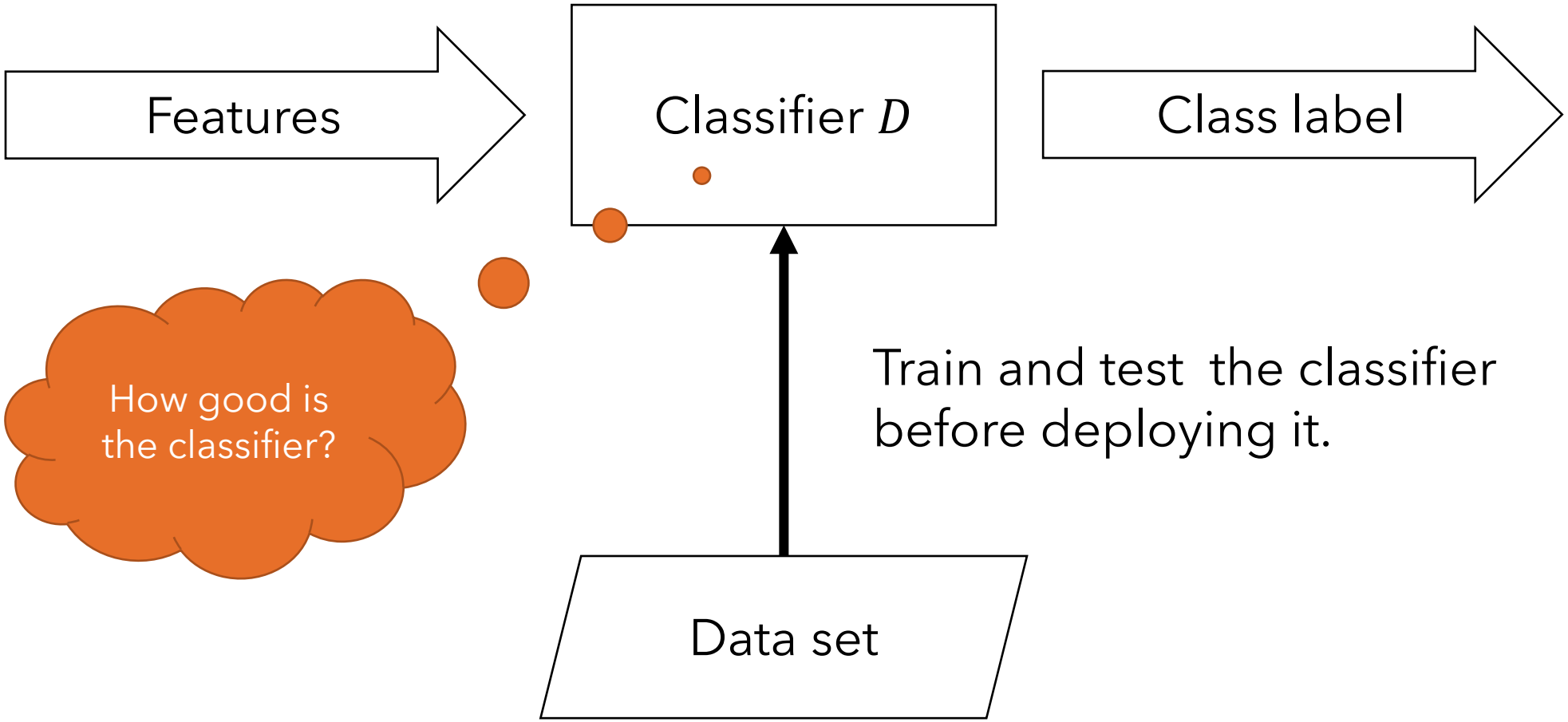
CSE 435 Pattern Recognition

Confusion matrix

Tanvir Azhar
Lecturer, EDU

Confusion matrix

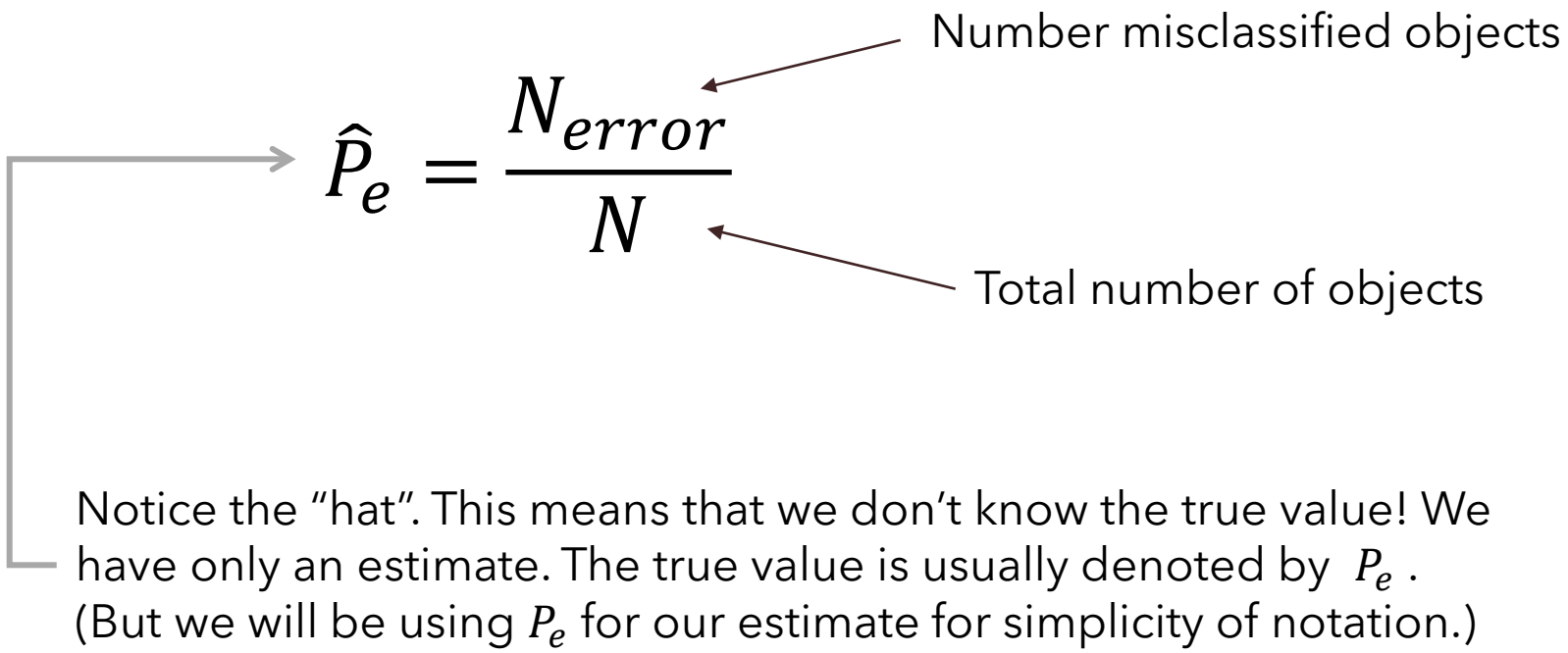
- Classification error
- What is the confusion matrix?
- Examples


