Foundations of data science - Solutions to comprehension questions on classification with k-nearest neighbours

QUESTION 2

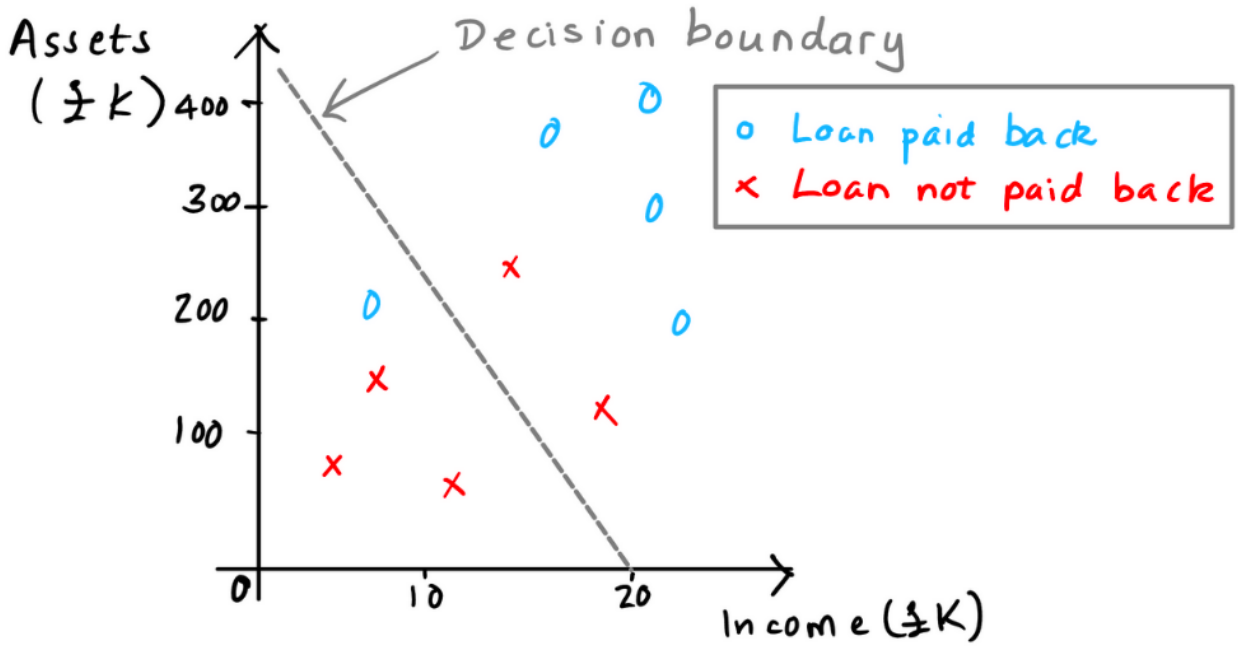
k-nearest neighbours assigns data points to *k* clusters

