Pacmann

K-Nearest Neighbor

Introduction to Machine Learning

Referensi

- An Introduction to Statistical Learning
  — *Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani*

Outcomes

- Siswa dapat **memahami** cara kerja KNN
- Siswa dapat **memahami** KNN untuk menyelesaikan regresi & klasifikasi
- Siswa dapat **memahami** cara memilih hyperparameter KNN

Outline

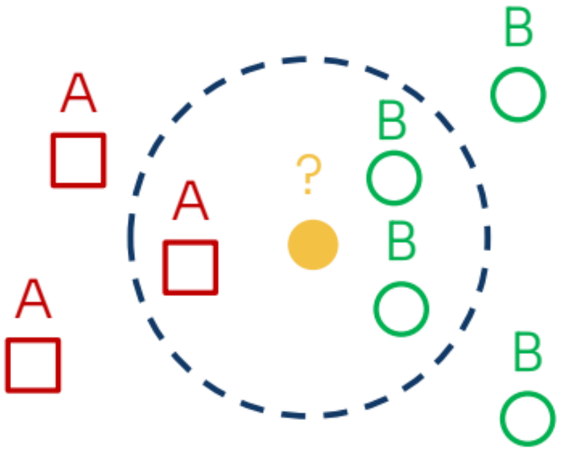
- Simple Memory-Based learning
  - Neighbor
  - Nearest-Neighbor
- K-NN Regressor
- K-NN Classifier
- Calculating the Distance
- Choosing the Best Hyperparameter Value for K
- Curse of Dimensionality

Simple Memory Based Method

K-Nearest Neighbor

Algoritma sederhana

Mengestimasi nilai berbasis **titik** menggunakan **tetangga terdekat** (nearest neighbor)

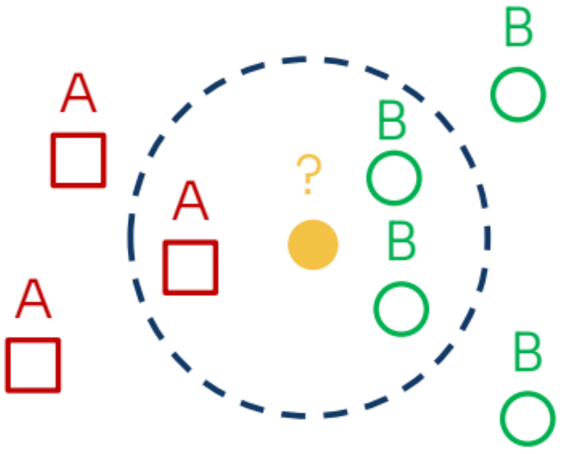


K-Nearest Neighbor

Algoritma sederhana

Mengestimasi nilai berbasis **titik** menggunakan **tetangga terdekat** (nearest neighbor)

Apa itu nearest neighbors?



Example

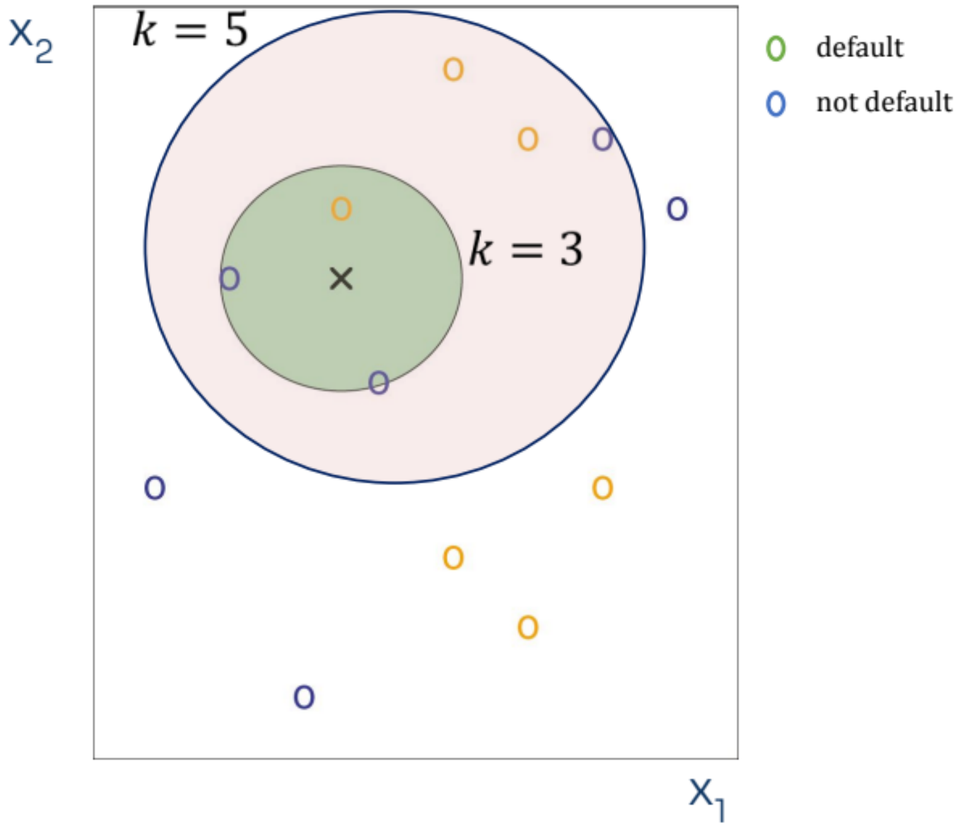
1. **Case:** Credit approval

Applicant Information

Age	23 years
Gender	male
Annual salary	$30,000
Years in residence	1 year
Years in job	1 year
Current debt	$15,000
...	...

K-Nearest Neighbor

Sejumlah k observasi (**k-observations**) yang memiliki **jarak** terdekat dengan titik, misal **x**.

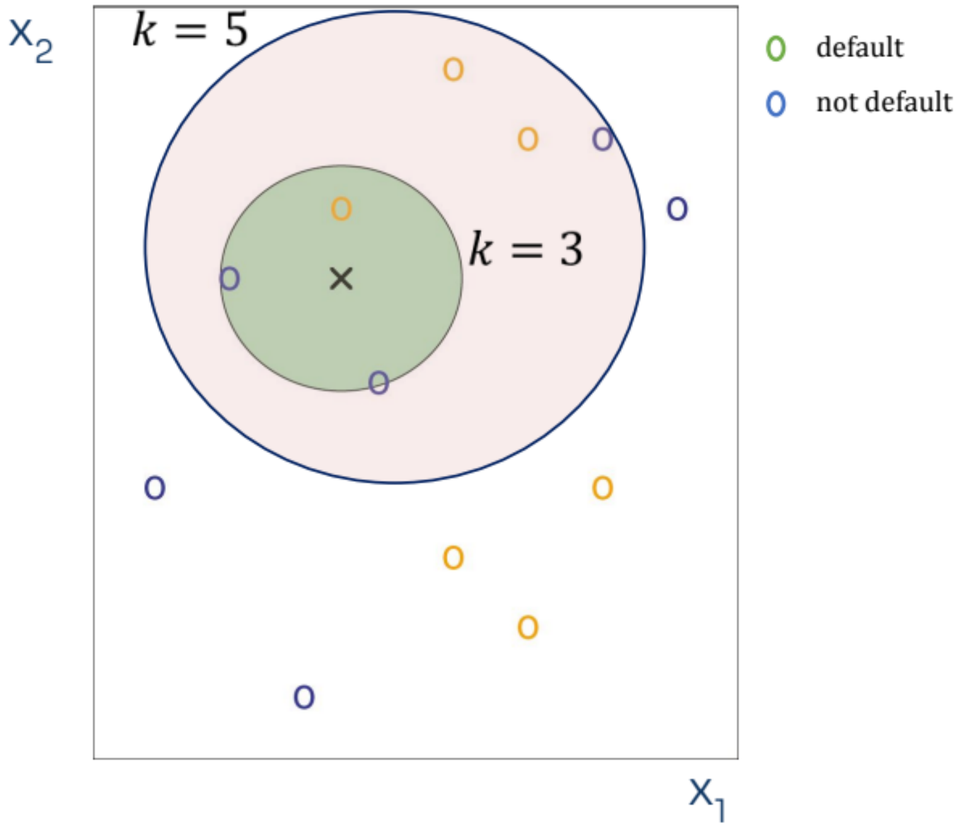


Picture: Introduction to Statistical Learning

K-Nearest Neighbor

Sejumlah k observasi (**k-observations**) yang memiliki **jarak** terdekat dengan titik, misal **x**.

Bagaimana cara prediksi nilai **x**?



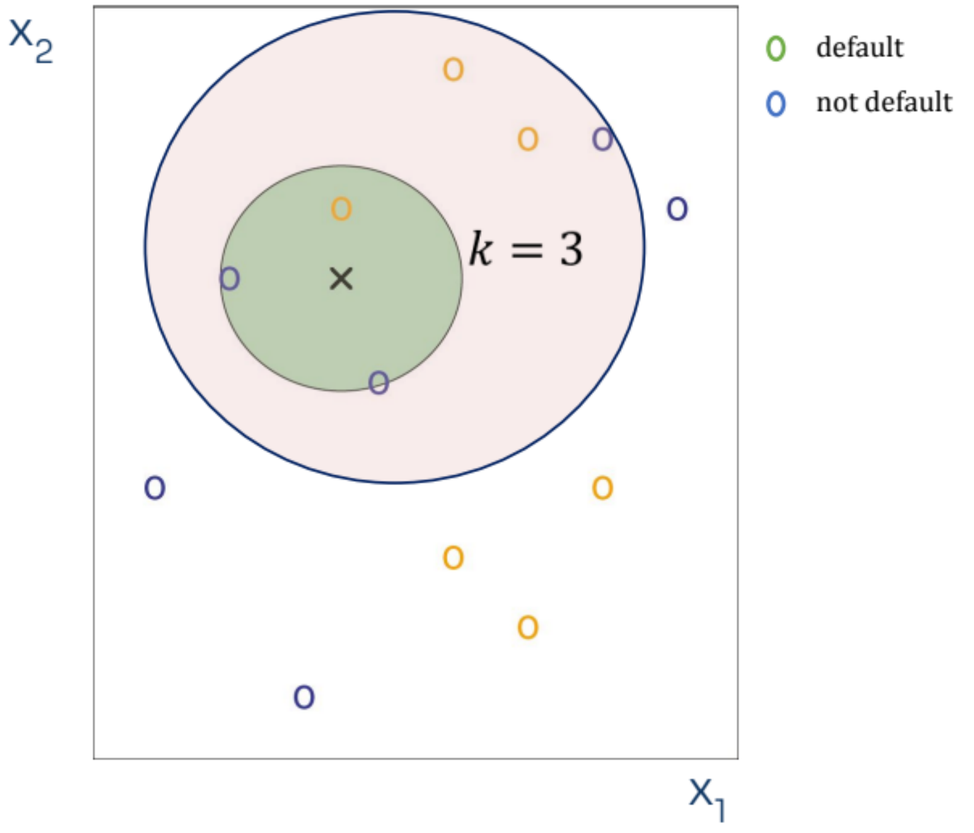
Picture: Introduction to Statistical Learning

K-Nearest Neighbor

Sejumlah k observasi (**k-observations**) yang memiliki **jarak** terdekat dengan titik, misal **x**.

