

# ASSIGNMENT 4 SOLUTION

- Nayan Man Singh Pradhan

## Excercise 1. (Calculating EPE and regressor)

In this task, we revise Examples 3.2 and 3.3 from the lecture notes for a different setup:

Let  $X : \Omega \rightarrow \mathbb{R}$  be an input variable and  $Y : \Omega \rightarrow \mathbb{R}$  be an output variable. For the input variable we assume that it follows the uniform distribution as  $X \sim \mathcal{U}[-1, 1]$ . (Note the other range!) Moreover, we make the very strong assumption to know the “true” dependency between  $X$  and  $Y$ . Specifically we define  $Y$  via

$$Y := g(X), \quad \text{with } g(x) = x^4.$$

Now we look for a function  $f$  that shall approximate that (usually unknown) relationship between  $X$  and  $Y$ . We claim that

$$f(x) = x^3,$$

is a good approximation to  $Y = g(X)$ .

- Calculate the expected (squared) prediction error for  $f$ . Why is the error that large compared to the lecture example?
- Find the regressor for  $Y = g(x)$ .

(4 Points)

a)

$$\begin{aligned} \text{EPE}(f) &= E(L_2(Y, f(x))) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} (y - f(x))^2 \cdot p(x, y) dx dy \end{aligned}$$

We know (from knowledge 3.1) that when  $X, Y$  is a continuous random variable and  $Y$  is given as  $Y = g(X)$ , where  $g$  is a function on the range of  $X$  that needs to fulfill some properties. We can give the joint probability distribution density for  $X$  and  $Y$  by

$$p(x, y) = p_x(x) \cdot \delta(y - g(x))$$

This gives us,

$$\begin{aligned} &= \int_{\mathbb{R}} \int_{\mathbb{R}} (y - f(x))^2 \cdot p_x(x) \delta(y - g(x)) dx dy \\ &= \int_{\mathbb{R}} p_x(x) \int_{\mathbb{R}} (y - x^3)^2 \delta(y - x^4) dy \cdot dx \\ &\quad \int_{\mathbb{R}} f(y) \cdot \delta(y - y_0) dy = f(y_0) \end{aligned}$$

$$\begin{aligned}
&= \int_{\mathbb{R}} p_x(x) (x^4 - x^3)^2 dx \\
&= \int_{-1}^1 1 \cdot (x^4 - x^3)^2 dx \\
&= \int_{-1}^1 (x^8 - 2x^7 + x^6) dx \\
&= \int_{-1}^1 x^8 dx - 2 \int_{-1}^1 x^7 dx + \int_{-1}^1 x^6 dx \\
&= \left[ \frac{x^9}{9} \right]_{-1}^1 - 2x \left[ \frac{x^8}{8} \right]_{-1}^1 + \left[ \frac{x^7}{7} \right]_{-1}^1 \\
&= \frac{1}{9} + \frac{1}{9} - 2 \times \frac{1}{8} + \frac{1}{8} \times 2 + \frac{1}{7} + \frac{1}{7} \\
&= \frac{32}{63} \approx 0.5079
\end{aligned}$$

The error is larger as compared to the lecture because in this question we are working with  $g(x)$  with degree 9 and  $f(x)$  with degree 3 as compared to the lecture where we were working with  $g(x)$  with degree 2 and  $f(x)$  with degree 9.

b) We know,

$$p(y|x) = \frac{p(x,y)}{p_x(x)} = \frac{p_x(x) \delta(y - g(x))}{p_x(x)} = \delta(y - x^4)$$

Given this, we evaluate the regressor to figure out what would be the best predictor under the given circumstances:

$$\begin{aligned}
E(Y|X=x) &= \int_{\mathbb{R}} y \cdot p(y|x) dy = \int_{\mathbb{R}} y \cdot \delta(y - x^4) dy = x^4 // \\
&\quad \int_{\mathbb{R}} f(y) \delta(y - y_0) dy = f(y_0)
\end{aligned}$$

∴ The regressor for  $Y=g(X)$  is  $x^4$ .

**Excercise 2.** (Regressor derivation)

In the lecture slides, you have seen that the expected (squared) prediction error  $EPE(f)$  is given by

$$EPE(f) = E(L_2(Y, f(X))).$$

Theorem 3.1 then states that the function  $f$  that minimizes the expected (squared) prediction error  $EPE(f)$  is given by

$$f(x) = E(Y|X = x)$$

Note that in the middle of the proof, we encounter an equation

$$\begin{aligned} EPE(f) &= E_X E_{Y|X} [(f(X) - E[Y|X])^2 + 2(f(X) - E[Y|X])(E[Y|X] - Y) \\ &\quad + (E[Y|X] - Y)^2 | X], \end{aligned}$$

where the second term can be dropped. Now, prove that this is possible, i.e. that it holds

$$2E_X E_{Y|X} [(f(X) - E[Y|X])(E[Y|X] - Y) | X] = 0.$$

(4 Points)

