The Data Science Process

1.) What question do you want to answer?

2.) Get the data

3.) Explore the data

4.) Clean the data

5.) **Machine Learning**

Explore the Data


STUDENT LOAN
APPLICATION
Personal Information
(Last)
(First)
(Middle Initial)
(City)
Home Telephone
Other Telephone

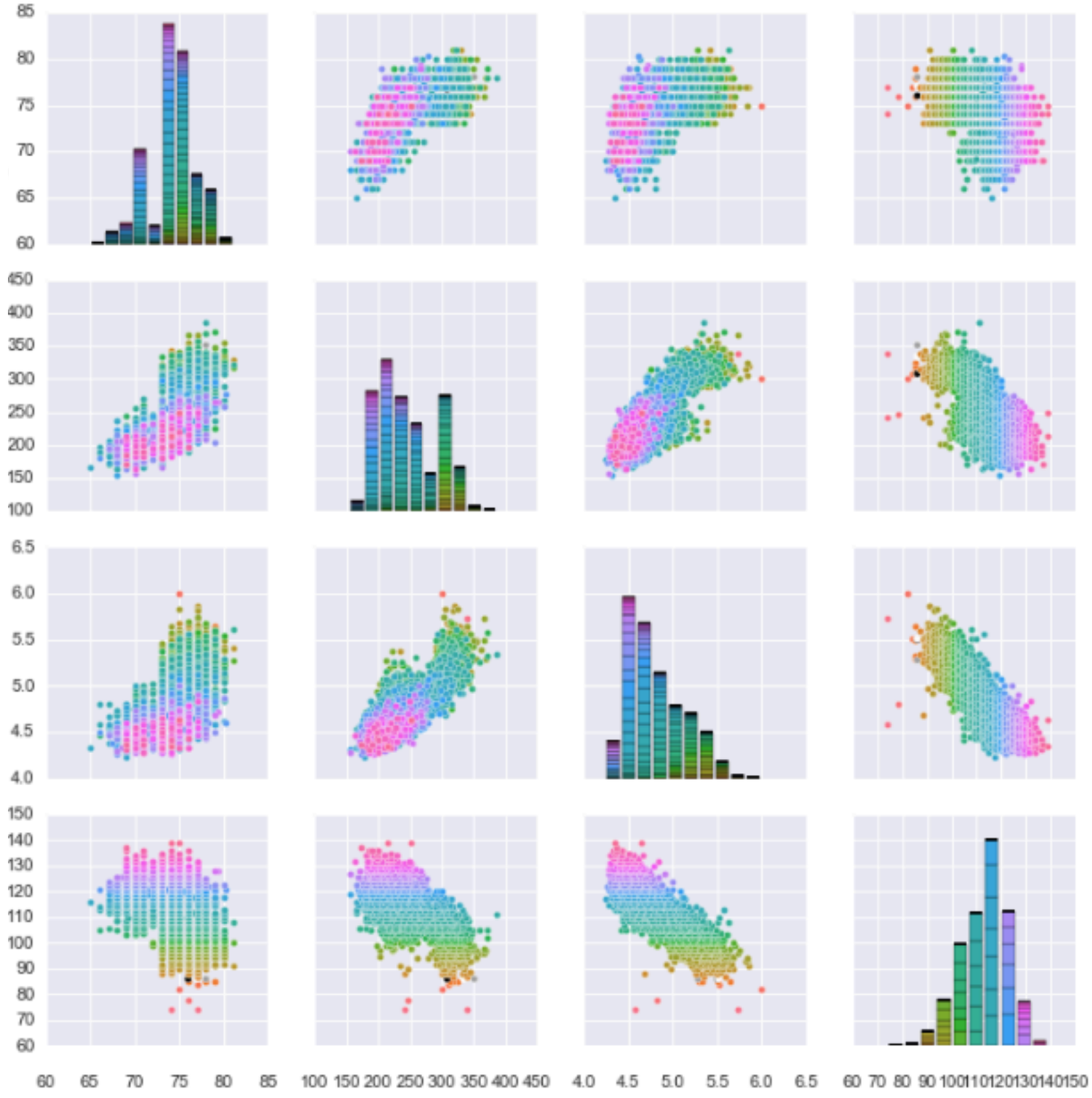
Student Loan Data

Age	Years Since Graduation	Income	Loan Balance	Next Payment	Defaulted
22	0	39247	14947.63	123	N
37	9	57981	27172.43	365	N
26	5	49031	18102.54	273	Y
42	16	63989	93429.79	491	Y
26	1	41250	13456.54	215	N
29	7	56482	29048.47	316	N
43	9	71202	26279.13	385	N
30	11	27649	6826.10	55	N