Istilah baru:

1. k-observations
2. distance (jarak)



Picture: Introduction to Statistical Learning

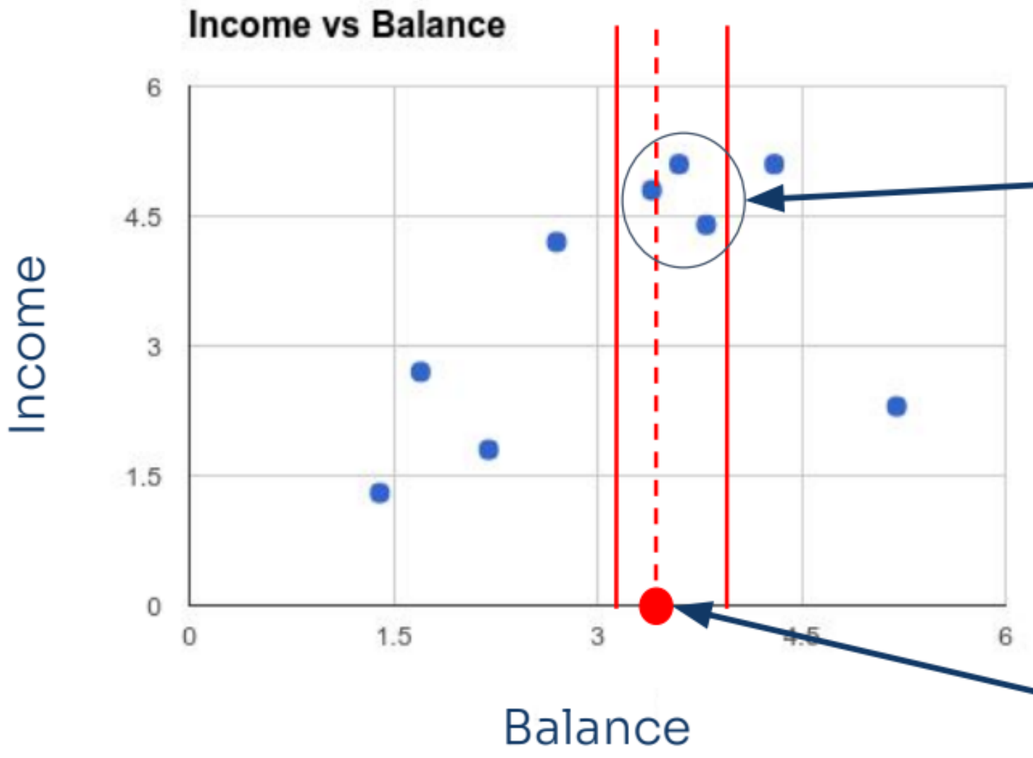
Regression vs Classification

K-Nearest Neighbor

Regression	Classification
Mean/Average: $$f(x) = \frac{1}{n}\sum_{i=1}^{n} y_i$$	Majority Vote: $$f(x) = \frac{1}{n}\sum_{i=1}^{n} I(y_i = 1)$$

Regression

$k = 3$



2

Pilih 3 observasi dari tetangga terdekat, kemudian hitung income dari 3 observasi ini.