Prove that:

$$2 E_X E_{Y|X} [(f(x) - E[Y|x])(E[Y|x] - Y) | X] = 0$$

$$\begin{aligned} \text{LHS} &= E_X \left[ 2 \cdot (f(x) - E(Y|x)) (E[Y|x] - Y) \right] \\ &= 2 \int_R (f(x) - E(Y|x)) (E[Y|x] - Y) \cdot p_x(x) dx \\ &= 0 \\ &= \text{RHS} \end{aligned}$$

$$\therefore \text{LHS} = \text{RHS}$$

Proved

**Excercise 3.** (kNN regression)

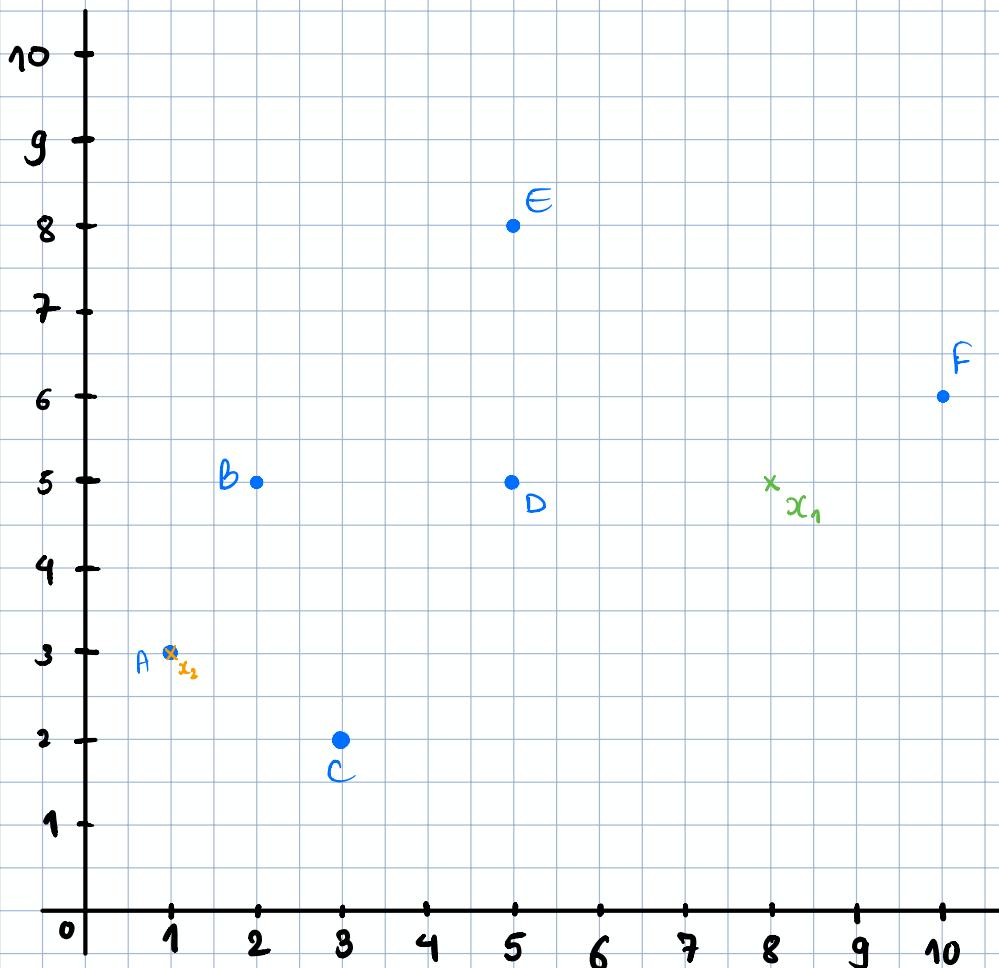
You are given the following training data:

$$\mathcal{T} = \{(A, 77), (B, 47), (C, 55), (D, 59), (E, 72), (F, 60)\}$$

Manually carry a kNN regression prediction for  $x_1 = (8, 5)^\top$ ,  $x_2 = (1, 3)$  and  $k = 1$ ,  $k = 3$ . As part of the task, you have to draw the points in a scatter plot (on paper) and mark the respective neighborhoods that contribute to the final result.