Features
Classifier D
Class label
How good is the classifier?
Train and test the classifier before deploying it.
Data set

How do we know how good classifier D is?

We can estimate **the probability of error** of D on unseen data.

Calculate this probability using a data set with N objects:



The probability of error P_e

= error rate of the classifier

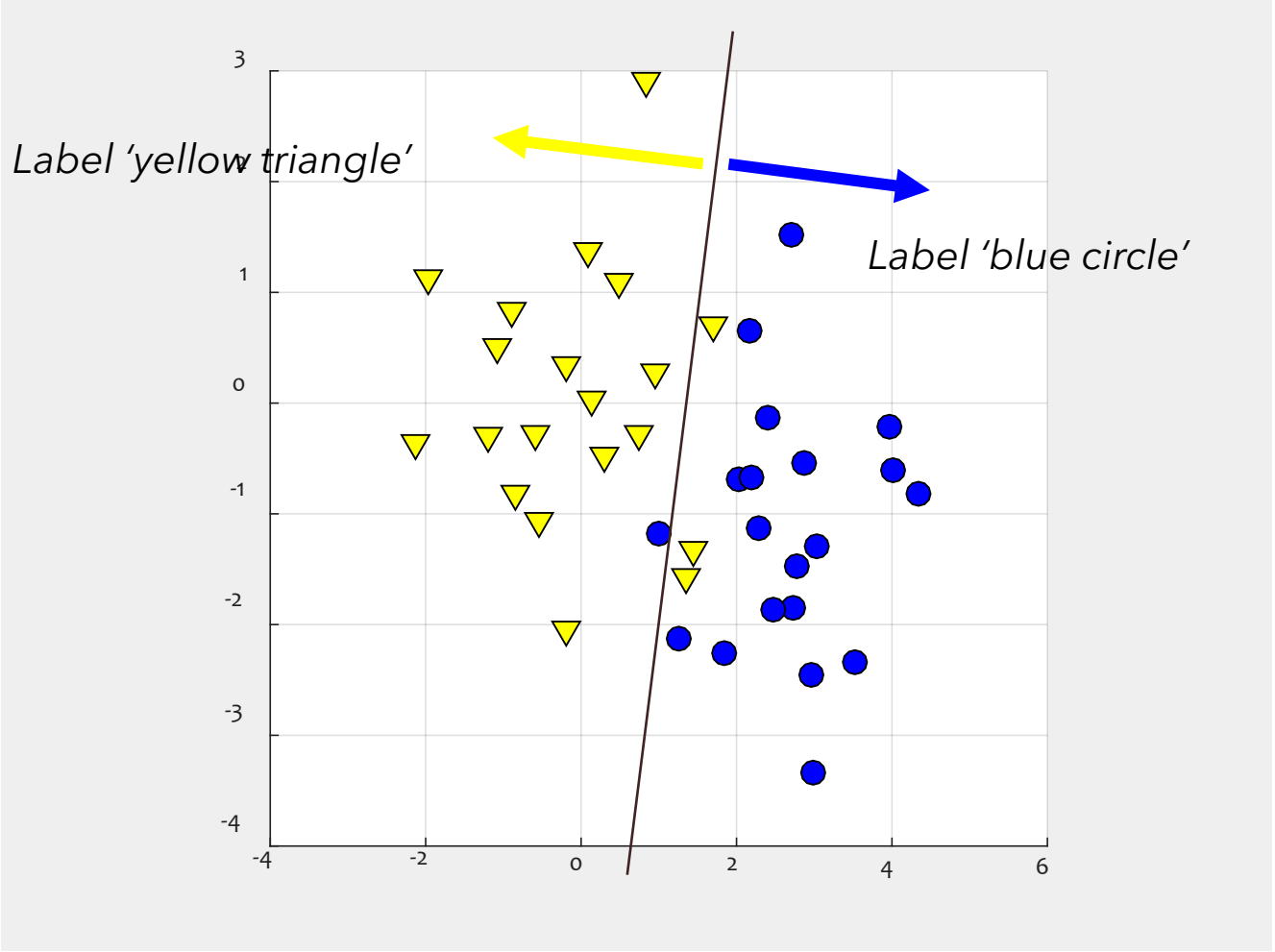
= error of the classifier

The probability of correct classification P_c

= accuracy of the classifier

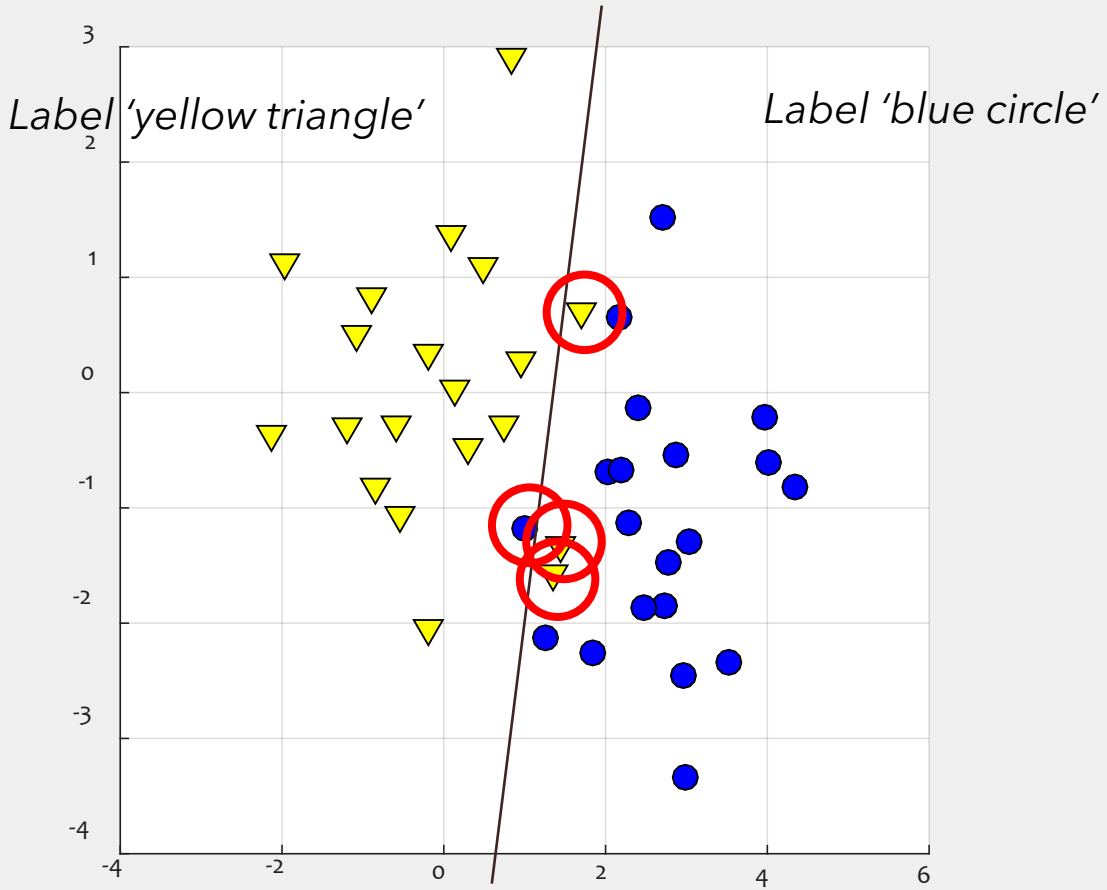
$$P_c = 1 - P_e = \frac{N_{correct}}{N} = 1 - \frac{N_{error}}{N}$$

Classifier D_1 defined by the linear discriminant function



How many errors does this classifier commit for the given data set?

There are 40 objects in total. What is the estimated error rate of this classifier?



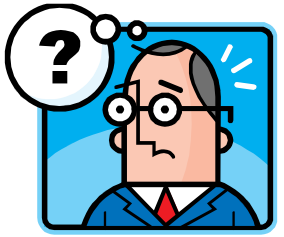
How many errors does this classifier commit for the given data set?

4

There are 40 objects in total. What is the estimated error rate of this classifier?

Error rate = 4/40 = 0.1

Confusion matrix

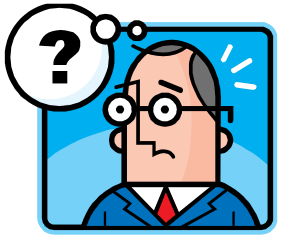


Guessed labels (assigned by D)

True labels	ω_1	ω_j	ω_c
ω_1	a_{11}		
ω_i	…	a_{ij}	
ω_c			

The number of elements of the testing set whose true label is ω_i and D has assigned them to ω_j

Confusion matrix



Guessed labels (assigned by *D*)

