KNN Practice Problem

Notice: *These problems are NOT Exam 3's problems. These problems are SIMILAR to Exam 3's problems.*

Problem 1

Given the data.

	Age	Sex	Survived
A	27	M	0
B	30	F	1
C	80	F	1
D	50	M	0
E	60	F	0

a. Let F be a female of 40 years old. Use 1NN to predict whether F is survived (`Survived =1`) or not (`Survived = 0`). Does the prediction change if used 3NN

b. Given the following data, use 1NN and 3NN to predict the salary for F.

	Age	Sex	Salary k
A	27	M	80
B	30	F	70
C	80	F	90
D	50	M	60
E	60	F	10

Solution

a.

Step 1: Standardize the `Age` column.

$$\text{Standardized} Age = \frac{Age - Age_{min}}{Age_{max} - Age_{min}} = \frac{Age - 27}{80 - 27}$$

We have the follows.

	Age	Sex	Survived
A	0	M	0
B	0.057	F	1
C	1	F	1
D	0.434	M	0

	Age	Sex	Survived
E	0.623	F	0
F	0.245	F	

Step 2: Calculate the distances from F to all the samples.

We have

$$AF = \sqrt{(0 - 0.245)^2 + 1^2} = 1.029575 \quad BF = \sqrt{(0.057 - 0.245)^2 + 0^2} = 0.188 \quad CF = \sqrt{(1 - 0.245)^2 + 0^2} = 0.755 \quad DF = \sqrt{(0.43\ldots}$$

We have BF < EF < CF < DF < AF. B is the nearest neighbor and B, E, C are the three nearest neighbor.

Since the nearest neighbor of F is B (BF is the smallest between AF, BF, CF, DF, and EF), thus we predicts F has the same outcome as B, which is Survived.

If we use 3NN, the prediction of F would be the majority outcome of the 3 nearest neighbors. The three nearest neighbor of F are B, E and C. The majority outcomes of B, E and C is still Survived. Thus, 3NN also predicts F Survived. The prediction does not change from 1NN to 3NN in this case.
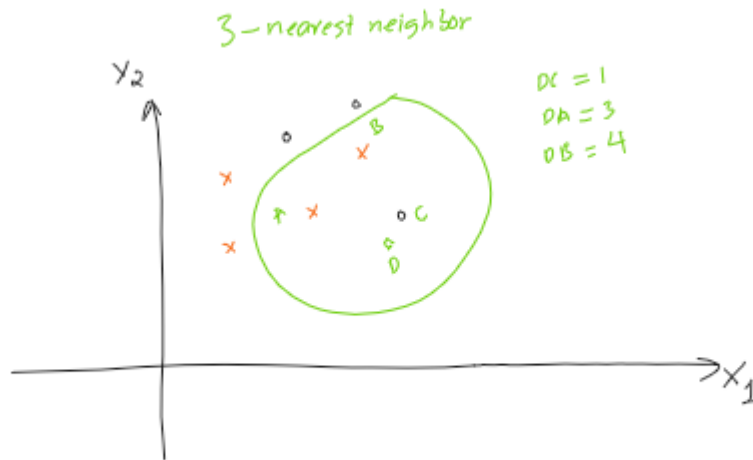
    b. Since the distances does not change, the neareast neighbors are still the same as part a.

For 1NN, the predicted salary of F is the same as B, 70k.

