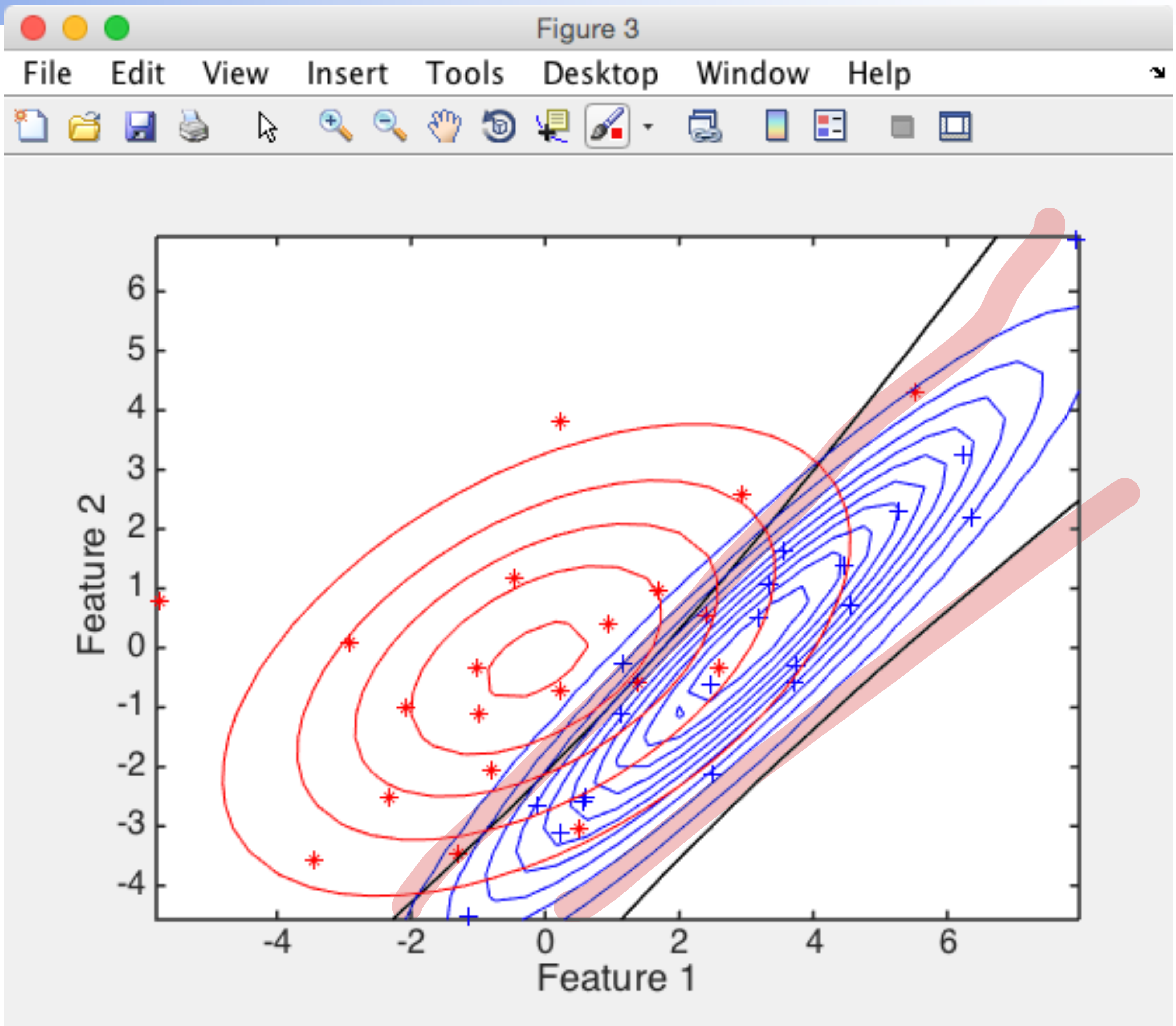


Figure 3
File Edit View Insert Tools Desktop Window Help
Feature 1
Feature 2
-4 -2 0 2 4 6
6 5 4 3 2 1 0 -1 -2 -3 -4