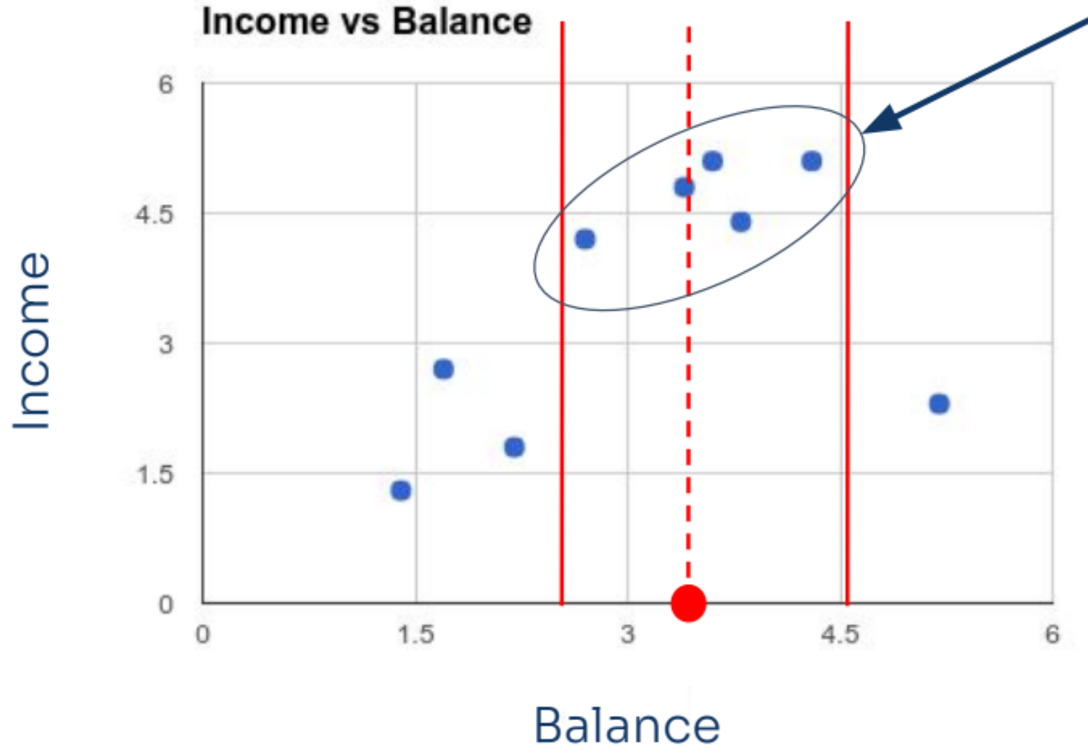
1

Prediksi Income when Balance = 3.3

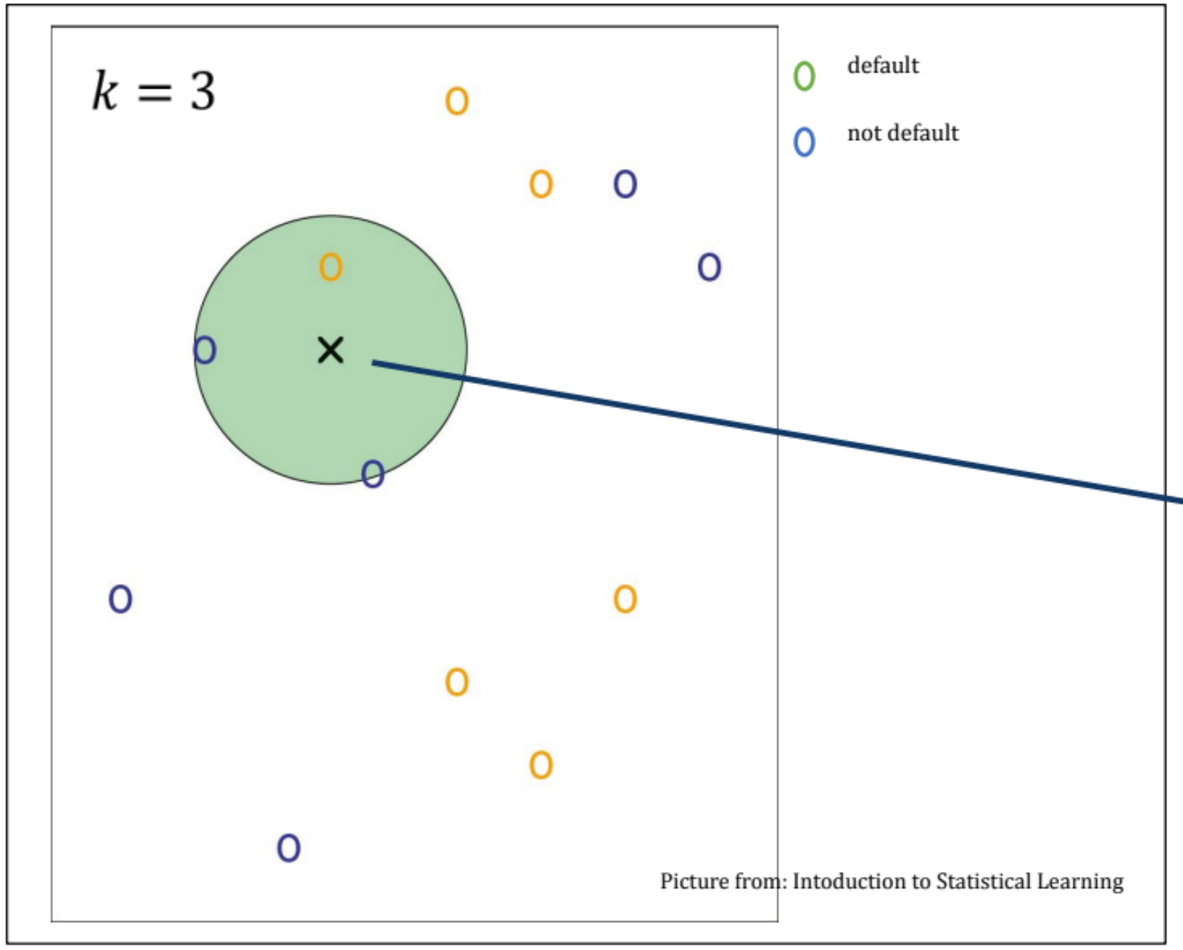
Regression

Hitung Income dari 5 observasi ini

$k = 5$

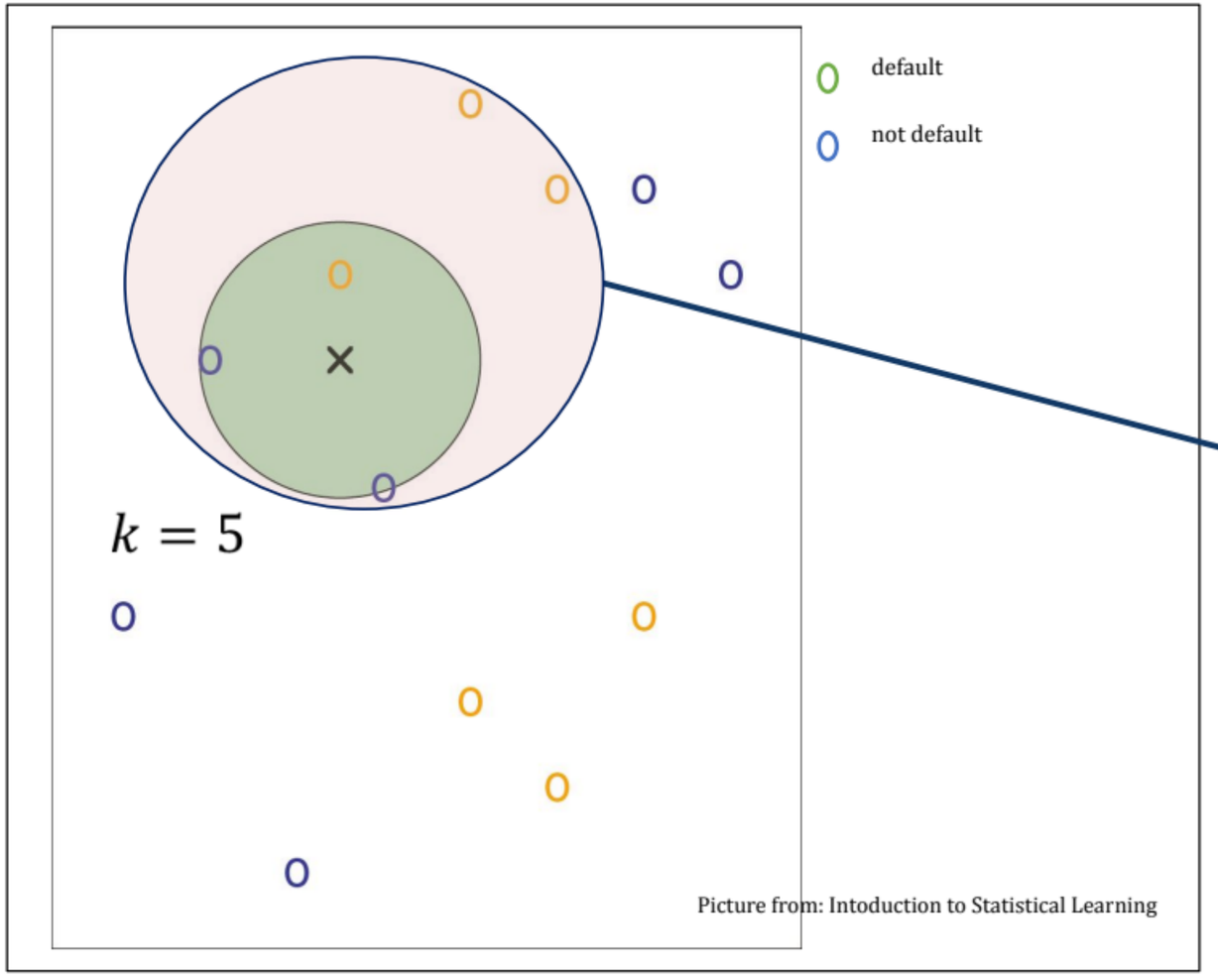


Classification



dengan probabilitas 2/3, prediksi dari "**X**" adalah not default

Classification

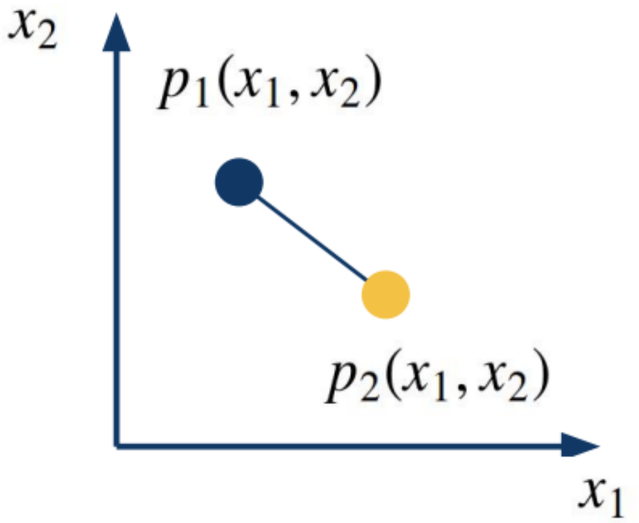


dengan probabilitas 3/5,
prediksi dari "**X**" adalah default

Calculate Distance

Distance

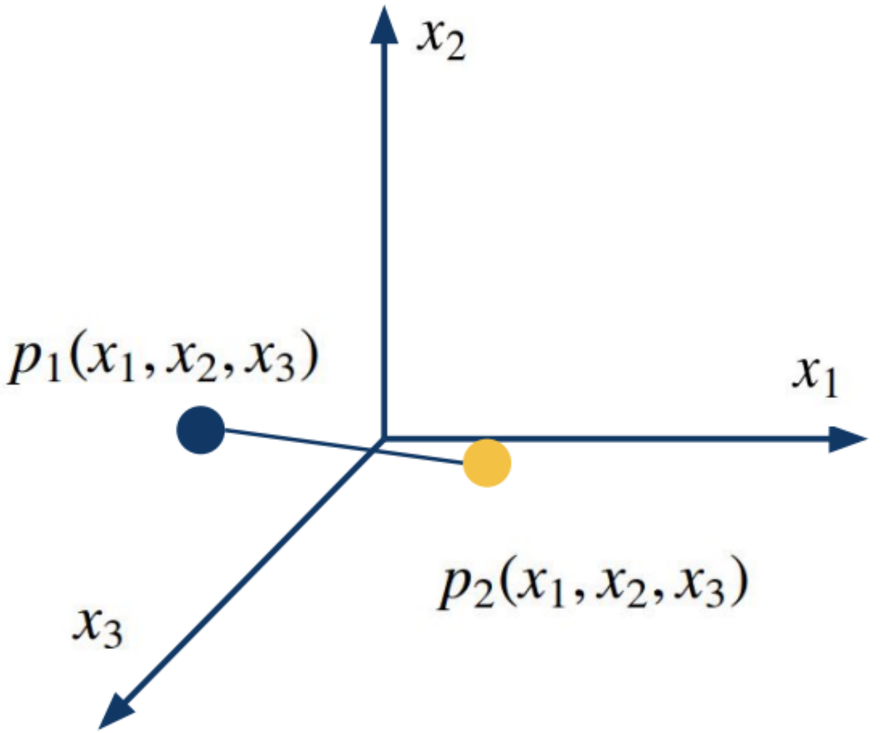
Most Common: **Euclidean Distance**



$$d(p_1, p_2) = \sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2}$$

Distance

Most Common: **Euclidean Distance**

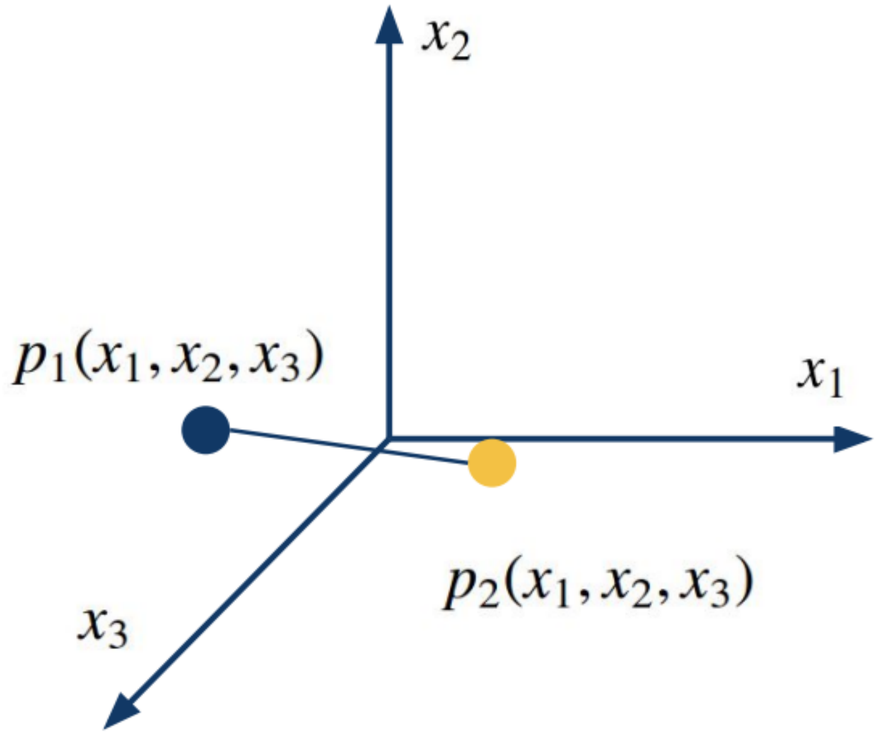


$$d(p_1, p_2) = \sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2 + (x_{31} - x_{32})^2}$$

Distance

Most Common: **Euclidean Distance**

Bingung ????



$$d(p_1, p_2) = \sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2 + (x_{31} - x_{32})^2}$$

Distance

Diberikan Dataset:

Obs	y	x_1	x_2
1	20	500	2000
2	50	300	3000
3	30	200	1000
4	60	600	7000
5	10	100	1500

Kita ingin prediksi:

6	?	200	2500

Distance

Diberikan Dataset:

Obs	y	x_1	x_2
1	20	500	2000
2	50	300	3000
3	30	200	1000
4	60	600	7000
5	10	100	1500

Kita ingin prediksi:

6	?	200	2500

Hitung distance :

$$d(p_1, p_6) = \sqrt{(500 - 200)^2 + (2000 - 2500)^2} = 583.095$$