For 3NN, the predicted salary of F is the average salaries of B, E and C, which is: $\frac{70+90+10}{3} = 56.67k$
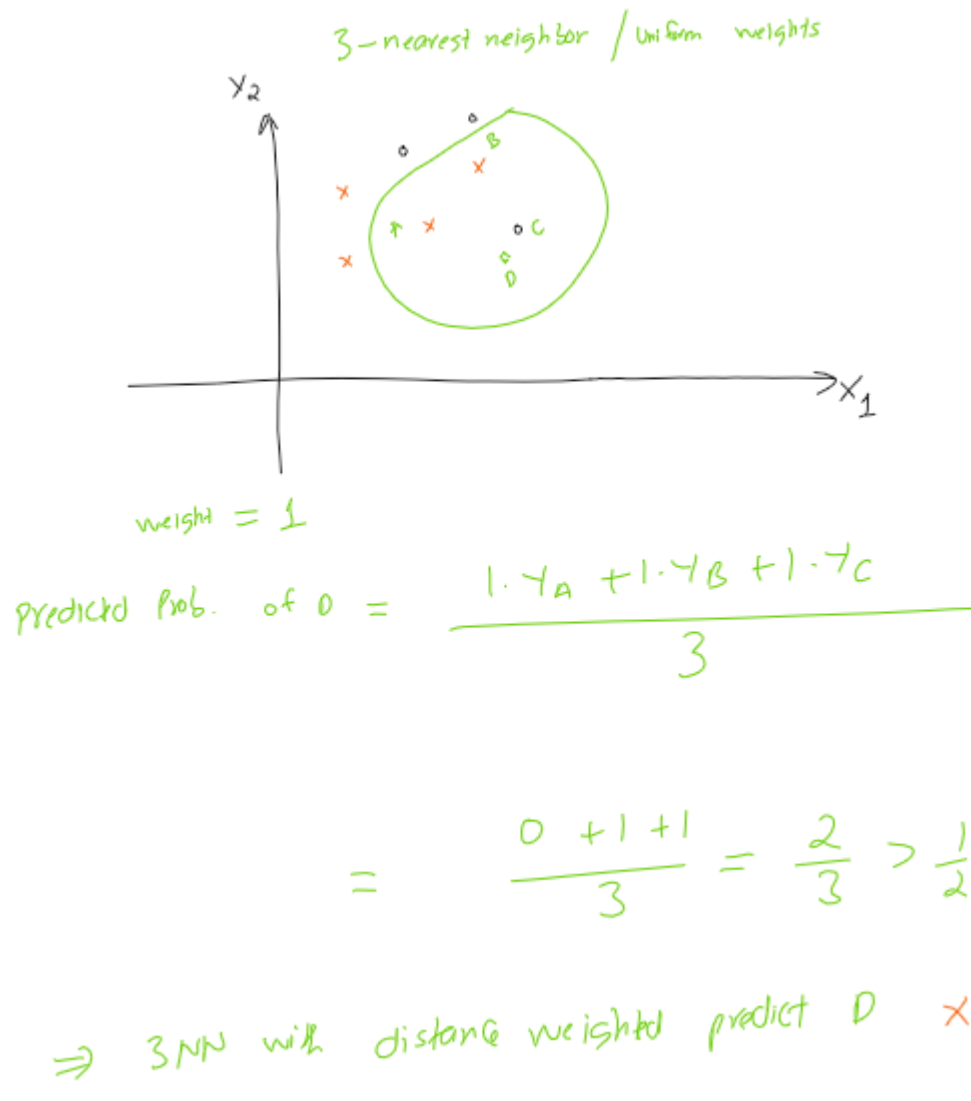
Problem 2

Given the data. Consider x as 1 and o as 0.

    a. Use the **uniform weights** to calculate the predicted probability and the prediction of **3NN** for D.
    b. Use the **distance weights** to calculate the predicted probability and the prediction of **3NN** for D.
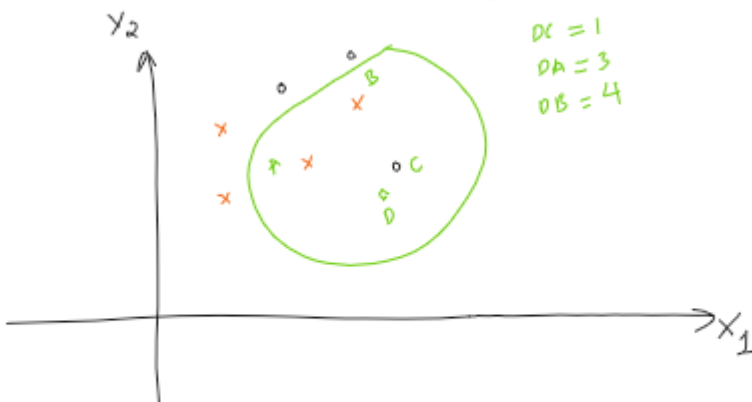
Solution