- ○ True
- ○ False

False: K-means assigns data points to K clusters

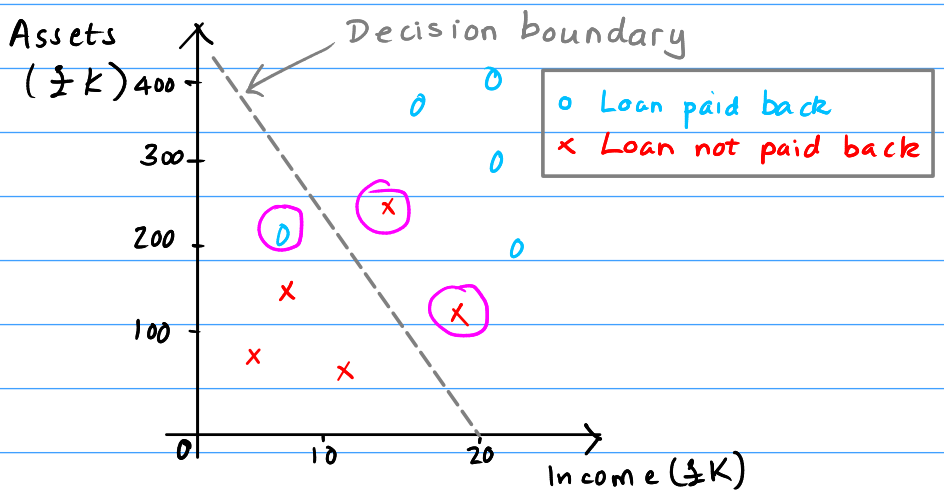
QUESTION 3

This and the next question relate to the figure below, which represents imaginary data collected by a bank. Each data point comprises of the income and assets of a number of previous borrowers, at the time they took out loans with the bank. The points are labelled with the classes "Loan paid back" and "Loan not paid back". The bank has used data to train a classifier with a linear decsion boundary. Any points above the decision boundary will be classified as "Loan paid back" and those below will be classified as "Loan not paid back".



How many of the training data points are missclassified?

3 training data points are misclassified, circled below:



QUESTION 4

What is the training error rate of the classifier in the figure above, as a percentage? (Don't type the percentage sign).

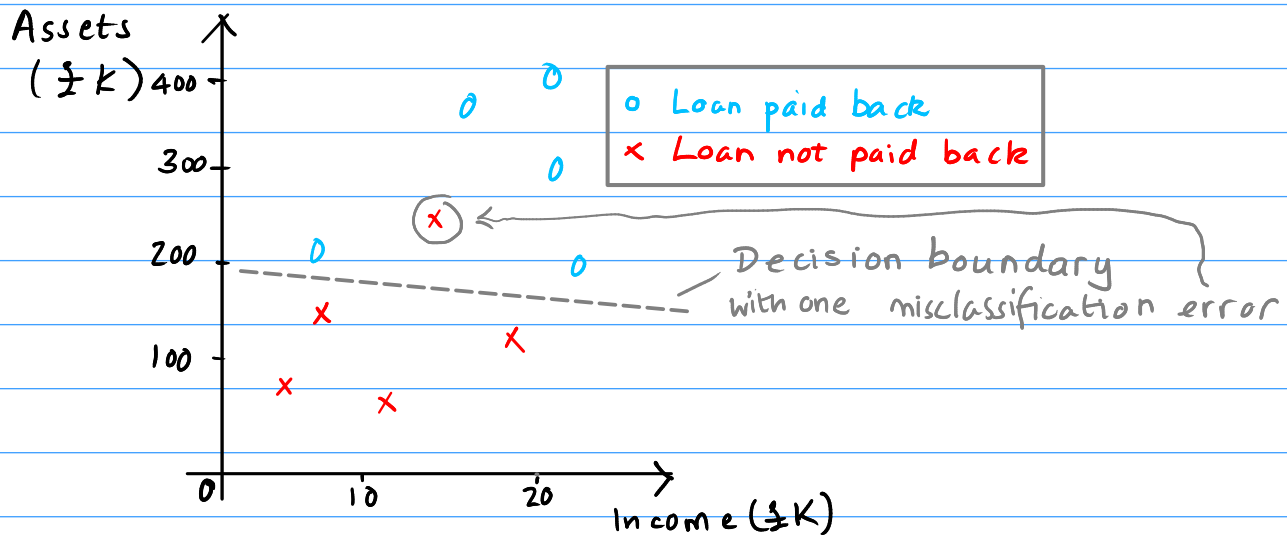
$$\text{Training error rate} = \frac{\#\text{ training points misclassified}}{\#\text{ training points}}$$

