k-Nearest Neighbors (*k*NN)

MACHINE LEARNING

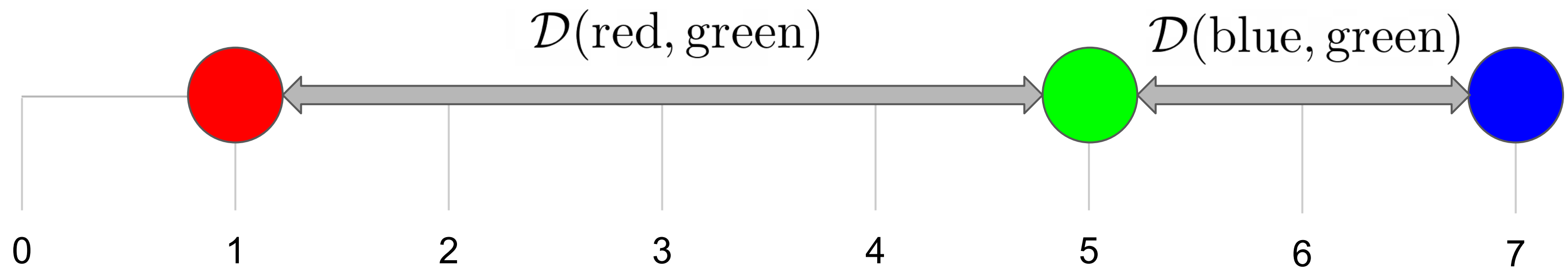
Pakarat Musikawan

Distance Metrics

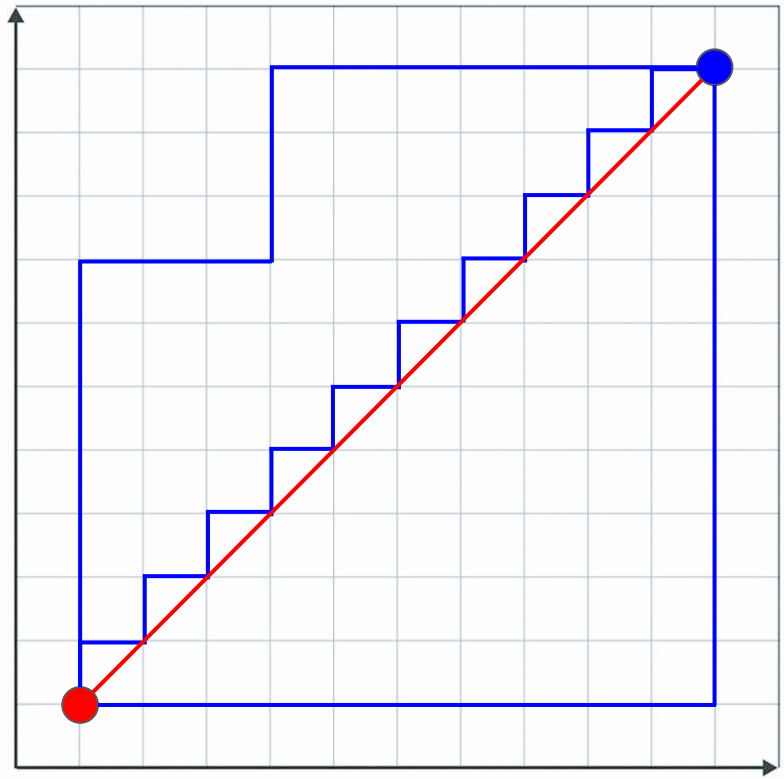
$$\mathcal{D}(\text{red}, \text{green}) = |\text{red} - \text{green}|$$

$$\mathcal{D}(\text{blue}, \text{green}) = |\text{blue} - \text{green}|$$

$$\mathcal{D}(\text{red}, \text{green}) > \mathcal{D}(\text{blue}, \text{green})$$



Distance Metrics



$$\mathcal{D}(\text{red}, \text{blue}) = d(\Delta x, \Delta y)$$
$$= d(x_{\text{red}} - x_{\text{blue}}, y_{\text{red}} - x_{\text{blue}})$$