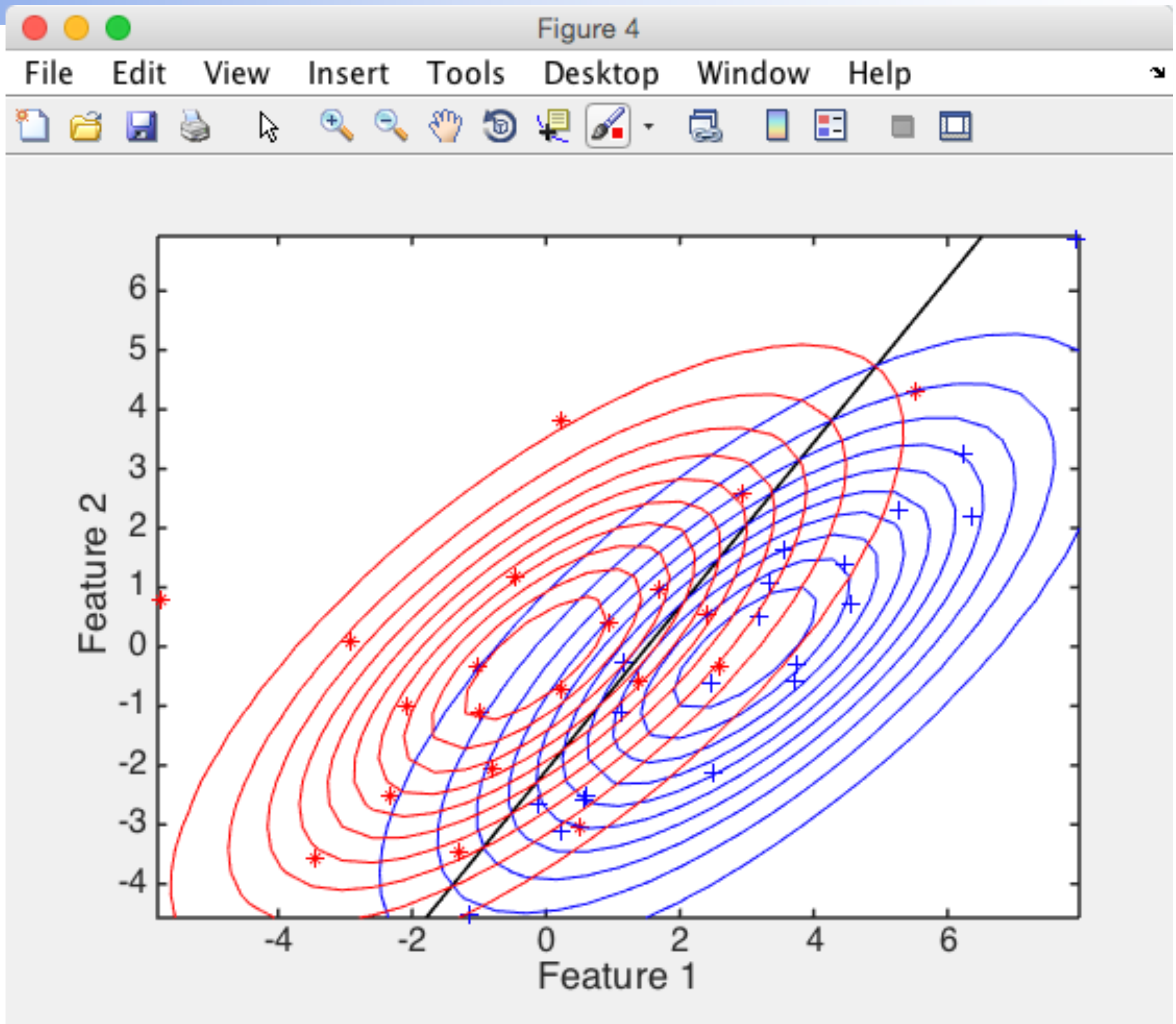


Figure 4
File Edit View Insert Tools Desktop Window Help
Feature 2
Feature 1
6
5
4
3
2
1
0
-1
-2
-3
-4
-4
-2
0
2
4
6