Distance

Diberikan Dataset:

Obs	y	x_1	x_2
1	20	500	2000
2	50	300	3000
3	30	200	1000
4	60	600	7000
5	10	100	1500

Kita ingin prediksi:

6	?	200	2500

Hitung distance :

$$d(p_1, p_6) = \sqrt{(500 - 200)^2 + (2000 - 2500)^2} = 583.095$$

$$d(p_2, p_6) = \sqrt{(300 - 200)^2 + (3000 - 2500)^2} = 509.902$$

Distance

Diberikan Dataset:

Obs	y	x_1	x_2
1	20	500	2000
2	50	300	3000
3	30	200	1000
4	60	600	7000
5	10	100	1500

Kita ingin prediksi:

6	?	200	2500

Hitung distance :

$$d(p_1, p_6) = \sqrt{(500-200)^2 + (2000-2500)^2} = 583.095$$

$$d(p_2, p_6) = \sqrt{(300-200)^2 + (3000-2500)^2} = 509.902$$

$$d(p_3, p_6) = \sqrt{(200-200)^2 + (1000-2500)^2} = 1500$$

$$d(p_4, p_6) = \sqrt{(600-200)^2 + (7000-2500)^2} = 4517.743$$

$$d(p_5, p_6) = \sqrt{(100-200)^2 + (1500-2500)^2} = 1004.988$$

Distance

Diberikan Dataset:

Obs	y	x_1	x_2
1	20	500	2000
2	50	300	3000
3	30	200	1000
4	60	600	7000
5	10	100	1500

Kita ingin prediksi:

6	?	200	2500

Hitung distance :

$$d(p_1, p_6) = \sqrt{(500-200)^2 + (2000-2500)^2} = 583.095$$

$$d(p_2, p_6) = \sqrt{(300-200)^2 + (3000-2500)^2} = 509.902$$

$$d(p_3, p_6) = \sqrt{(200-200)^2 + (1000-2500)^2} = 1500$$

$$d(p_4, p_6) = \sqrt{(600-200)^2 + (7000-2500)^2} = 4517.743$$

$$d(p_5, p_6) = \sqrt{(100-200)^2 + (1500-2500)^2} = 1004.988$$

Distance

Diberikan Dataset:

Obs	y	x_1	x_2
1	20	500	2000
2	50	300	3000
3	30	200	1000
4	60	600	7000
5	10	100	1500

Kita ingin prediksi:

6	?	200	2500

Hitung distance :

$$d(p_1, p_6) = \sqrt{(500-200)^2 + (2000-2500)^2} = 583.095$$

$$d(p_2, p_6) = \sqrt{(300-200)^2 + (3000-2500)^2} = 509.902$$

$$d(p_3, p_6) = \sqrt{(200-200)^2 + (1000-2500)^2} = 1500$$

$$d(p_4, p_6) = \sqrt{(600-200)^2 + (7000-2500)^2} = 4517.743$$

$$d(p_5, p_6) = \sqrt{(100-200)^2 + (1500-2500)^2} = 1004.988$$

Distance

Diberikan Dataset:

Obs	y	x_1	x_2
1	20	500	2000
2	50	300	3000
3	30	200	1000
4	60	600	7000
5	10	100	1500

Kita ingin prediksi:

6	?	200	2500

Hitung distance :

$$d(p_1, p_6) = \sqrt{(500-200)^2 + (2000-2500)^2} = 583.095$$

$$d(p_2, p_6) = \sqrt{(300-200)^2 + (3000-2500)^2} = 509.902$$

$$d(p_3, p_6) = \sqrt{(200-200)^2 + (1000-2500)^2} = 1500$$

$$d(p_4, p_6) = \sqrt{(600-200)^2 + (7000-2500)^2} = 4517.743$$

$$d(p_5, p_6) = \sqrt{(100-200)^2 + (1500-2500)^2} = 1004.988$$

$k = 1 \longrightarrow y = 50$

$k = 3 \longrightarrow y = (20 + 50 + 10)/3 = 26.6667$

Distance

Diberikan Dataset:

Obs	y	x_1	x_2
1	20	500	2000
2	50	300	3000
3	30	200	1000
4	60	600	7000
5	10	100	1500

Kita ingin prediksi:

6	?	200	2500

Di standarisasi

	x_1	x_2
Mean	340	2900
Std.Dev	207.36	2408.32

???

Obs	y	x_1	x_2
1	20	0.772	-0.374
2	50	-0.193	0.042
3	30	-0.675	-0.789
4	60	1.254	1.702
5	10	-1.157	-0.581

6	?	-0.675	-0.166

Distance

Dataset Baru:

Obs	y	x_1	x_2
1	20	0.772	-0.374
2	50	-0.193	0.042
3	30	-0.675	-0.789
4	60	1.254	1.702
5	10	-1.157	-0.581

6	?	-0.675	-0.166

Lakukan hal yang sama:

Distance

Dataset Baru:

Obs	y	x_1	x_2
1	20	0.772	-0.374
2	50	-0.193	0.042
3	30	-0.675	-0.789
4	60	1.254	1.702
5	10	-1.157	-0.581

6	?	-0.675	-0.166

Lakukan hal yang sama:

$$d(p_1, p_6) = \sqrt{(0.772 - (-0.675))^2 + ((-0.374) - (-0.166))^2} = 1.4615$$

$$d(p_2, p_6) = \sqrt{((-0.193) - (-0.675))^2 + (0.042 - (-0.166))^2} = 0.5250$$

$$d(p_3, p_6) = \sqrt{((-0.675) - (-0.675))^2 + ((-0.789) - (-0.166))^2} = 0.6228$$

$$d(p_4, p_6) = \sqrt{(1.254 - (-0.675))^2 + (1.702 - (-0.166))^2} = 2.6856$$

$$d(p_5, p_6) = \sqrt{((-1.157) - (-0.675))^2 + ((-0.581) - (-0.166))^2} = 0.6364$$

Distance

Dataset Baru:

Obs	y	x_1	x_2
1	20	0.772	-0.374
2	50	-0.193	0.042
3	30	-0.675	-0.789
4	60	1.254	1.702
5	10	-1.157	-0.581

6	?	-0.675	-0.166

Lakukan hal yang sama:

$$d(p_1, p_6) = \sqrt{(0.772 - (-0.675))^2 + ((-0.374) - (-0.166))^2} = 1.4615$$

$$d(p_2, p_6) = \sqrt{((-0.193) - (-0.675))^2 + (0.042 - (-0.166))^2} = 0.5250$$

$$d(p_3, p_6) = \sqrt{((-0.675) - (-0.675))^2 + ((-0.789) - (-0.166))^2} = 0.6228$$

$$d(p_4, p_6) = \sqrt{(1.254 - (-0.675))^2 + (1.702 - (-0.166))^2} = 2.6856$$

$$d(p_5, p_6) = \sqrt{((-1.157) - (-0.675))^2 + ((-0.581) - (-0.166))^2} = 0.6364$$

Distance

Dataset Baru:

Obs	y	x_1	x_2
1	20	0.772	-0.374
2	50	-0.193	0.042
3	30	-0.675	-0.789
4	60	1.254	1.702
5	10	-1.157	-0.581

6	?	-0.675	-0.166

Lakukan hal yang sama:

$$d(p_1, p_6) = \sqrt{(0.772 - (-0.675))^2 + ((-0.374) - (-0.166))^2} = 1.4615$$

$$d(p_2, p_6) = \sqrt{((-0.193) - (-0.675))^2 + (0.042 - (-0.166))^2} = 0.5250$$

$$d(p_3, p_6) = \sqrt{((-0.675) - (-0.675))^2 + ((-0.789) - (-0.166))^2} = 0.6228$$

$$d(p_4, p_6) = \sqrt{(1.254 - (-0.675))^2 + (1.702 - (-0.166))^2} = 2.6856$$

$$d(p_5, p_6) = \sqrt{((-1.157) - (-0.675))^2 + ((-0.581) - (-0.166))^2} = 0.6364$$

Distance

Dataset Baru:

Obs	y	x_1	x_2
1	20	0.772	-0.374
2	50	-0.193	0.042
3	30	-0.675	-0.789
4	60	1.254	1.702
5	10	-1.157	-0.581

6	?	-0.675	-0.166

Lakukan hal yang sama:

$$d(p_1, p_6) = \sqrt{(0.772 - (-0.675))^2 + ((-0.374) - (-0.166))^2} = 1.4615$$

$$d(p_2, p_6) = \sqrt{((-0.193) - (-0.675))^2 + (0.042 - (-0.166))^2} = 0.5250$$

$$d(p_3, p_6) = \sqrt{((-0.675) - (-0.675))^2 + ((-0.789) - (-0.166))^2} = 0.6228$$

$$d(p_4, p_6) = \sqrt{(1.254 - (-0.675))^2 + (1.702 - (-0.166))^2} = 2.6856$$

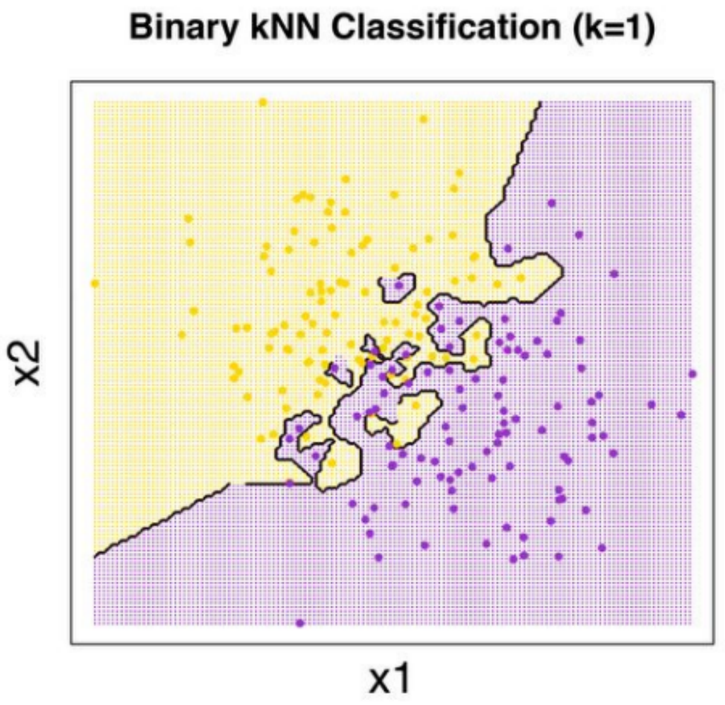
$$d(p_5, p_6) = \sqrt{((-1.157) - (-0.675))^2 + ((-0.581) - (-0.166))^2} = 0.6364$$

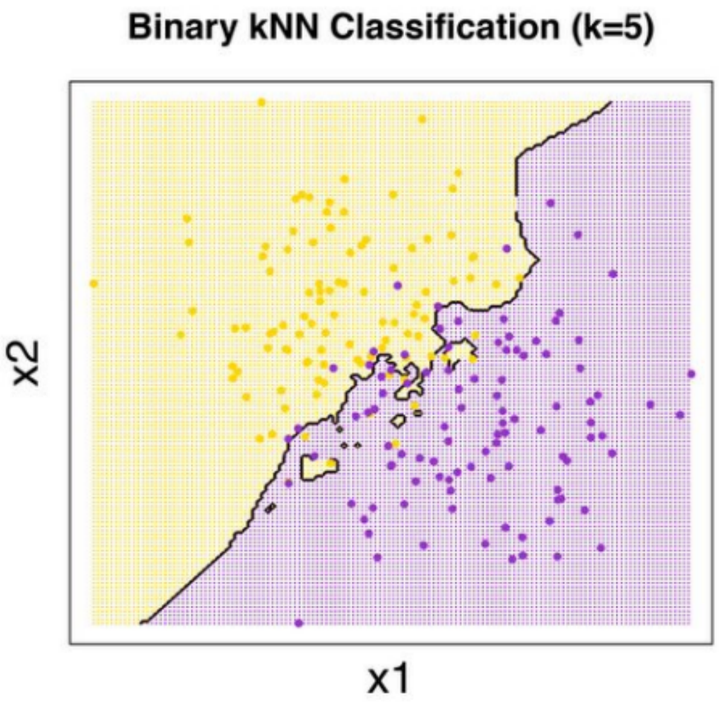
$k = 1 \longrightarrow y = 50$

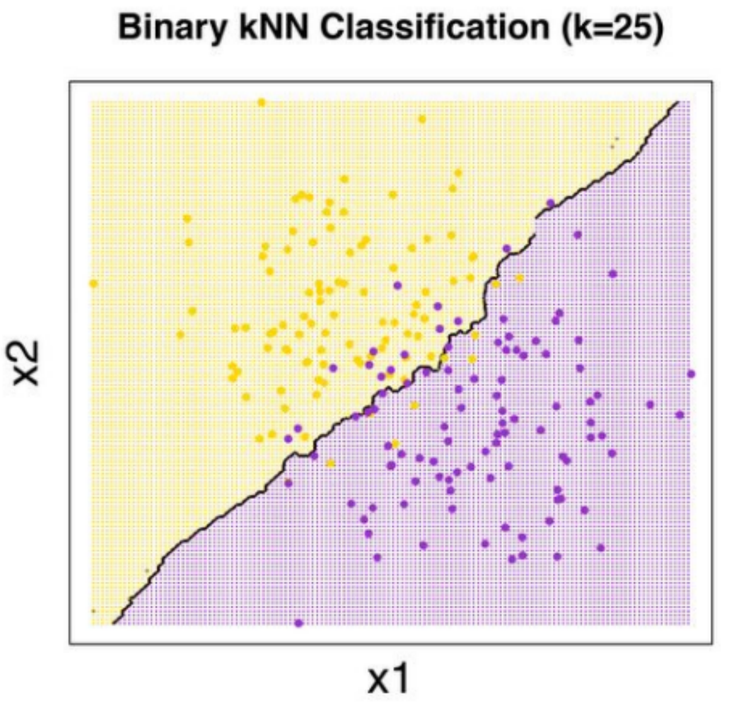
$k = 3 \longrightarrow y = (50 + 30 + 10)/3 = 30$