$$= \frac{3}{10}$$

$$= 30\%$$

QUESTION 5

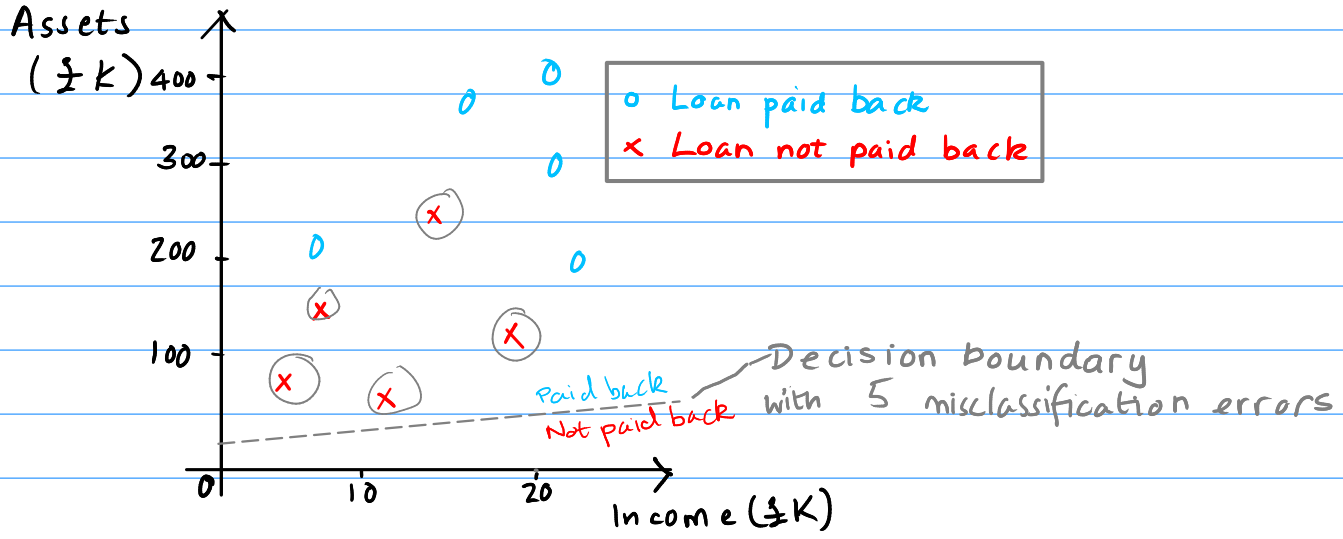
Suppose now we try to construct a better classifier with a straight-line decision boundary. What is the number of misclassifications made by the best possible classifier?



By inspection, the best straight-line decision boundary has only one misclassification error. (The diagram shows one such boundary. There is at least one more.)

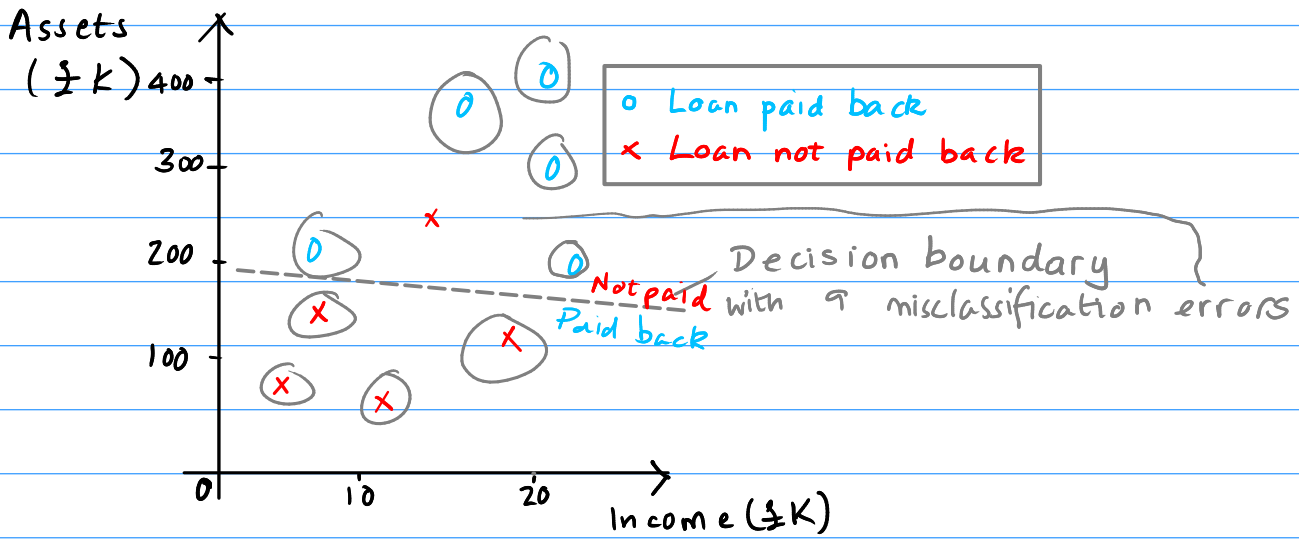
QUESTION 6

What is the highest number of misclassification errors on the training set for any classifier with a linear decision boundary?