3-nearest neighbor / Uniform weights

y_2

weight $= 1$

predicted Prob. of 0 = $\dfrac{1 \cdot y_A + 1 \cdot y_B + 1 \cdot y_C}{3}$

$= \dfrac{0 + 1 + 1}{3} = \dfrac{2}{3} > \dfrac{1}{2}$

$\Rightarrow$ 3NN with distance weighted predict D ✗

a.

Notice that if the probability is greater or equal to 1/2, we predict x.

3-nearest neighbor / Distance weights

y_2

3-nearest neighbor region (as drawn)

$DC = 1$
$DA = 3$
$DB = 4$

x_1

$$weight = \frac{1}{Distance}$$

$$predicted\ Prob.\ of\ D = \frac{\frac{1}{DC} \cdot Y_D + \frac{1}{DA} \cdot Y_A + \frac{1}{DB} \cdot Y_B}{\frac{1}{DC} + \frac{1}{DA} + \frac{1}{DB}}$$

$$= \frac{1 \cdot 0 + \frac{1}{3} \cdot 1 + \frac{1}{4} \cdot 1}{1 + \frac{1}{3} + \frac{1}{4}}$$

$$= 0.368 < \frac{1}{2}$$

$\Rightarrow$ 3NN with distance weighted predict D as ○

b.

Problem 3

Given the utility matrix

	Item 1	Item 2	Item 3	Item 4	Item 5
Alice	5	3	4	4	
User 1	3	1	2	3	3
User 2	4	3	4	3	5
User 3	3	3	1	4	4
User 4	1	5	5	2	1

Should we recommend Item 5 to Alice? Calculate her estimated rating on Item 5 to answer the question.

4

Recommend the item if Alice's rating is 4 or above.

    a. Use user-based KNN, with $k = 2$ and Manhantan distance.

    b. Use item-based KNN, with $k = 2$ and cosine similarity.

Solution

    a. We calculate the Manhantance distance from Alice to other users. Let d_{A1} be the Manhantan distance from Alice to User 1 (exclude Item 5). We have.

$$d_{A1} = |5 - 3| + |3 - 1| + |4 - 2| + |4 - 3| = 7$$

Similarly,

$$d_{A2} = |5 - 4| + |3 - 3| + |4 - 4| + |4 - 3| = 2$$
$$d_{A3} = |5 - 3| + |3 - 3| + |4 - 1| + |4 - 4| = 5$$
$$d_{A4} = |5 - 1| + |3 - 5| + |4 - 5| + |4 - 2| = 9.$$

Thus, the two nearest neighbors are User 2 and User 3. Alice's rating on Item 5 is predicted as the average ratings of these two Users on Item 5, which is (5+4)/2= 4.5.

We recommend this Item to Alice.

    b. We calculate the cosine similarities of Item 5 to other items (excluding Alice). Let s_{51} be the cosine similarity of Item 5 and Item 1. We have.

$$s_{51} = \frac{3 \cdot 3 + 5 \cdot 4 + 4 \cdot 3 + 1 \cdot 1}{\sqrt{3^2 + 5^2 + 4^2 + 1^2} \cdot \sqrt{3^2 + 4^2 + 3^2 + 1^2}} = 0.9941$$

Similarly,

$$s_{52} = 0.7388$$
$$s_{53} = 0.7226$$
$$s_{54} = 0.9540$$

Since s_{51} and s_{54} are the largest, the two nearest neighbors (largest similarities) of Item 5 are Item 1 and Item 4. Alice's rating on Item 5 is predicted as the average ratings of her ratings for Item 1 and Item 4, which is (5+4)/2= 4.5.

We also recommend this Item to Alice.