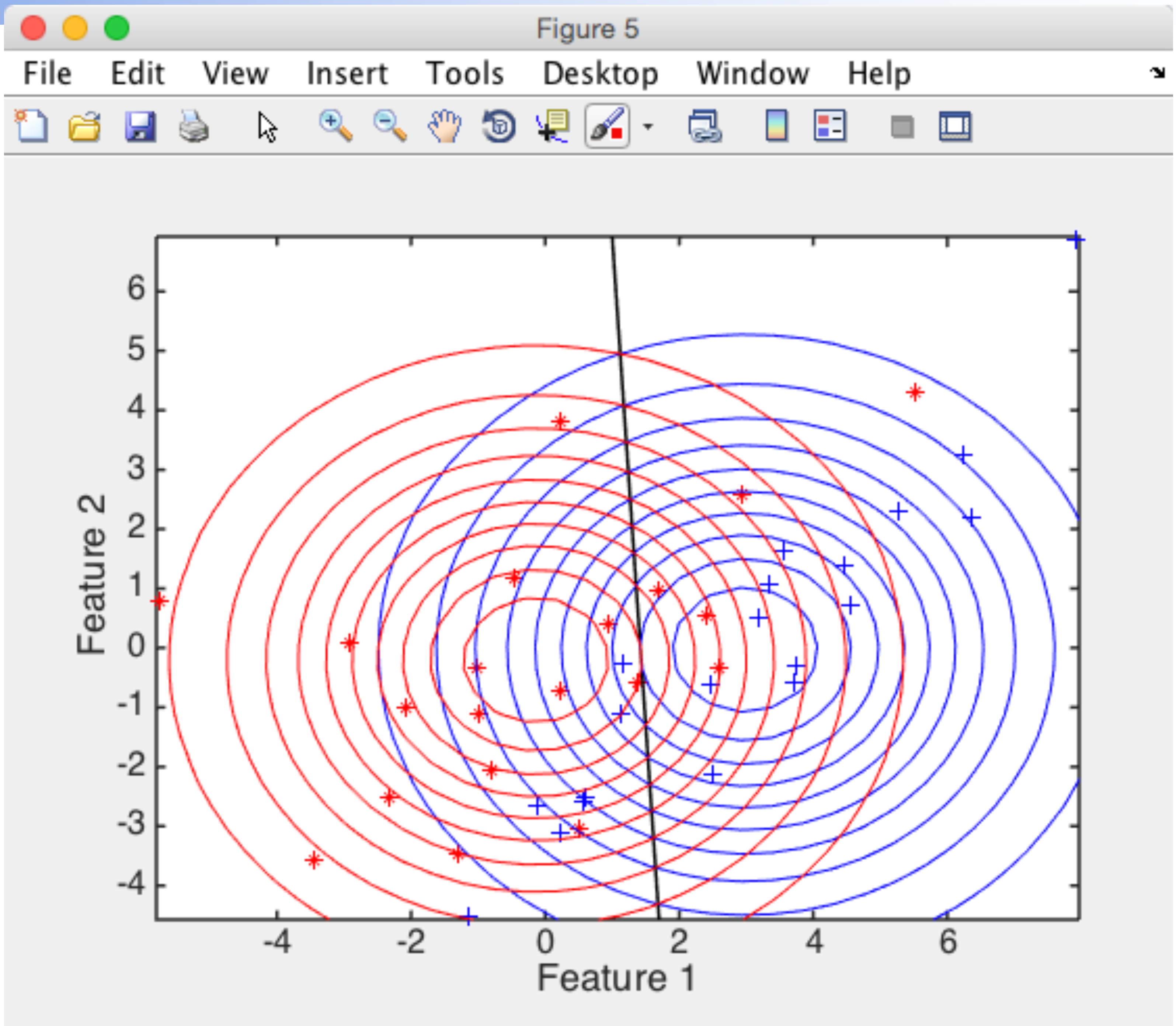


Figure 5
File Edit View Insert Tools Desktop Window Help
Feature 2
Feature 1
6
5
4
3
2
1
0
-1
-2
-3
-4
-4
-2
0
2
4
6