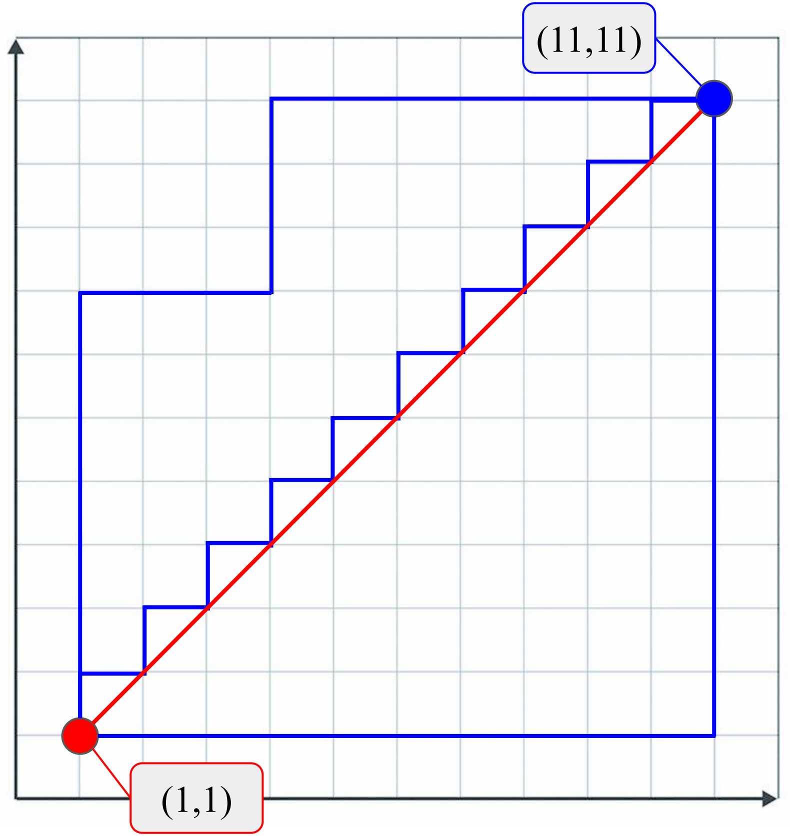
Euclidean

$$\mathcal{D}(\text{red}, \text{blue}) = d(\Delta x, \Delta y)$$
$$= d(x_{\text{red}} - x_{\text{blue}}, y_{\text{red}} - x_{\text{blue}})$$
$$= \sqrt{(x_{\text{red}} - x_{\text{blue}})^2 + (y_{\text{red}} - y_{\text{blue}})^2}$$

Manhattan

$$\mathcal{D}(\text{red}, \text{blue}) = d(\Delta x, \Delta y)$$
$$= d(x_{\text{red}} - x_{\text{blue}}, y_{\text{red}} - x_{\text{blue}})$$
$$= |x_{\text{red}} - x_{\text{blue}}| + |y_{\text{red}} - y_{\text{blue}}|$$

Distance Metrics



$$\begin{aligned}\mathcal{D}(\text{red}, \text{blue}) &= d(\Delta x, \Delta y) \\ &= d(x_{\text{red}} - x_{\text{blue}}, y_{\text{red}} - x_{\text{blue}}) \\ &= d(1 - 11, 1 - 11)\end{aligned}$$

