Exercise 1.

Supervised

- Identify diseases
- Decide the content of a picture
- Predict stock values, ~~[illegible]~~
- identify sentiments in sentences

Unsupervised

- customer segmentation. Group customers based on their characteristics
- 2D image segmentation. Cluster the colors
- 3D data segmentation. Cluster shapes and 3D masses

Exercise 2.

Binary

- Classify whether someone has a disease
- Outlier detection. Classify wheter something is normal or not

Supervised

- identify the content of a image.
- Iris dataset.

Clustering

- check unsupervised part in previous section
-

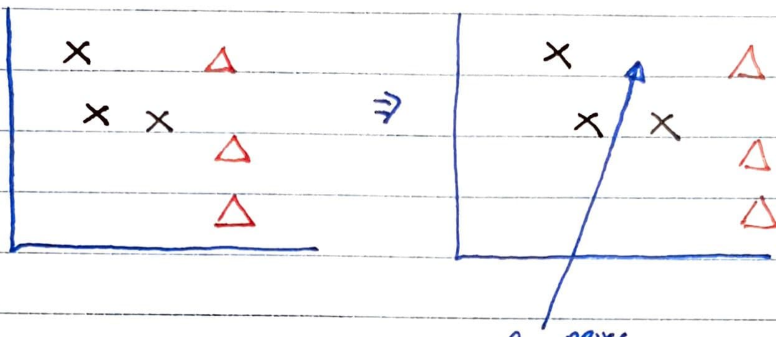
Regression

- Forecast the value of a stock in the future
- Credit scoring.

Machine learning Tutorial notes

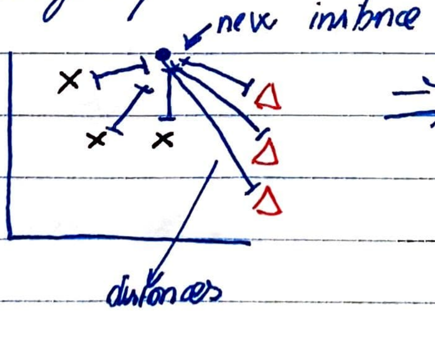
3.

a) There is not training as such. We identify the closest neighbours of an istance. Then, we classify that instance based on the neighbours category.



a new instance here has more probability to be classified as X because proximity To These points

b) The method checks the distance with all the points in the space and the instance we want to classify. Then, the method checks the class of the "K" closest neighbours and ~~decide~~ decides the class of the new instance based on the majority.



⟹ base on the 2 closest neighbours the new instance is X

c) The more neighbours the more robust

Exercise 4.

→ 20 instances
→ 15 instances properly identified
→ 1 positive } we can change what is positive and negative
→ 0 negative }

→ accuracy = $\frac{15 \text{ (instances properly classified)}}{20 \text{ (total number of instances)}} = 0.75$

b)
→ TP: positive instances that we classified properly = 8
→ TN: negative instances that we classified properly = 7
→ FP: negative ~~[illegible]~~ instances we predicted as positive = 2
→ FN: ~~for~~ positive instances we predicted as negative = 3

	1	0 ← real
1	8	2
0	3	7

↑ predicted

8 → 7: main diagonal

c) If the numbers in the main diagonal are higher than the other numbers in the matrix, the classifier is doing a good job. The confusion matrix can indicate the balance of our classier. Thus, we can know if our classifier commits more mistakes predicting negative or positive samples.

d) Precision = $\frac{TP}{TP+FP} = \frac{8}{8+2} = 0.2$
↓
among all the positive predictions, how many are correct?

4. d) recall $= \frac{TP}{TP + FN}$

↓

among all the positive instances, how many have we predicted?

f) $F_1 = 2 \times \frac{precision \times recall}{precision + recall} = 2 \times \frac{0.8 \times 0.73}{0.8 + 0.73} = 0.76$

↓ aims to be a measure between precision and recall

Exercise 5 → K-nearest neighbours

instance 1: (x_1, y_1)

instance 2: (x_2, y_2)

Manhattan distance: $|x_1 - x_2| + |x_2 - y_2|$

instance 1: (3, 1)

	class	Distance
X_1	C_1	$\lvert 3-1\rvert + \lvert 1-1\rvert =$ ②
X_2	C_1	$\lvert 3-2\rvert + \lvert 1-2\rvert =$ ②
X_3	C_1	$\lvert 3-1\rvert + \lvert 1-3\rvert = 4$
X_4	C_1	$\lvert 3-4\rvert + \lvert 1-2\rvert =$ ②
X_5	C_2	$\lvert 3-1\rvert + \lvert 1-6\rvert = 7$
X_6	C_2	$\lvert 3-2\rvert + \lvert 1-4\rvert = 4$
X_7	C_2	$\lvert 3-2\rvert + \lvert 1-5\rvert = 5$
X_8	C_2	$\lvert 3-3\rvert + \lvert 1-4\rvert = 3$
X_9	C_2	$\lvert 3-5\rvert + \lvert 1-4\rvert = 5$

Nearest Neighbours all are C1 → Instance 1 : C_1

Instance 2: (4,5)

		Distance
X_1	C_1	7
X_2	C_1	5
X_3	C_1	5
X_4	C_1	3
X_5	C_2	4
X_6	C_2	3
X_7	C_2	②
X_8	C_2	②
X_9	C_2	②

Nearest Neighbour all are C_2 Instance 2 → C_2

Instance 3: (2,3)

		Distance
X_1	C_1	3
X_2	C_1	①
X_3	C_1	①
X_4	C_1	3
X_5	C_2	4
X_6	C_2	①
X_7	C_2	2
X_8	C_2	2
X_9	C_2	4

3 Nearest Neighbours .....

C_1, C_1, C_2

→ 9

Exercise 5

Instance 3

$P(C_1) = \frac{2}{3}$, $P(C_2) = \frac{1}{3}$