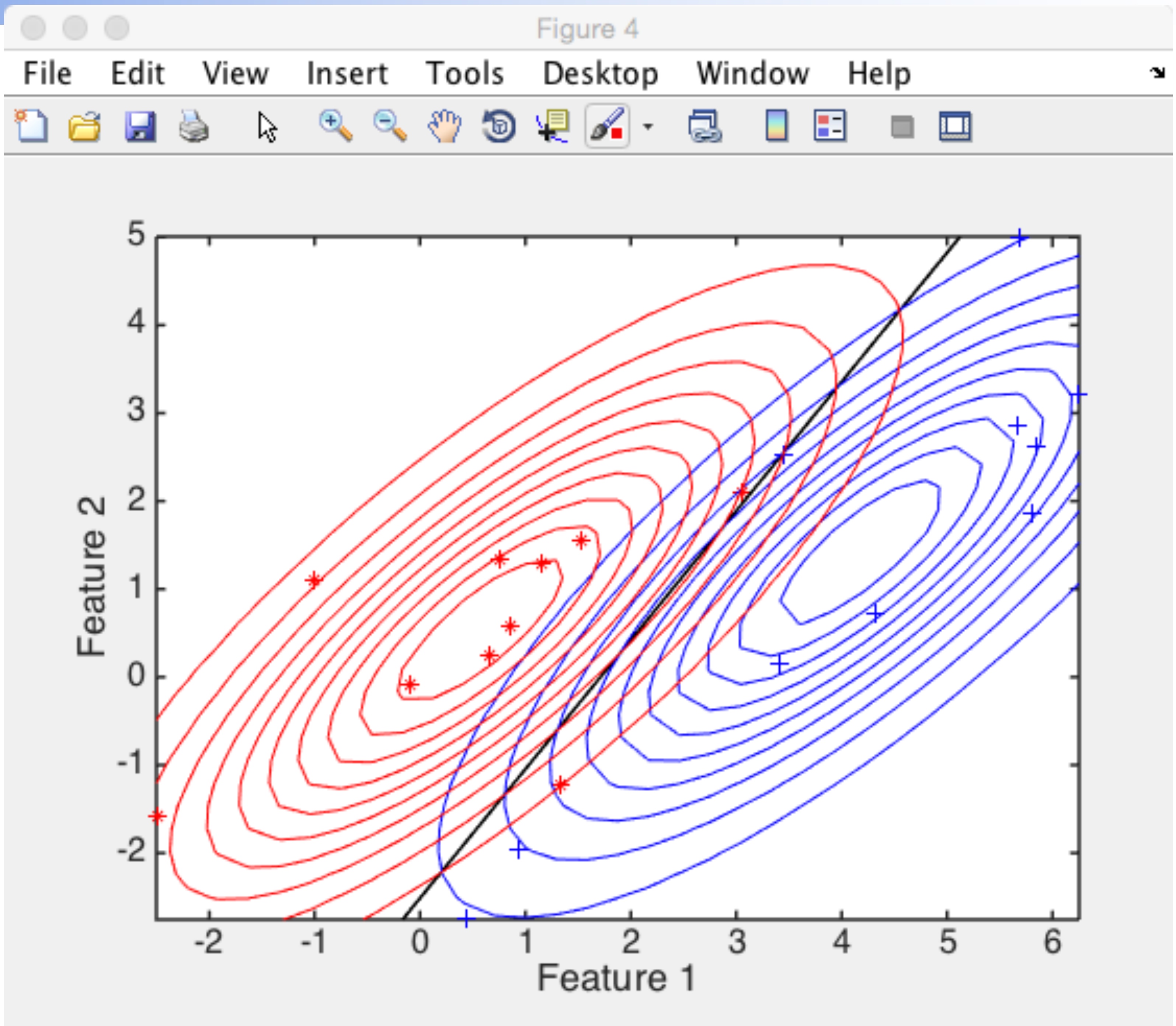


Figure 4
File Edit View Insert Tools Desktop Window Help
Feature 1
Feature 2
-2 -1 0 1 2 3 4 5 6
5 4 3 2 1 0 -1 -2