Student Loan Data Summary

	Age	Years Since Graduation	Income	Loan Balance	Next Payment
count	4947	4947	4947	4947	4933
mean	32.17	6.33	57981.98	37172.43	376.51
std	6.12	4.21	9031.62	18102.54	137.09
min	2	0.00	0.00	93.00	5.00
25%	26	3	41250.00	13456.00	112.00
50%	29	7	56482.00	39084.00	381.00
75%	43	9	71202.00	66212.00	485.00
max	67	31	276498.00	269910.00	755.00

Pair Plot


85
80
75
70
65
60
450
400
350
300
250
200
150
100
6.5
6.0
5.5
5.0
4.5
4.0
150
140
130
120
110
100
90
80
70
60
60 65 70 75 80 85
100 150 200 250 300 350 400 450
4.0 4.5 5.0 5.5 6.0 6.5
60 70 80 90 100 110 120 130 140 150

Clean the Data

Student Loan Data

Loan ID#	Graduated	Loan Type	Loan Balance	Next Payment	Months Delinquent	Defaulted
10148975	Yes	1	31567	327	0	N
19773966	Yes	3	27909	-	3	N
25220947	Yes	2	11,463	243	25	Y
17090812	No	2	29801	255	15	Y
23956341	Yes	3	18755	173	0	N
12680900	Yes	1	16,211	122	7	N
23435111	No	1	5064	84	0	N

Student Loan Data

Name	Graduated	Loan Type	Loan Balance	Next Payment	Months Delinquent	Defaulted
Ken	Yes	1	31567	327	0	N
Mary	Yes	3	27909	-	3	N
Reuben	Yes	2	11,463	243	25	Y
Amulya	No	2	29801	255	15	Y
Diane	Yes	3	18755	173	0	N
Timothy	Yes	1	16,211	122	7	N
Michelle	No	1	5064	84	0	N

Student Loan Data

Graduated	Loan Type	Loan Balance	Next Payment	Months Delinquent	Defaulted
Yes	1	31567	327	0	N
Yes	3	27909	-	3	N
Yes	2	11,463	243	25	Y
No	2	29801	255	15	Y
Yes	3	18755	173	0	N
Yes	1	16,211	122	7	N
No	1	5064	84	0	N

Student Loan Data

Graduated	Loan Type	Loan Balance	Next Payment	Months Delinquent	Defaulted
Yes	1	31567	327	0	N
Yes	3	27909	-	3	N
Yes	2	11463	243	25	Y
No	2	29801	255	15	Y
Yes	3	18755	173	0	N
Yes	1	16211	122	7	N
No	1	5064	84	0	N

Student Loan Data

Graduated	Loan Type	Loan Balance	Next Payment	Months Delinquent	Defaulted
Yes	1	31567	327	0	N
Yes	3	27909	NaN	3	N
Yes	2	11463	243	25	Y
No	2	29801	255	15	Y
Yes	3	18755	173	0	N
Yes	1	16211	122	7	N
No	1	5064	84	0	N

Student Loan Data

Graduated	Loan Type	Loan Balance	Next Payment	Months Delinquent	Defaulted
Yes	1	31567	327	0	N
Yes	3	27909	200	3	N
Yes	2	11463	243	25	Y
No	2	29801	255	15	Y
Yes	3	18755	173	0	N
Yes	1	16211	122	7	N
No	1	5064	84	0	N

Transform the Data

Student Loan Data

Graduated	Loan Type	Loan Balance	Next Payment	Months Delinquent	Defaulted
Yes	1	31567	327	0	N
Yes	3	27909	200	3	N
Yes	2	11463	243	25	Y
No	2	29801	255	15	Y
Yes	3	18755	173	0	N
Yes	1	16211	122	7	N
No	1	5064	84	0	N