Euclidean

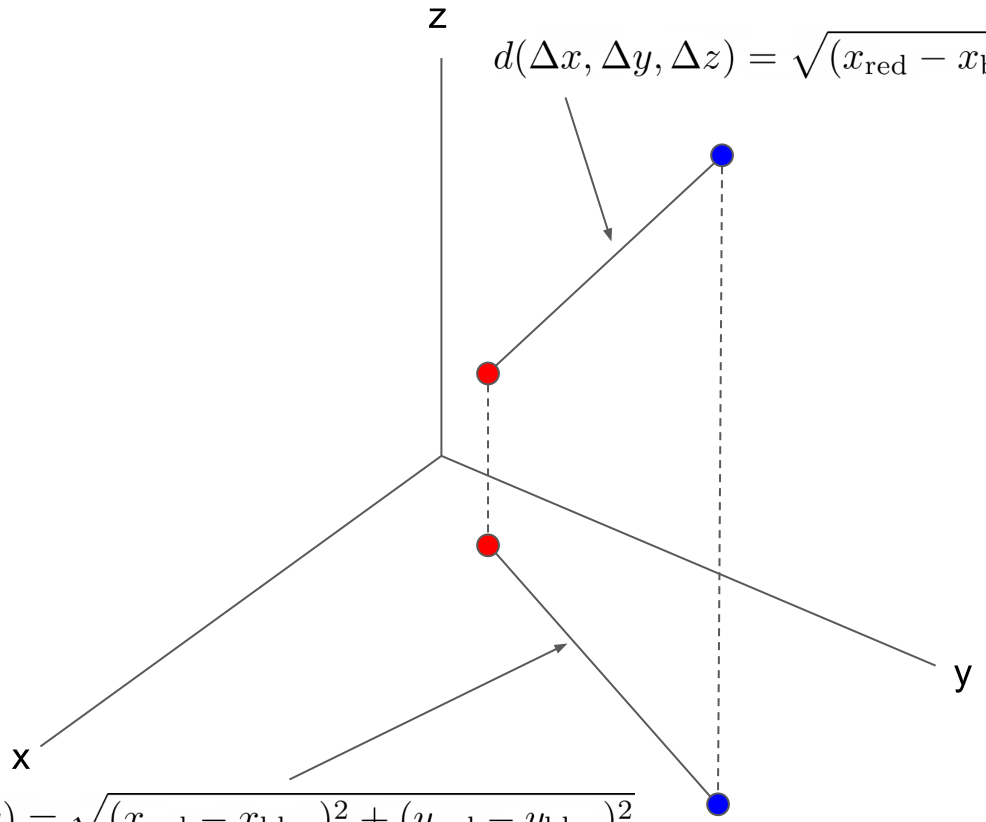
$$\begin{aligned}\mathcal{D}(\text{red}, \text{blue}) &= d(\Delta x, \Delta y) \\ &= d(x_{\text{red}} - x_{\text{blue}}, y_{\text{red}} - x_{\text{blue}}) \\ &= \sqrt{(1 - 11)^2 + (1 - 11)^2} = 14.14\end{aligned}$$

Manhattan

$$\begin{aligned}\mathcal{D}(\text{red}, \text{blue}) &= d(\Delta x, \Delta y) \\ &= d(x_{\text{red}} - x_{\text{blue}}, y_{\text{red}} - x_{\text{blue}}) \\ &= |1 - 11| + |1 - 11| = 20\end{aligned}$$

Euclidean Distance

$$d(\Delta x, \Delta y, \Delta z) = \sqrt{(x_{\text{red}} - x_{\text{blue}})^2 + (y_{\text{red}} - y_{\text{blue}})^2 + (z_{\text{red}} - z_{\text{blue}})^2}$$



$$d(\Delta x, \Delta y) = \sqrt{(x_{\text{red}} - x_{\text{blue}})^2 + (y_{\text{red}} - y_{\text{blue}})^2}$$

Euclidean Distance

$$d(\Delta x) = \sqrt{(x_A - x_B)^2}$$

$$d(\Delta x, \Delta y) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

$$d(\Delta x, \Delta y, \Delta z) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2}$$

$$\boldsymbol{A} = [A_1, A_2, \ldots, A_n] \qquad \boldsymbol{B} = [B_1, B_2, \ldots, B_n]$$

$$\begin{aligned} \mathcal{D}(\boldsymbol{A}, \boldsymbol{B}) &= \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + \ldots + (A_n - B_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} = \|\boldsymbol{A} - \boldsymbol{B}\| \end{aligned}$$