Choosing best hyperparameter value for K

Choosing *k* : **Classification**

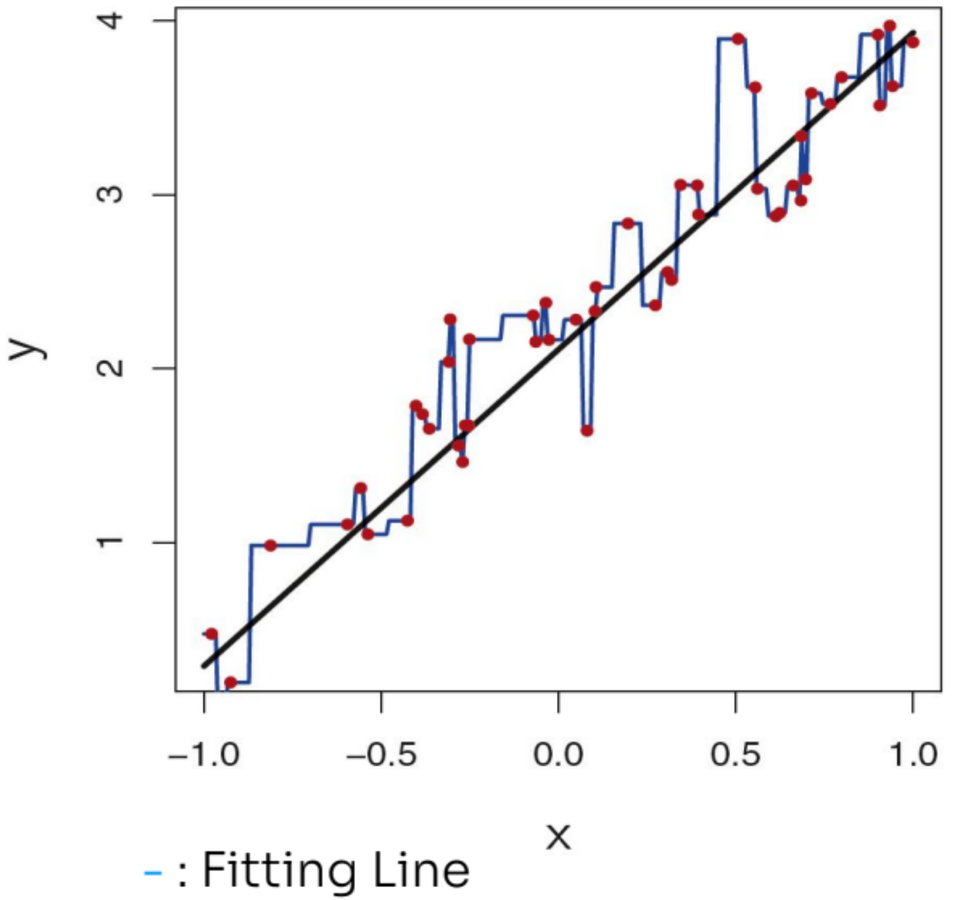




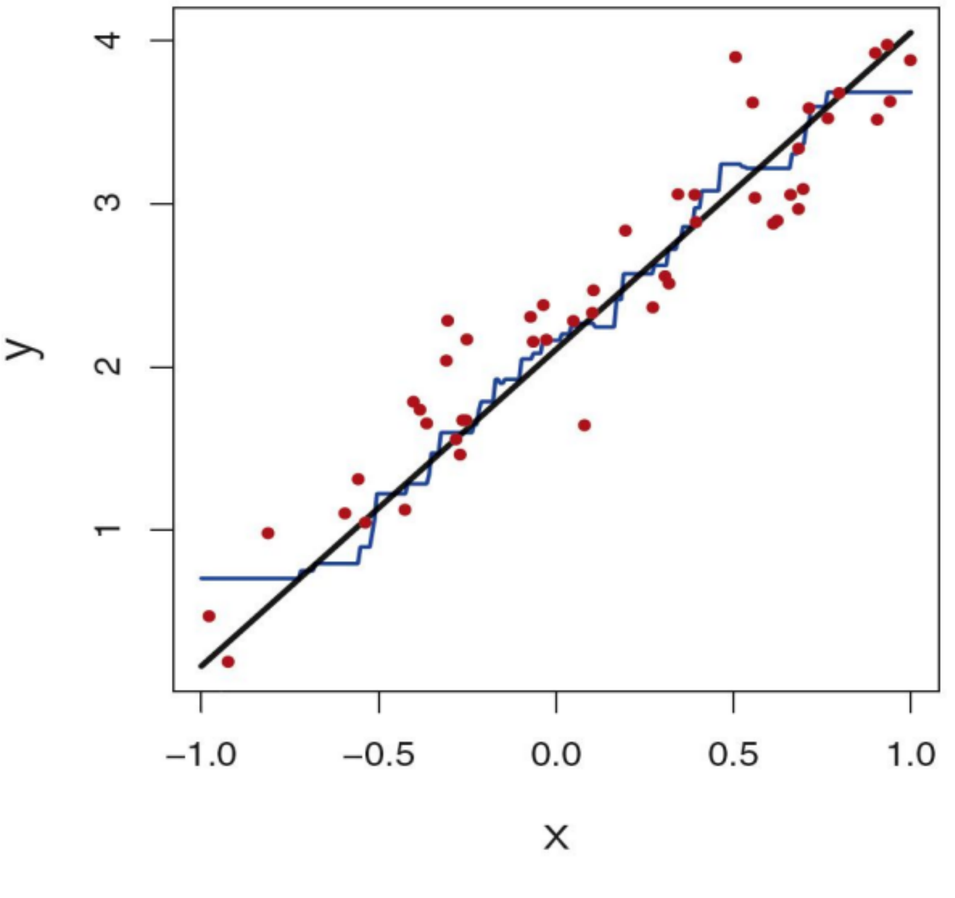


Choosing *k* : **Regression**

K kecil

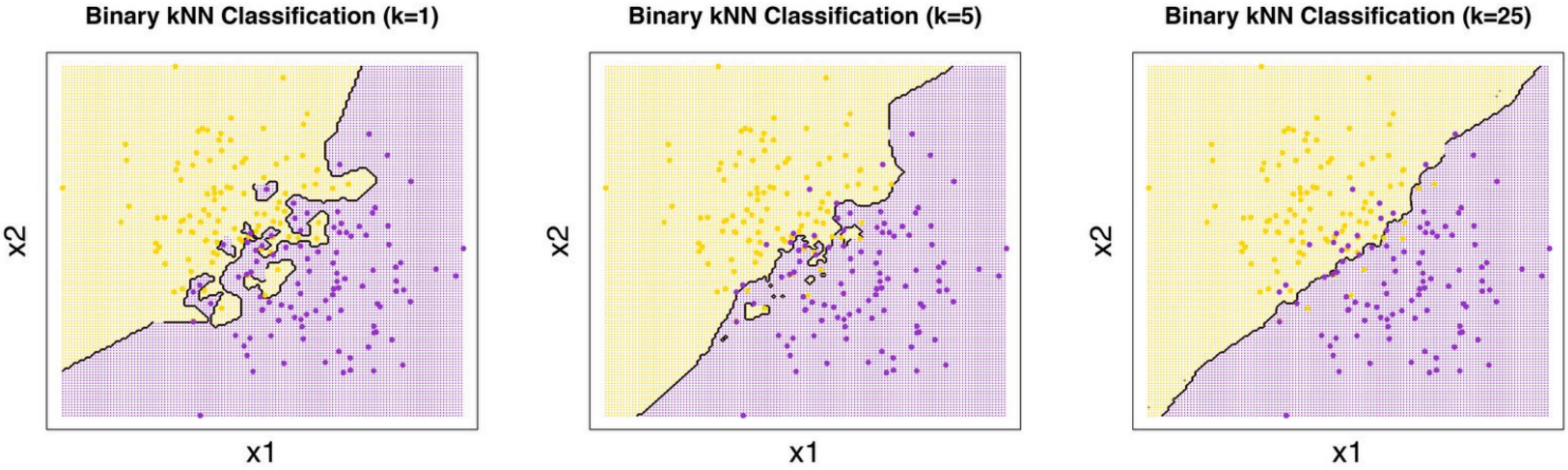


K besar

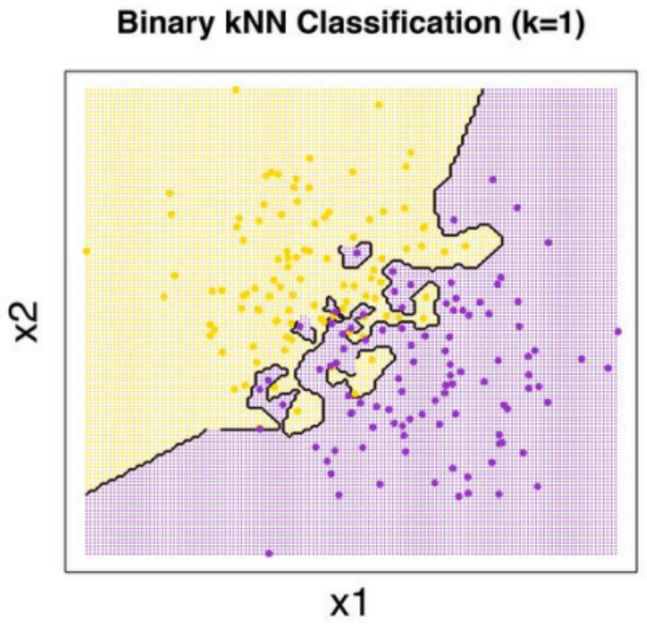


– : Fitting Line

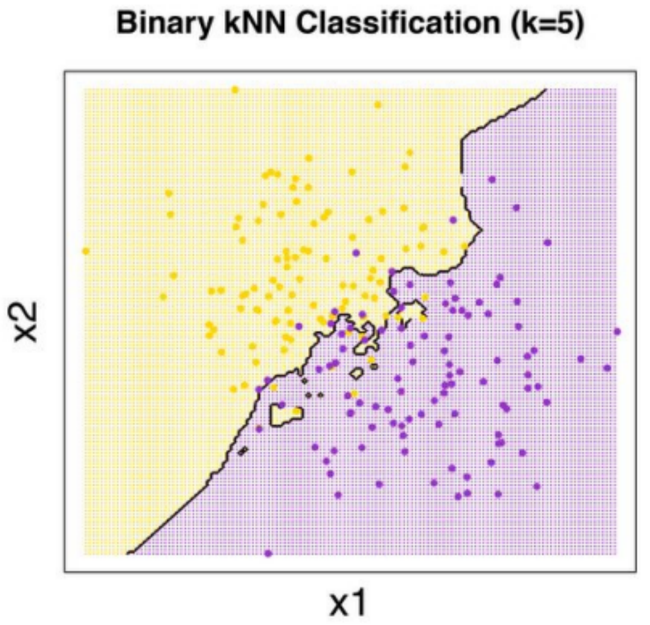
Large *k:* **Less Sensitive to Noise**



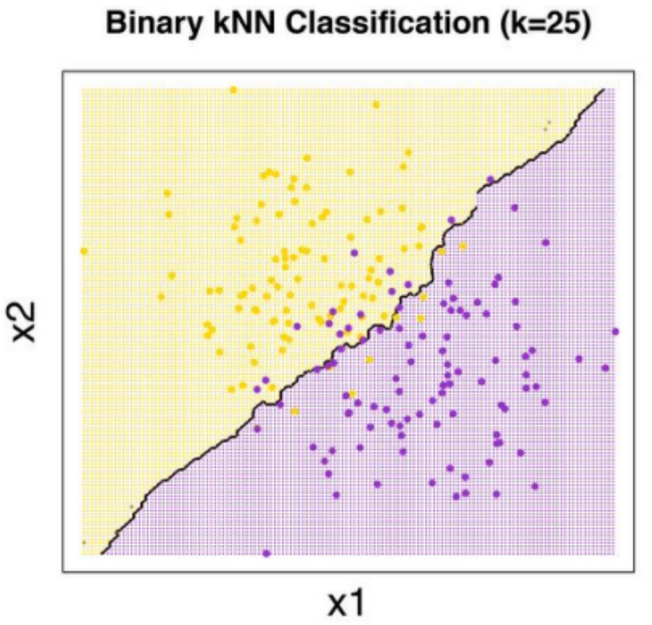
Large *k*: **Better probability estimates**