True labels	ω_1	ω_j	ω_c
ω_1	☑	☒	☒
	☒	☑	☒
ω_i	☒	☒	☑
	☒	☒	☒
ω_c	☒	☒	☒

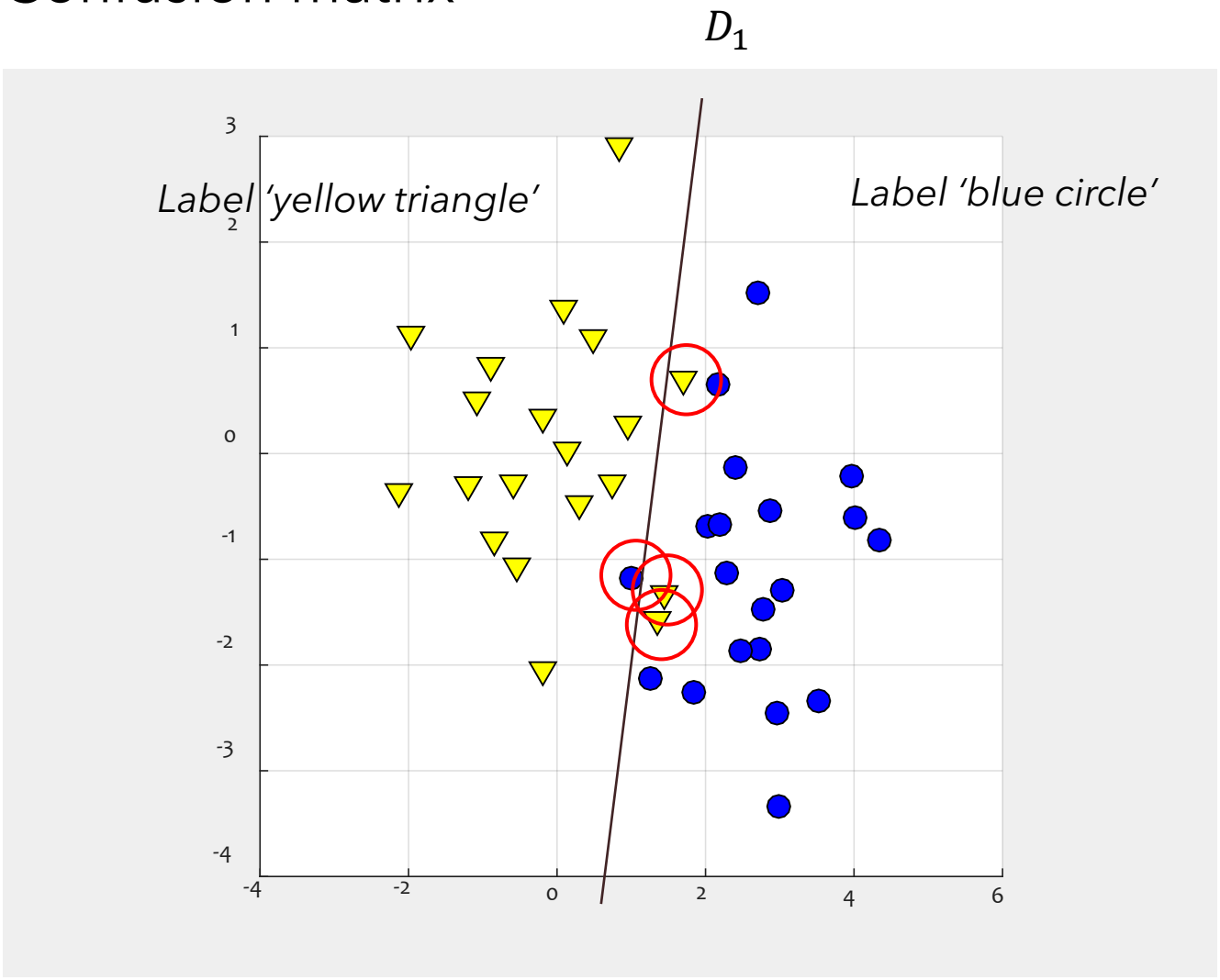
Correct

(diagonal elements)

Wrong

(off-diagonal elements)

Confusion matrix



c = 2 classes

n = 2 features

N = 40 objects

(20 from each class)