Minkowski Distance

$$d(\Delta x) = \left((x_A - x_B)^q\right)^{\frac{1}{q}}$$

$$d(\Delta x, \Delta y) = \left((x_A - x_B)^q + (y_A - y_B)^q\right)^{\frac{1}{q}}$$

$$d(\Delta x, \Delta y, \Delta z) = \left((x_A - x_B)^q + (y_A - y_B)^q + (z_A - z_B)^q\right)^{\frac{1}{q}}$$

$$\boldsymbol{A} = [A_1, A_2, \ldots, A_n] \qquad \boldsymbol{B} = [B_1, B_2, \ldots, B_n]$$

$$\mathcal{D}(\boldsymbol{A}, \boldsymbol{B}) = \left((A_1 - B_1)^q + (A_2 - B_2)^q + \ldots + (A_n - B_n)^q\right)^{\frac{1}{q}}$$

$$= \left(\sum_{i=1}^{n} (A_i - B_i)^q\right)^{\frac{1}{q}}$$

Manhattan Distance

$$d(\Delta x) = |x_A - x_B|$$

$$d(\Delta x, \Delta y) = |x_A - x_B| + |y_A - y_B|$$

$$d(\Delta x, \Delta y, \Delta z) = |x_A - x_B| + |y_A - y_B| + |z_A - z_B|$$

$$\boldsymbol{A} = [A_1, A_2, \ldots, A_n] \qquad \boldsymbol{B} = [B_1, B_2, \ldots, B_n]$$

$$\mathcal{D}(\boldsymbol{A}, \boldsymbol{B}) = |A_1 - B_1| + |A_2 - B_2| + \ldots + |A_n - B_n|$$

$$= \sum_{i=1}^{n} |A_i - B_i|$$

Hamming Distance

$$d(\Delta x) = x_A \oplus x_B$$

$$d(\Delta x, \Delta y) = (x_A \oplus x_B) + (y_A \oplus y_B)$$

$$d(\Delta x, \Delta y, \Delta z) = (x_A \oplus x_B) + (y_A \oplus y_B) + (z_A \oplus z_B)$$

$$\boldsymbol{A} = [A_1, A_2, \ldots, A_n] \qquad \boldsymbol{B} = [B_1, B_2, \ldots, B_n]$$

$$\begin{aligned} \mathcal{D}(\boldsymbol{A}, \boldsymbol{B}) &= (A_1 \oplus B_1) + (A_2 \oplus B_2) + \ldots + (A_n \oplus B_n) \\ &= \sum_{i=1}^{n} (A_i \oplus B_i) \end{aligned}$$

Hamming Distance

0	1	1	0	1
1	1	1	0	0

$$
\begin{aligned}
\mathcal{D}(\boldsymbol{A}, \boldsymbol{B}) &= \sum_{i=1}^{5} (A_i \oplus B_i) \\
&= (0 \oplus 1) + (1 \oplus 1) + (1 \oplus 1) + (0 \oplus 0) + (1 \oplus 0) \\
&= 2
\end{aligned}
$$

k-Nearest Neighbor (*k*NN)

- *k*-Nearest Neighbor (*k*NN)
- Memory-based Reasoning
- Example-based Reasoning
- Instance-based Learning
- Lazy Learning

k-Nearest Neighbor (*k*NN)

Customer	Age	Loan	Default
A	25	40000	No
B	35	60000	No
C	45	80000	No
D	20	20000	No
E	35	120000	No
F	52	18000	No
G	23	95000	Yes
H	40	62000	Yes
I	60	100000	Yes
J	48	220000	Yes
K	33	150000	Yes
L	48	142000	?

$$\boxed{\begin{aligned}\mathcal{D}(L, A) &= \sqrt{(48 - 25)^2 + (142000 - 40000)^2} \\ &= 102000.003\end{aligned}}$$

k-Nearest Neighbor (*k*NN)

Customer	Age	Loan	Default
A	25	40000	No
B	35	60000	No
C	45	80000	No
D	20	20000	No
E	35	120000	No
F	52	18000	No
G	23	95000	Yes
H	40	62000	Yes
I	60	100000	Yes
J	48	220000	Yes
K	33	150000	Yes
L	48	142000	?

	Distance	
D(L,A)	102000.003	9
D(L,B)	82000.001	8
D(L,C)	62000.000	5
D(L,D)	122000.003	10
D(L,E)	22000.004	2
D(L,F)	124000.000	11
D(L,G)	47000.007	4
D(L,H)	80000.000	7
D(L,I)	42000.002	3
D(L,J)	78000.000	6
D(L,K)	8000.014	1

k-Nearest Neighbor (kNN)

$$x'_{j,i} = \frac{x_{j,i} - \min(X_i)}{\max(X_i) - \min(X_i)}$$

Customer	Age	Loan	Default
A	25	40000	No
B	35	60000	No
C	45	80000	No
D	20	20000	No
E	35	120000	No
F	52	18000	No
G	23	95000	Yes
H	40	62000	Yes
I	60	100000	Yes
J	48	220000	Yes
K	33	150000	Yes
L	48	142000	?