Yellow = 1

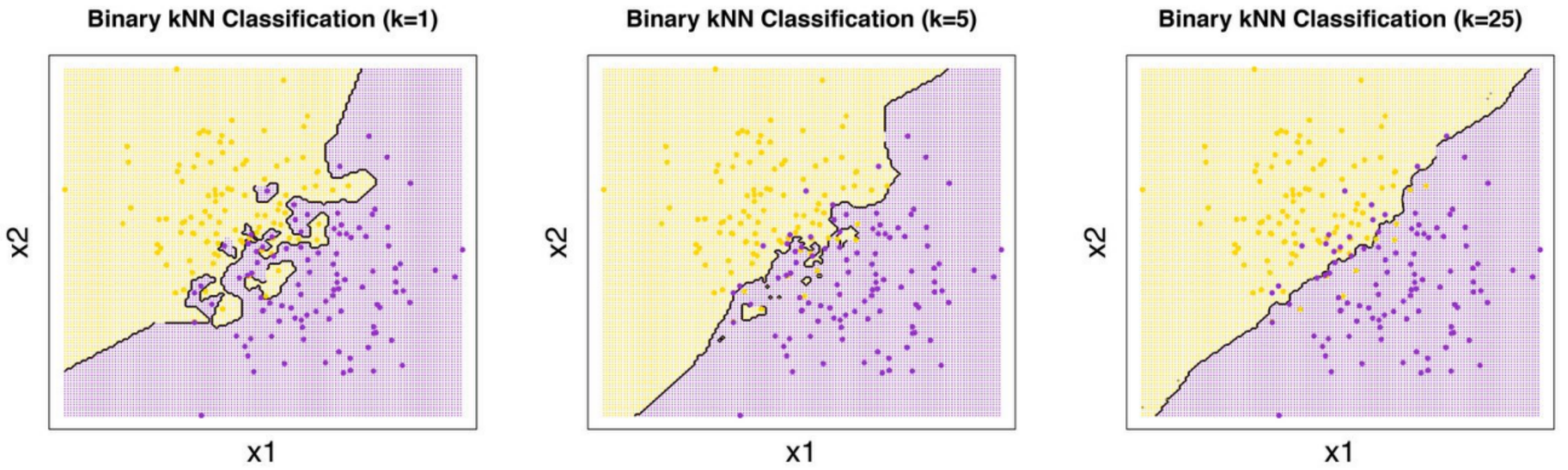


Yellow = 2/5 = 0.4



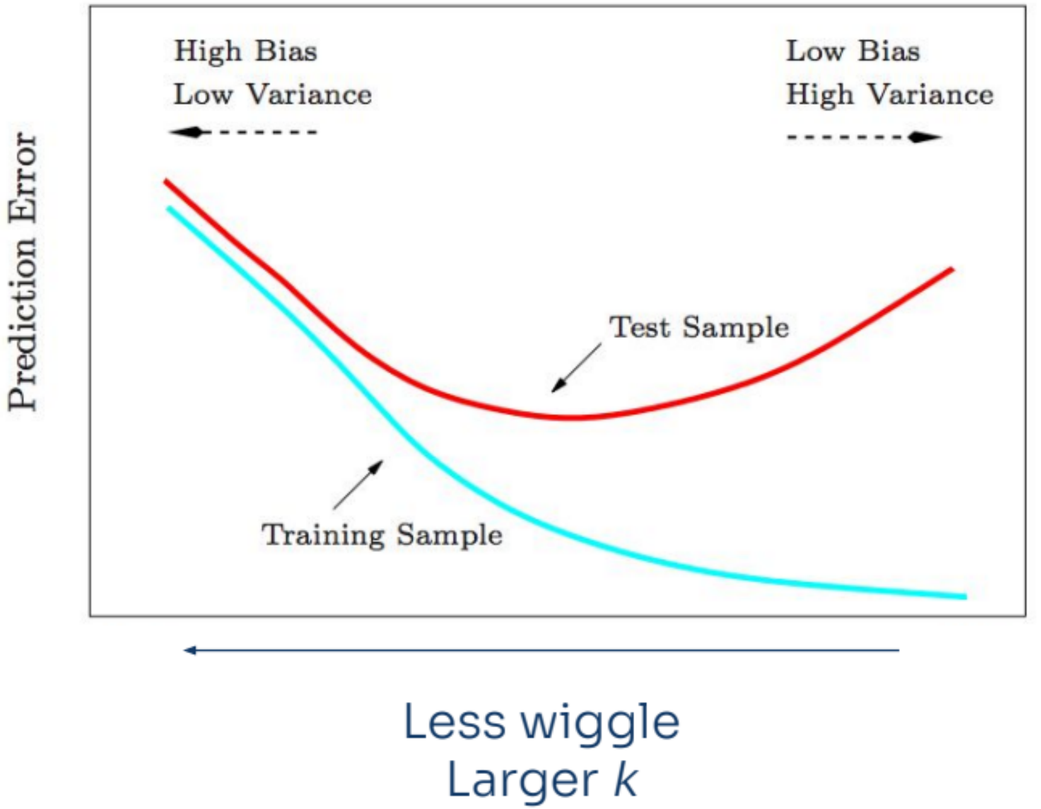
Yellow = 16/25 = 0.64

Large *k:* **Less Complex Model**

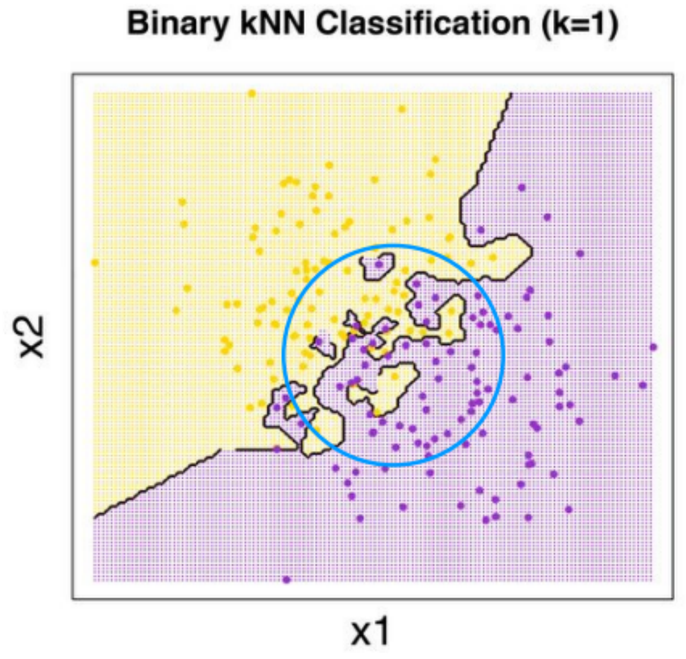


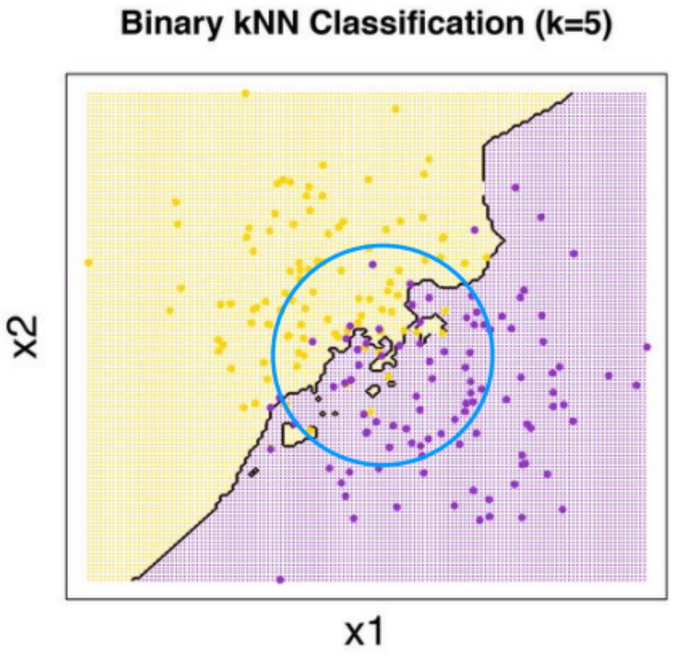
Less wiggle

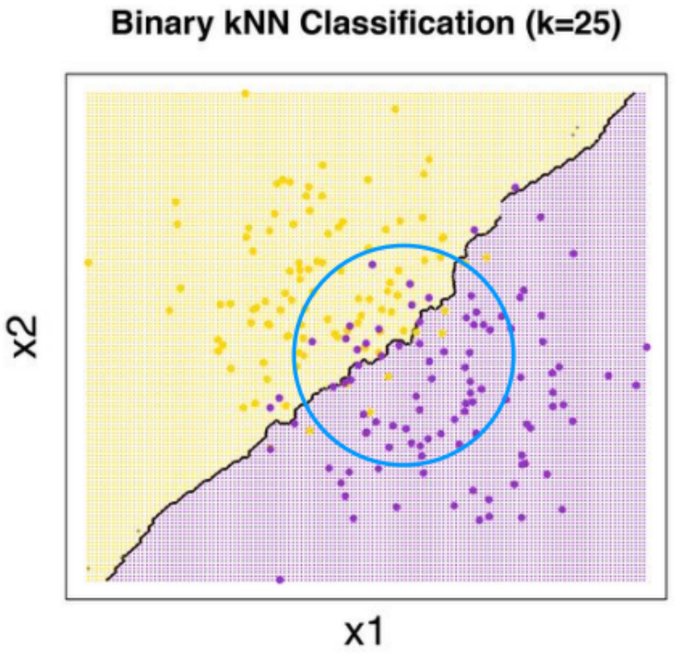
Large *k*: **Large Estimation Bias**



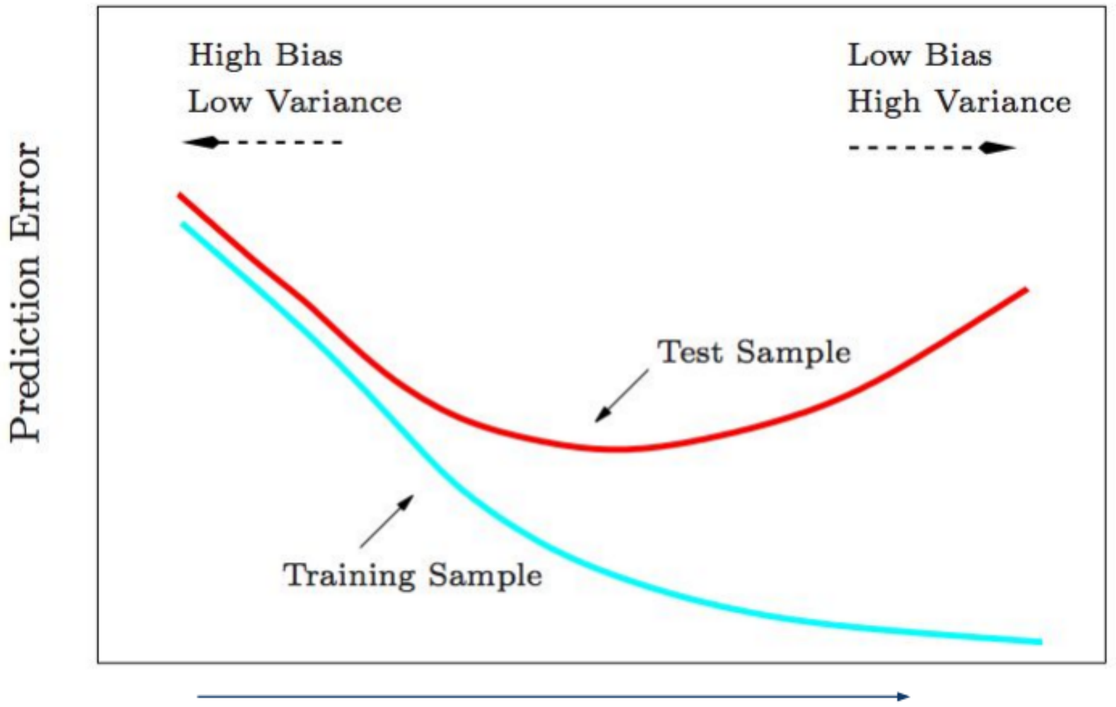
Small *k*: **Better at Capturing Problem**







Small *k*: **Increasing Model Complexity**

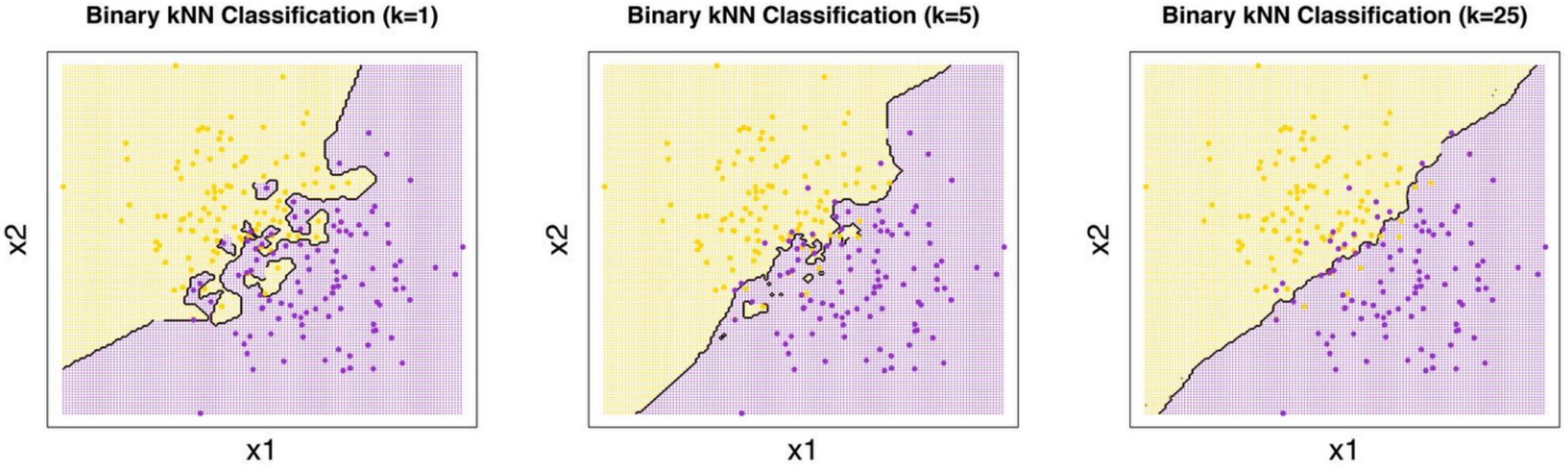


More wiggle
smaller *k*

Choosing *k*

- **Nilai terbaik** dari k bergantung pada data.
- **Cross validation** dapat digunakan untuk membandingkan k
- Gunakan **nilai k yang besar**, tapi berikan bobot tersendiri untuk tetangga yang lebih dekat dengan titik >> **distance-weighted k-nearest neighbors**

Choosing *k*



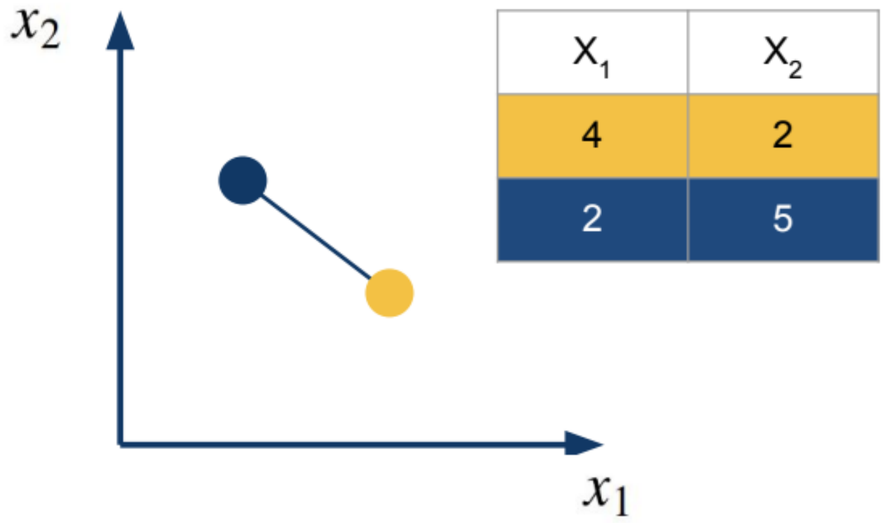
Curse of Dimensionality

Curse of Dimensionality

- Menariknya, **k-observasi** terdekat pada titik observasi “x”, mungkin menjadi **jauh** di dalam p-dimensional space ketika **p besar**

Curse of Dimensionality

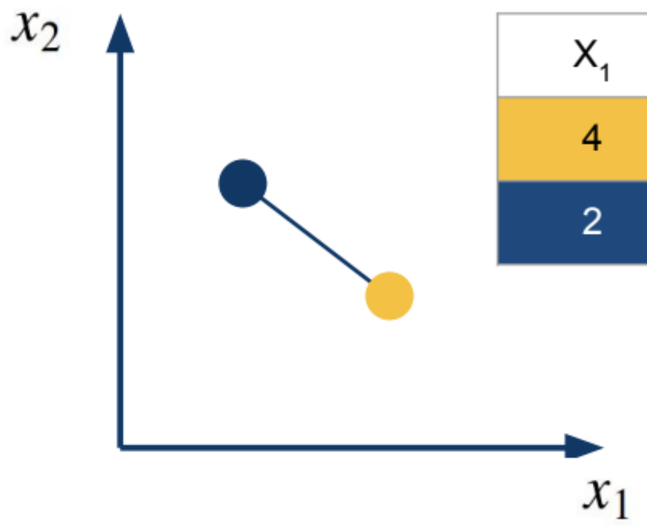
- Menariknya, **k-observasi** terdekat pada titik observasi “x”, mungkin menjadi **jauh** di dalam p-dimensional space ketika **p besar**