Assigned

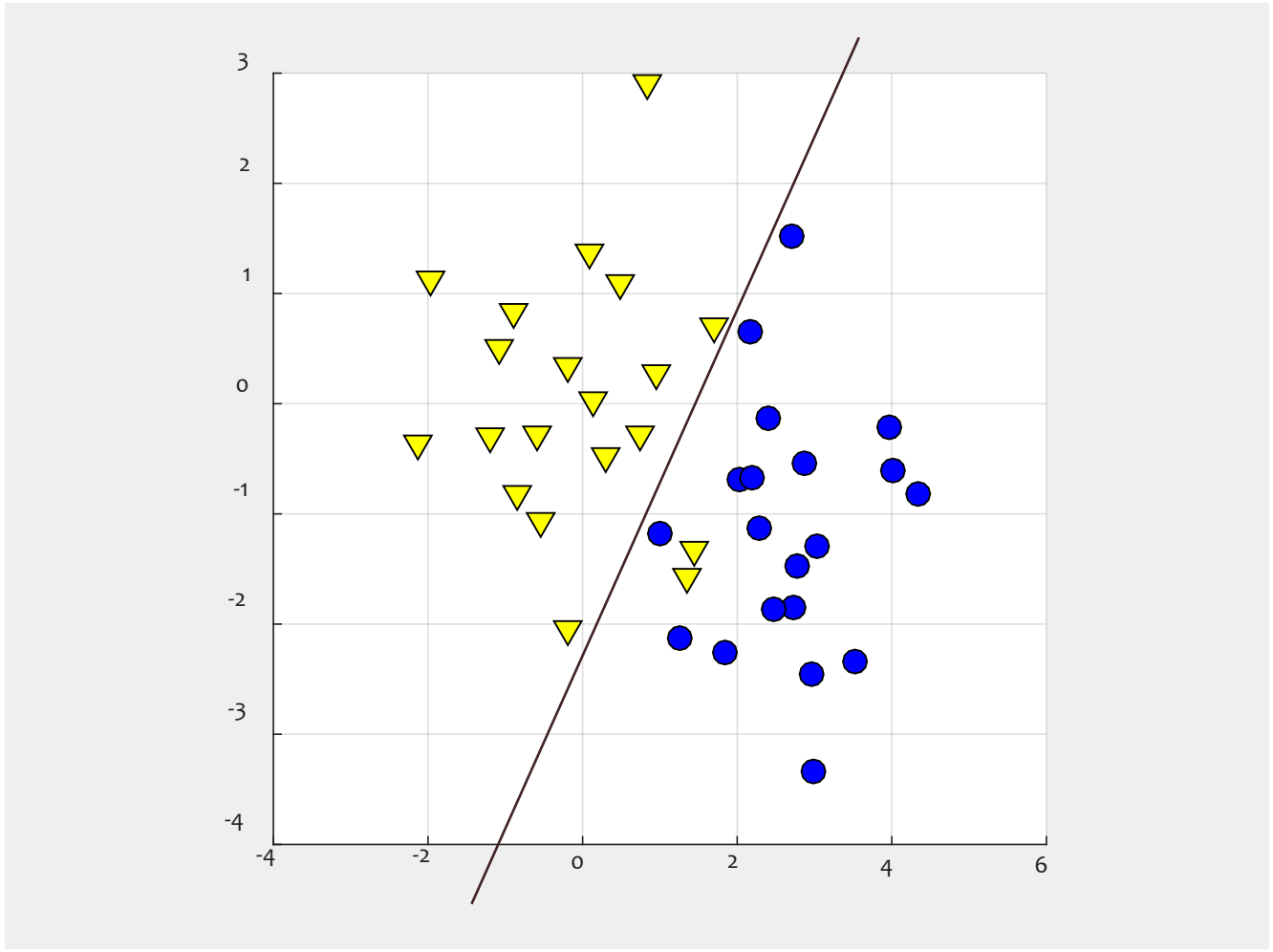
True

D_1	Yellow	Blue
Yellow	17	3
Blue	1	19

P_D = 4/40 = 0.10
(probability of error)

Accuracy = (17+19)/40 = 0.90

Confusion matrix



c = 2 classes

n = 2 features

N = 40 objects

(20 from each class)

Assigned

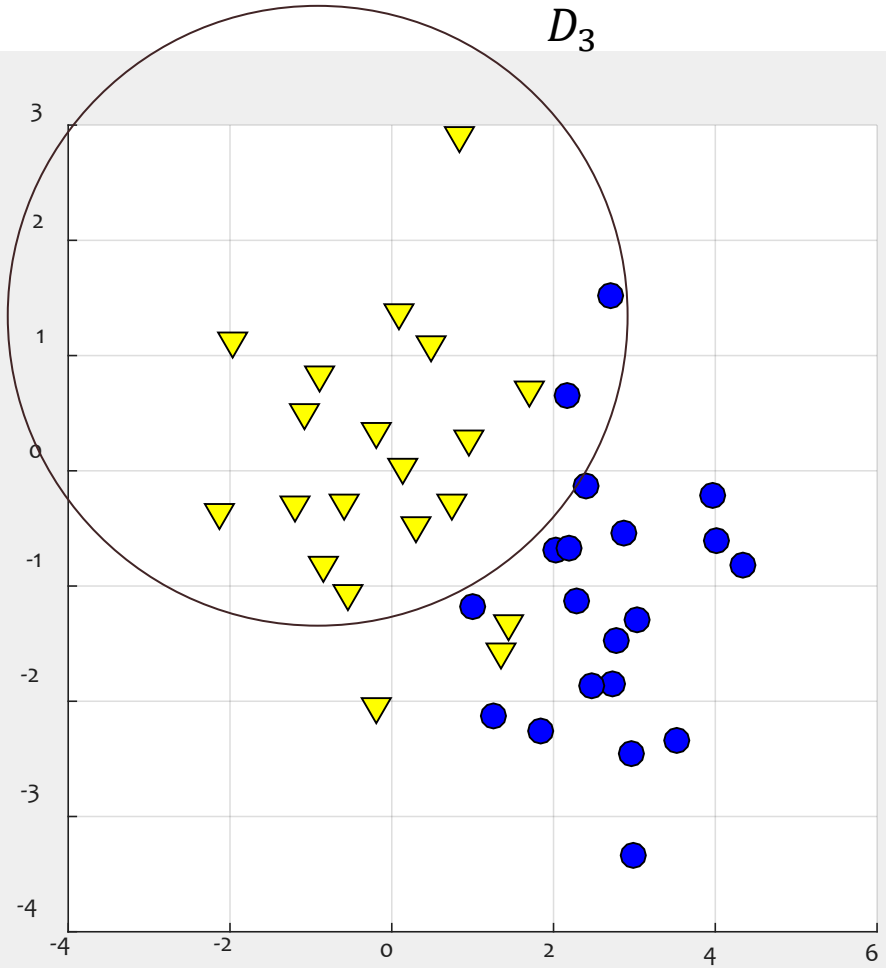
True	D_2	Yellow	Blue
	Yellow	18	2
	Blue	0	20

P_D = 2/40 = 0.05

(probability of error)

Accuracy = (18+20)/40 = 0.95

Confusion matrix



c = 2 classes

n = 2 features

N = 40 objects

(20 from each class)