Age	Loan
0.125	0.109
0.375	0.208
0.625	0.307
0.000	0.010
0.375	0.505
0.800	0.000
0.075	0.381
0.500	0.218
1.000	0.406
0.700	1.000
0.325	0.653
0.700	0.614

k-Nearest Neighbor (*k*NN)

Customer	Age	Loan	Default
A	0.125	0.109	No
B	0.375	0.208	No
C	0.625	0.307	No
D	0.000	0.010	No
E	0.375	0.505	No
F	0.800	0.000	No
G	0.075	0.381	Yes
H	0.500	0.218	Yes
I	1.000	0.406	Yes
J	0.700	1.000	Yes
K	0.325	0.653	Yes
L	0.700	0.614	?

	Distance	
D(L,A)	0.765	10
D(L,B)	0.520	7
D(L,C)	0.316	1
D(L,D)	0.925	11
D(L,E)	0.343	2
D(L,F)	0.622	8
D(L,G)	0.667	9
D(L,H)	0.444	6
D(L,I)	0.365	3
D(L,J)	0.386	5
D(L,K)	0.377	4

k-Nearest Neighbor (kNN)

k=3

Customer	Age	Loan	Default
A	0.125	0.109	No
B	0.375	0.208	No
C	0.625	0.307	No
D	0.000	0.010	No
E	0.375	0.505	No
F	0.800	0.000	No
G	0.075	0.381	Yes
H	0.500	0.218	Yes
I	1.000	0.406	Yes
J	0.700	1.000	Yes
K	0.325	0.653	Yes
L	0.700	0.614	**No**

	Distance	
D(L,A)	0.765	10
D(L,B)	0.520	7
D(L,C)	0.316	1
D(L,D)	0.925	11
D(L,E)	0.343	2
D(L,F)	0.622	8
D(L,G)	0.667	9
D(L,H)	0.444	6
D(L,I)	0.365	3
D(L,J)	0.386	5
D(L,K)	0.377	4

k-Nearest Neighbor (*k*NN)

k=5

Customer	Age	Loan	Default		Distance	
A	0.125	0.109	No	*D*(L,A)	0.765	10
B	0.375	0.208	No	*D*(L,B)	0.520	7
C	0.625	0.307	No	*D*(L,C)	0.316	1
D	0.000	0.010	No	*D*(L,D)	0.925	11
E	0.375	0.505	No	*D*(L,E)	0.343	2
F	0.800	0.000	No	*D*(L,F)	0.622	8
G	0.075	0.381	Yes	*D*(L,G)	0.667	9
H	0.500	0.218	Yes	*D*(L,H)	0.444	6
I	1.000	0.406	Yes	*D*(L,I)	0.365	3
J	0.700	1.000	Yes	*D*(L,J)	0.386	5
K	0.325	0.653	Yes	*D*(L,K)	0.377	4
L	0.700	0.614	**Yes**			

Workshop

Given k is 5, and *D*(.,.) is the Euclidean distance.
Scale both provided datasets into the range of [0-1].
Predict the classes for the unlabeled dataset in the right-hand table by using the training dataset in the left-hand table.

sepal length	sepal width	petal length	class
5.1	3.5	1.4	setosa
4.9	3	1.4	setosa
4.7	3.2	1.3	setosa
4.6	3.1	1.5	setosa
7	3.2	4.7	versicolor
6.4	3.2	4.5	versicolor
6.9	3.1	4.9	versicolor
5.5	2.3	4	versicolor
6.5	2.8	4.6	versicolor
6.3	3.3	6	virginica
5.8	2.7	5.1	virginica
7.1	3	5.9	virginica
6.3	2.9	5.6	virginica
6.5	3	5.8	virginica

sepal length	sepal width	petal length	class
4.8	3	1.4	?
6.6	3	4.4	?
6.7	3	5.2	?