Student Loan Data

Graduated	Loan Type	Loan Balance	Next Payment	Months Delinquent	Defaulted
1	1	31567	327	0	N
1	3	27909	200	3	N
1	2	11463	243	25	Y
0	2	29801	255	15	Y
1	3	18755	173	0	N
1	1	16211	122	7	N
0	1	5064	84	0	N

One-Hot Encoding

Student Loan Data

Graduated	Loan 1	Loan 2	Loan 3	Loan Balance	Next Payment	Months Delinquent	Defaulted
1	1	0	0	31567	327	0	N
1	0	0	1	27909	200	3	N
1	0	1	0	11463	243	25	Y
0	0	1	0	29801	255	15	Y
1	0	0	1	18755	173	0	N
1	1	0	0	16211	122	7	N
0	1	0	0	5064	84	0	N

Student Loan Data

Graduated	Loan 1	Loan 2	Loan 3	Loan Balance	Next Payment	Months Delinquent	Defaulted
1	1	0	0	31567	327	0	0
1	0	0	1	27909	200	3	0
1	0	1	0	11463	243	25	1
0	0	1	0	29801	255	15	1
1	0	0	1	18755	173	0	0
1	1	0	0	16211	122	7	0
0	1	0	0	5064	84	0	0

DataFrame

Student Loan Data

Graduated	Loan Balance	Next Payment	Months Delinquent	Defaulted
1	31567	327	0	0
1	27909	200	3	0
1	11463	243	25	1
0	29801	255	15	1
1	18755	173	0	0
1	16211	122	7	0
0	5064	84	0	0
0	17198	154	0	0
1	21309	201	0	0
0	14693	193	4	1

Student Loan Data

features → (X)

Graduated	Loan Balance	Next Payment	Months Delinquent	Defaulted
1	31567	327	0	0
1	27909	200	3	0
1	11463	243	25	1
0	29801	255	15	1
1	18755	173	0	0
1	16211	122	7	0
0	5064	84	0	0
0	17198	154	0	0
1	21309	201	0	0
0	14693	193	4	1

Student Loan Data

Graduated	Loan Balance	Next Payment	Months Delinquent	Defaulted
1	31567	327	0	0
1	27909	200	3	0
1	11463	243	25	1
0	29801	255	15	1
1	18755	173	0	0
1	16211	122	7	0
0	5064	84	0	0
0	17198	154	0	0
1	21309	201	0	0
0	14693	193	4	1

← **target** (y)

Student Loan Data

Graduated	Loan Balance	Next Payment	Months Delinquent	Defaulted
1	31567	327	0	0
1	27909	200	3	0
1	11463	243	25	1
0	29801	255	15	1
1	18755	173	0	0
1	16211	122	7	0
0	5064	84	0	0
0	17198	154	0	0
1	21309	201	0	0
0	14693	193	4	1

sample →

Student Loan Data

Graduated	Loan Balance	Next Payment	Months Delinquent	Defaulted
1	31567	327	0	0
1	27909	200	3	0
1	11463	243	25	1
0	29801	255	15	1
1	18755	173	0	0
1	16211	122	7	0
0	5064	84	0	0
0	17198	154	0	0
1	21309	201	0	0
0	14693	193	4	1

Student Loan Data

1	31567	327	0	0
1	27909	200	3	0
1	11463	243	25	1
0	29801	255	15	1
1	18755	173	0	0
1	16211	122	7	0
0	5064	84	0	0
0	17198	154	0	0
1	21309	201	0	0
0	14693	193	4	1