Assigned

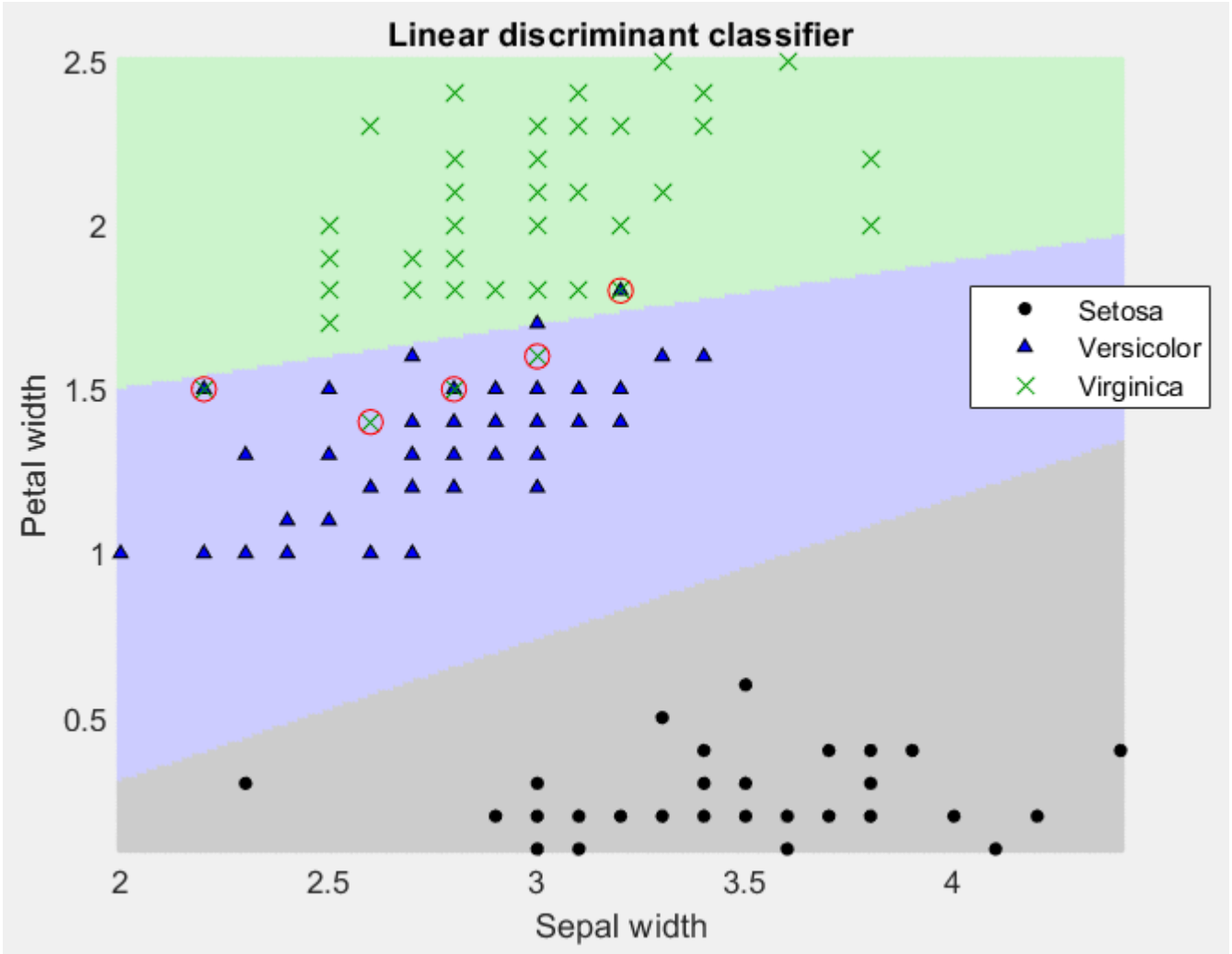
True	D_3	Yellow	Blue
	Yellow	17	3
	Blue	2	18

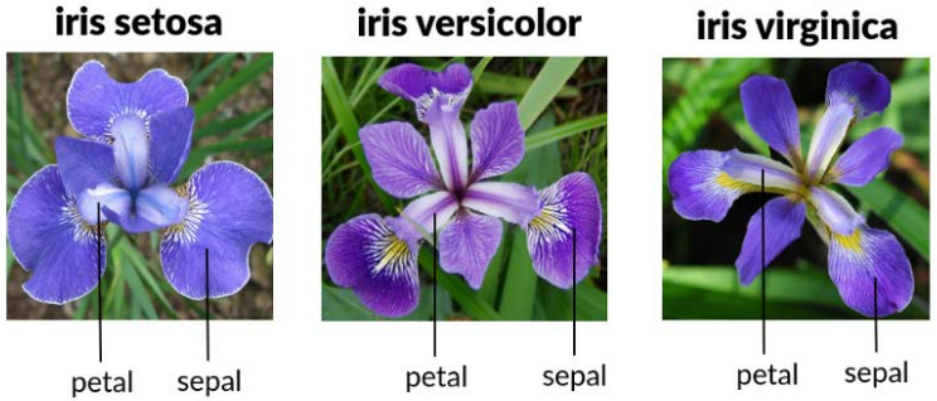
P_D = 5/40 = 0.125
(probability of error)

Accuracy = (17+18)/40 = 0.875

N = 150, 50 from each class

https://medium.com/analytics-vidhya/exploratory-data-analysis-iris-flower-dataset-a21c368a1f4