(4 Points)

let, A  $(1, 3)^\top$ , B  $(2, 5)^\top$ , C  $(3, 2)^\top$ , D  $(5, 5)^\top$ , E  $(5, 8)^\top$  and F  $(10, 6)^\top$



First, lets calculate the euclidean distance between  $x_1$  and all training data points.

$$dist(A, x_1) = \sqrt{(3-5)^2 + (1-8)^2} = 7.28$$

I have calculated the distances and illustrated it in Table 1

	A	B	C	D	E	F
$x_1$	7.28	6.0	5.83	3.0	4.24	2.23
$x_2$	0.0	2.23	2.23	4.47	6.40	9.98

Table(1)

We know general formula for KNN regression,

$$Y(x_i) = \frac{1}{K} \sum_{j=1}^K (y_j); \text{ where input: } \{(x_i, y_i)\}_{i=1}^N \text{ training data}$$

: k neighbor size

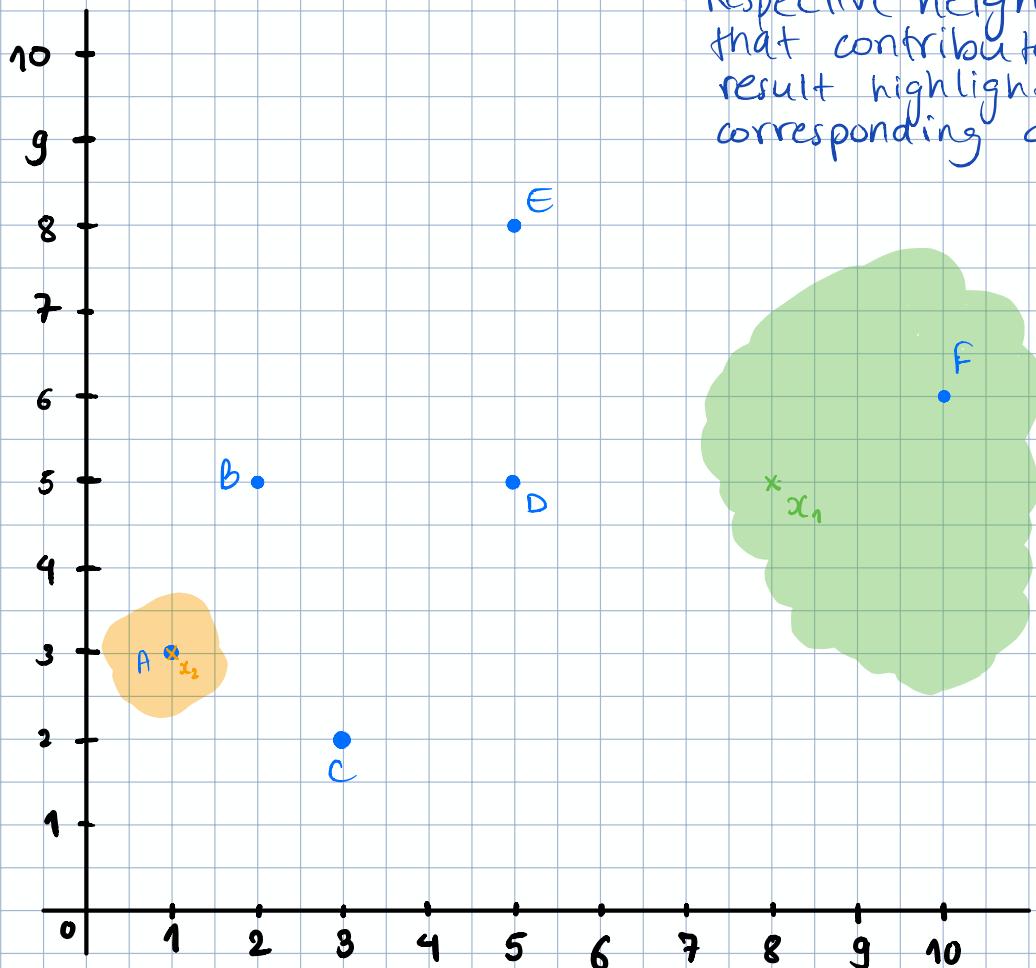
When  $K=1$ ,

For  $x_1$ , the closest neighbor is point F (from Table 1)

$$\therefore Y(x_1) = \frac{1}{1} \sum_{j=1}^1 60 = 60,$$

for  $x_2$ , the closest neighbor is point A (from Table 1)

$$\therefore Y(x_2) = \frac{1}{1} \sum_{j=1}^1 77 = 77,$$



\*Respective neighborhood that contribute to final result highlighted with corresponding color

When  $K = 3$ ,

For  $x_1$ , the closest neighbors are F, D and E (from Table 1)

$$\therefore Y(x_1) = \frac{1}{3} (60 + 59 + 72) = 63.67,$$

for  $x_2$ , the closest neighbors are A, B and C (from Table 1)

$$\therefore Y(x_2) = \frac{1}{3} (77 + 47 + 55) = 59.67,$$

\*Respective neighborhood that contribute to final result highlighted with corresponding color