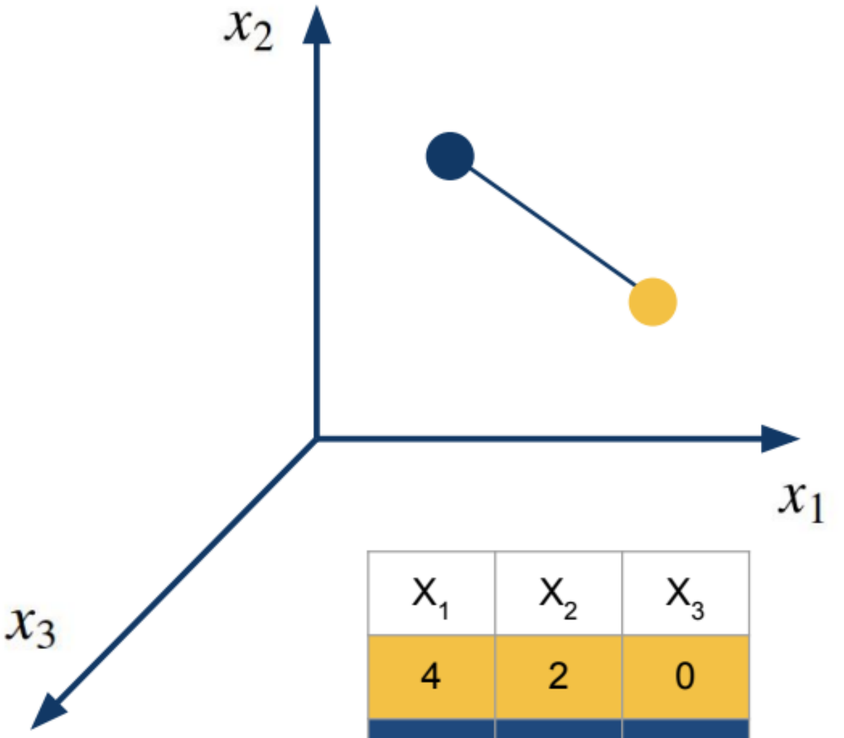
X_1	X_2
4	2
2	5

Curse of Dimensionality

- Menariknya, **k-observasi** terdekat pada titik observasi “x”, mungkin menjadi **jauh** di dalam p-dimensional space ketika **p besar**



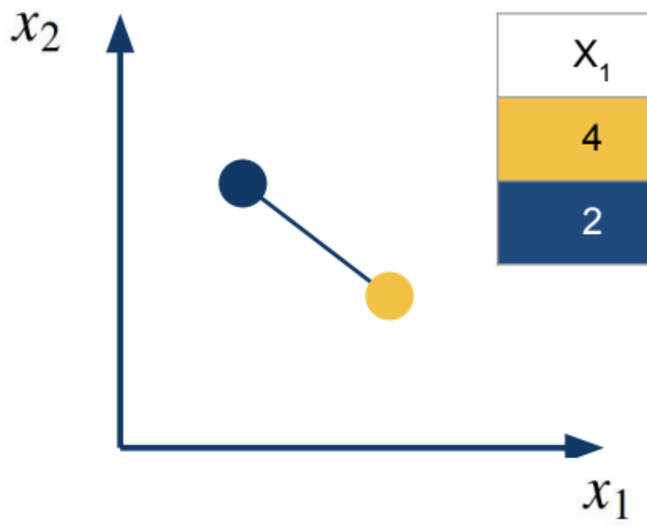
X_1	X_2
4	2
2	5



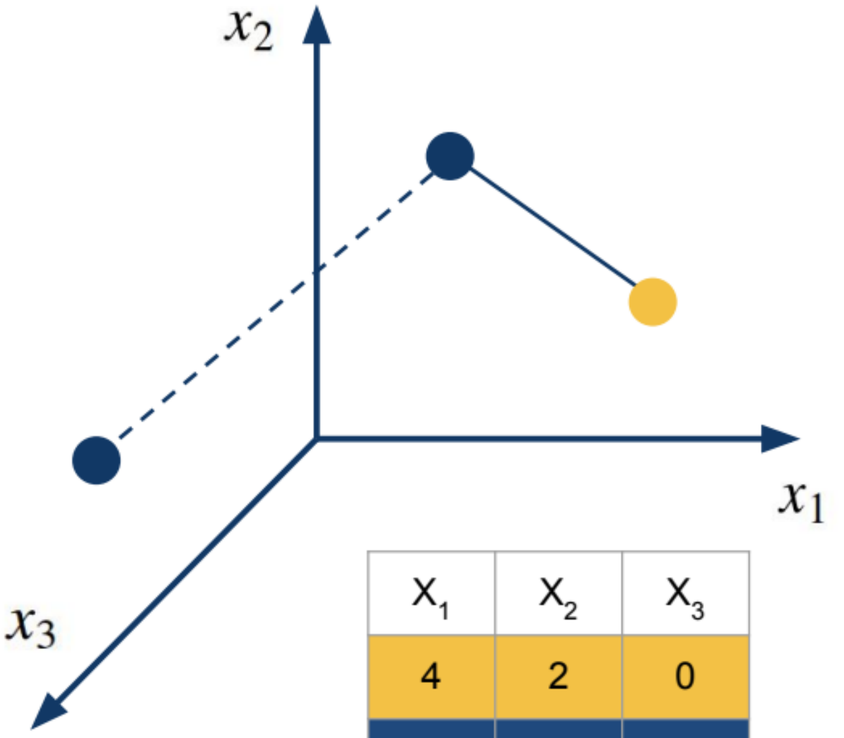
X_1	X_2	X_3
4	2	0
2	5	0

Curse of Dimensionality

- Menariknya, **k-observasi** terdekat pada titik observasi “x”, mungkin menjadi **jauh** di dalam p-dimensional space ketika **p besar**



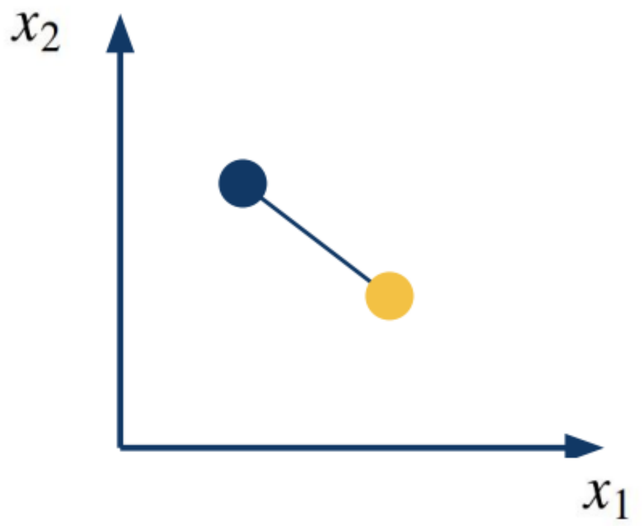
X_1	X_2
4	2
2	5

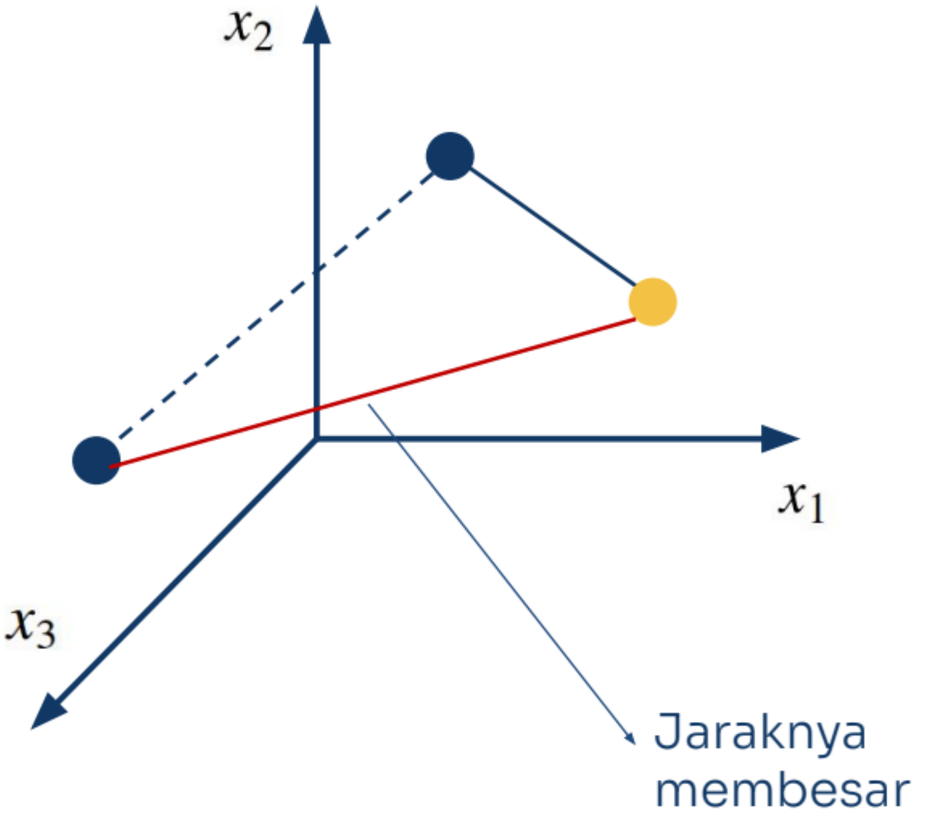


X_1	X_2	X_3
4	2	0
2	5	10

Curse of Dimensionality

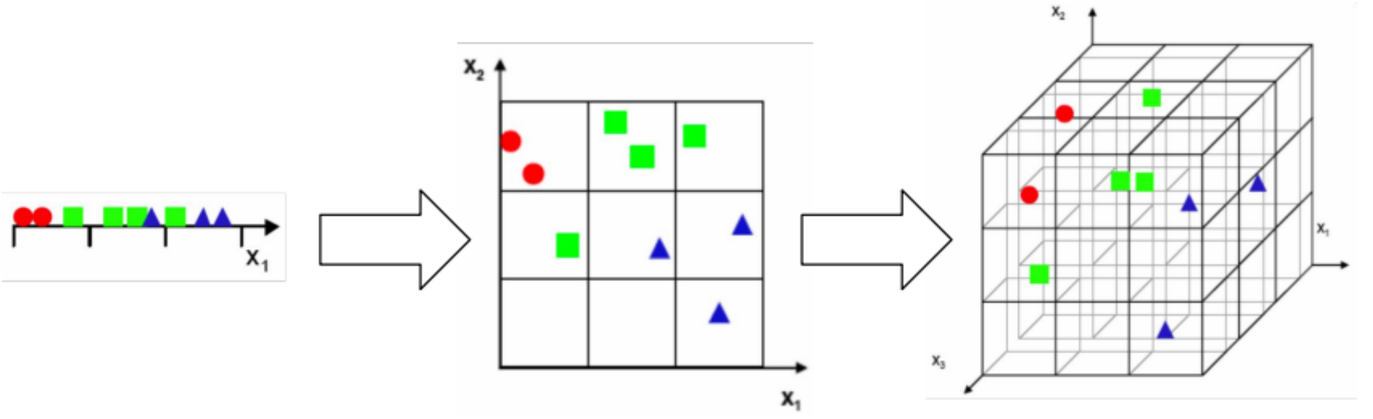
- Menariknya, **k-observasi** terdekat pada titik observasi “x”, mungkin menjadi **jauh** di dalam p-dimensional space ketika **p besar**





Curse of Dimensionality

- Dataset biasanya ada dalam dimensi tinggi.
- Saat dimensi bertambah: lebih sedikit observasi dapat ditemukan di tiap region-nya.
  - 1 Dimension : 3^1 region
  - 2 Dimension : 3^2 region
  - 3 Dimension : 3^3 region
  - n Dimension : 3^n region



Victor Lavrenko

Curse of Dimensionality

Solusi

- Feature selection
- More complex distance function
- Parametric methods (e.g. Linear Regression or Logistic Regression)

Victor Lavrenko

Thank You