	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	49	1
Virginica	0	4	46

Confusion matrix for a classifier D tested on data set Z

Assigned

True	1	2	3
1	10	1	2
2	0	8	2
3	2	2	9

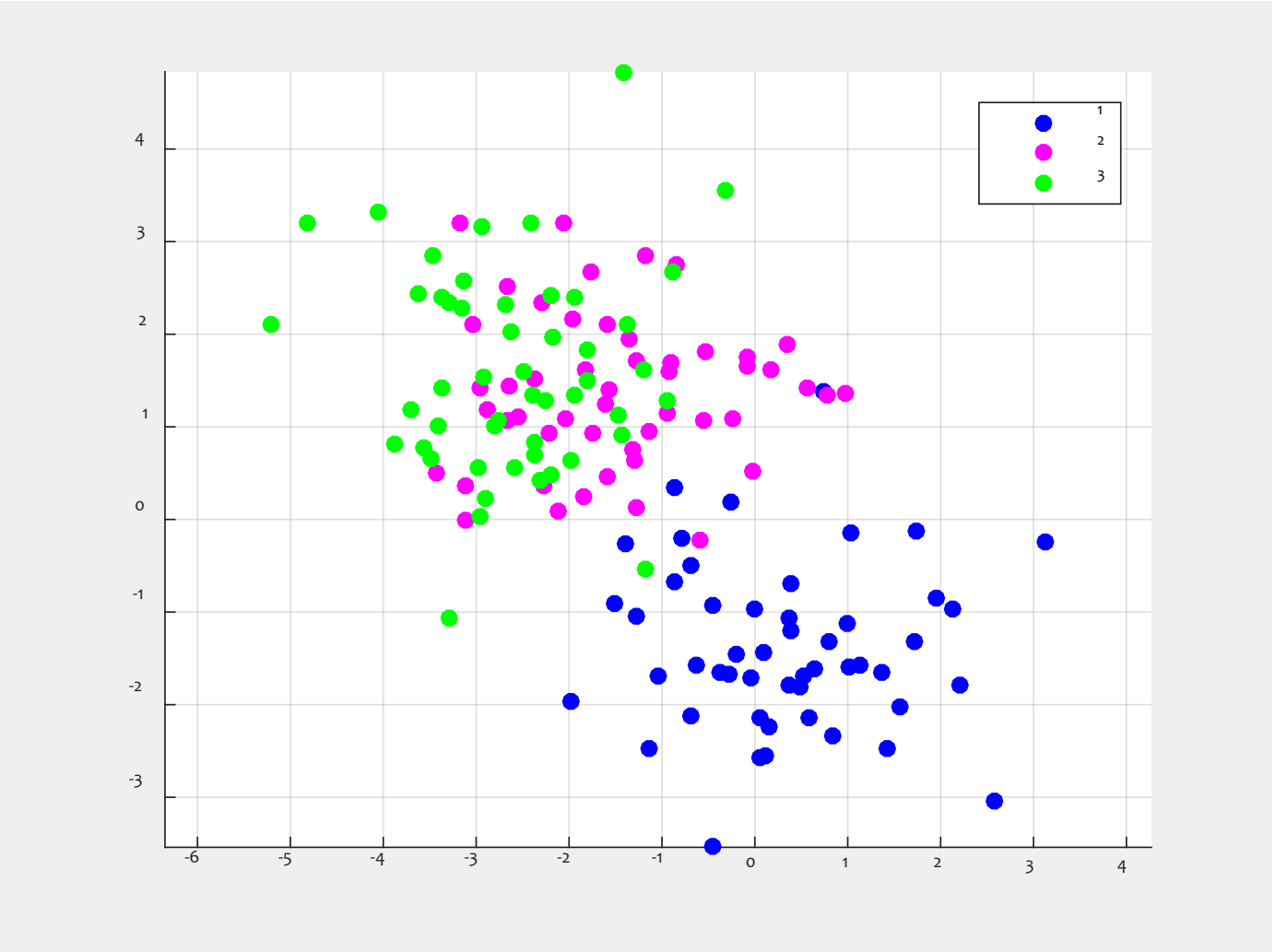
Q1. How many points are there in Z? (N =?)

Q2. How many points from Z have label Class 2?

Q3. Estimate the accuracy of the classifier.