The answer we were looking for in the comprehension

quiz was 5, as in the picture above. But actually we could have a linear decision boundary that has 9 misclassification errors:



We've switched which side of the boundary is "not paid" and which side is "paid back".

QUESTION 7

How many misclassification errors would you expect a 1-NN classifier to make on the training set?

Zero errors. Because each point is nearest to itself.

N.B. The original version of this question was wrong: it said "k-NN classifier". The answer would depend on k.

QUESTION 9

Consdier the following subset of some imaginary loan data.

Customer ID	Income (£10K)	Assets (£10K)	Class
A	2	6	Paid
B	2	2	Not paid
C	5	2	Paid
D	7	4	Paid
E	5	5	Not paid

Suppose a new customer appears who has an income of £30K and assets of £30K. Use Euclidean distance to order the training data points above from closest to furthest from the new customer.

Remember that the Eudidean distance of two points $\underline{u}$ and $\underline{v}$ is $d(\underline{u}, \underline{v}) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2}$

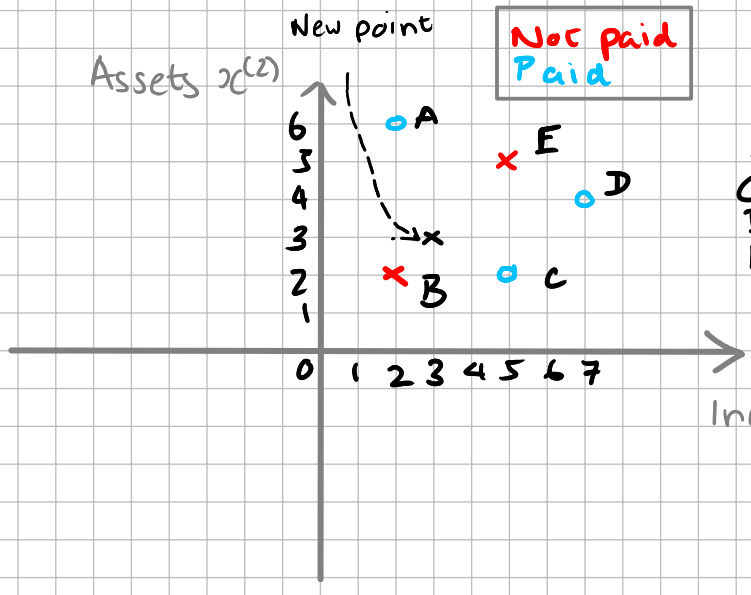
E.g. New point $\underline{u} = (3, 3)^T$

Point A $\underline{v} = (2, 6)$

$\Rightarrow$ Distance from new point to A is:

$d(\underline{u}, \underline{v}) = \sqrt{(3-2)^2 + (3-6)^2} = \sqrt{1^2 + 3^2} = \sqrt{10}$

Repeat for other points



	$x^{(1)}$	$x^{(2)}$	C	d
A	2	6	P	$\sqrt{10}$
B	2	2	N	$\sqrt{2}$
C	5	2	P	$\sqrt{5}$
D	7	4	P	$\sqrt{17}$
E	5	5	N	$\sqrt{8}$

Now order from closest to furthest from the new customer:

	$x^{(1)}$	$x^{(2)}$	C	d
B	2	2	N	$\sqrt{2}$
C	5	2	P	$\sqrt{5}$
E	5	5	N	$\sqrt{8}$
A	2	6	P	$\sqrt{10}$
D	7	4	P	$\sqrt{17}$

QUESTION 10

Suppose we try out *k*-Nearest Neighbours with different values of *k*. Match the value of *k* to the class chosen.

We can use the ordered points to help us compute K-NN with various values of K

	$x^{(1)}$	$x^{(2)}$	C	d	k=1	k=4	k=5
B	2	2	N	$\sqrt{2}$	N	N	N
C	5	2	P	$\sqrt{5}$		P	P
E	5	5	N	$\sqrt{8}$		N	N
A	2	6	P	$\sqrt{10}$		P	P
D	7	4	P	$\sqrt{17}$			P
					1 N ⇒ Majority vote is N	2 N, 2 P ⇒ Vote tied	2N, 3 P ⇒ Majority vote is P.

Hence the classes predicted are

1 NN = N
4 NN = Tie
5 NN = P