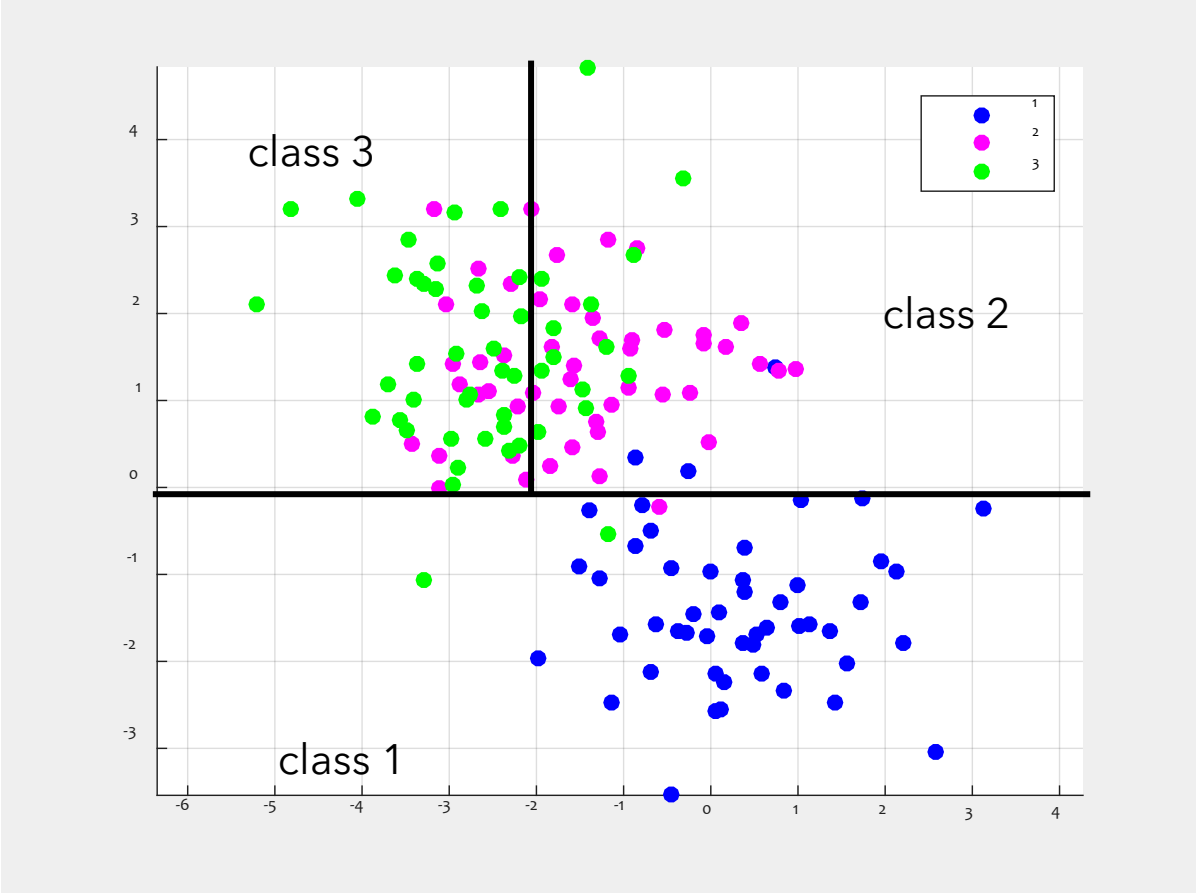
Q4. If the classifier predicts class label 2, what is the probability that this label is correct?

Q5. Write Python code to create three Gaussian classes (see Lab 1), two of which are highly overlapping, and the third is away from both as in the figure below:



Q6. Based on the way you picked the class means, program a rule-based classifier which splits the 2D feature space as shown below:

Note that the exact positions of the horizontal and the vertical lines depend on YOUR chosen class means



Q7. Write a Python function called `confusion_matrix` which will take two input arguments: the true labels and the assigned labels, and will output the confusion matrix.

Q8. Subsequently, apply your function to obtain the confusion matrix for your data and classifier. Calculate the accuracy and the error rate of your classifier.

Q9. Generate a second random data set from the same distribution as in Problem 5. Classify the data according to your rule-based classifier. Calculate the confusion matrix, the accuracy and the error of your classifier with the new data set.

Answers to some questions:

Q1. Add all elements. $N = |Z| = 36$

Q2. Sum up row # 2 (true labels = 2). $0 + 8 + 2 = 10$

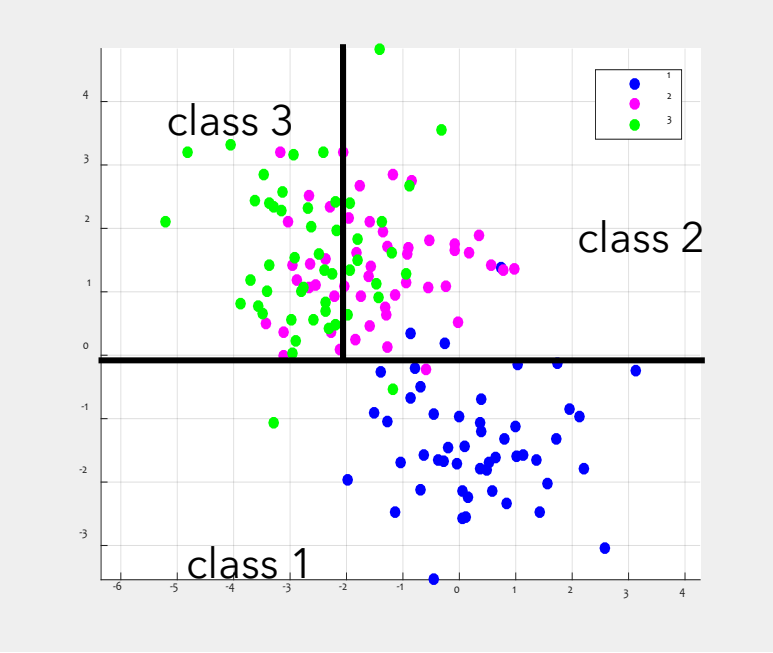
Q3. Sum up the diagonal entries and divide by N.

$$P_{acc} = \frac{10+8+9}{36} = \frac{27}{36} = 0.75$$

Q4. Conditional probability. The condition is "assigned label is 2". Total outcomes favourable for this event is the sum of column #2: $1 + 8 + 2 = 11$. Then the probability of being correct is number of correctly recognised objects from class 2 divided by this total:

$$P(Correct|Assigned\ label\ is\ 2) = \frac{8}{11} = 0.73$$

Q6. In any programming language:



$$If\ x_2 < -0.1\ then\ class\ 1$$
$$Else\ if\ x_1 \geq -2\ then\ class\ 2$$
$$Else, class\ 3$$