Student Loan Data

1	31567	327	0	0
1	27909	200	3	0
1	11463	243	25	1
0	29801	255	15	1

X =
features

1	18755	173	0
1	16211	122	7
0	5064	84	0
0	17198	154	0
1	21309	201	0
0	14693	193	4

y =
target

0
0
0
0
0
1

Pandas Intro

Matplotlib Intro

Machine Learning

Test/Train Split

DATA SET

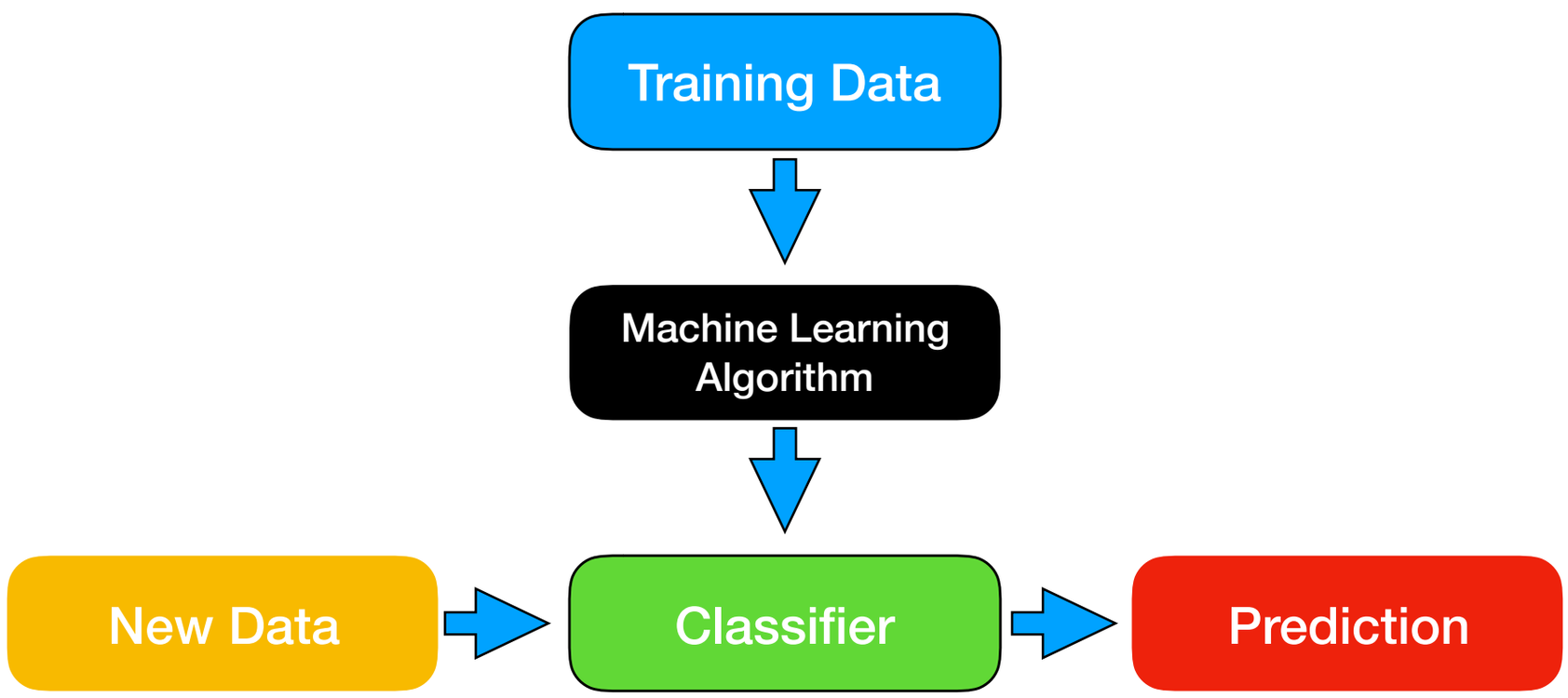
Data

100%

DATA SET

DATA SET

Learning Process



Common Machine Learning Algorithms

Linear Regression

Logistic Regression

Support Vector Machine

Decision Tree

K-Nearest Neighbor

K-Means

Neural Network

$$\hat{f}(X)$$

The Prediction

output

input

$$\hat{y} \equiv \hat{f}(x)$$

The Prediction

continuous value

output

$$\hat{y} = 66.5$$

The Prediction

probability

output

$$\hat{y} = .85$$

The Prediction

class membership

output

$$\hat{y} = 1$$

The Prediction

class membership

output

$$\hat{y} = 0$$

Linear Regression

equation of a line

$$y = mx + b$$

Linear Regression

$$y = \beta_0 + \beta_1 x$$

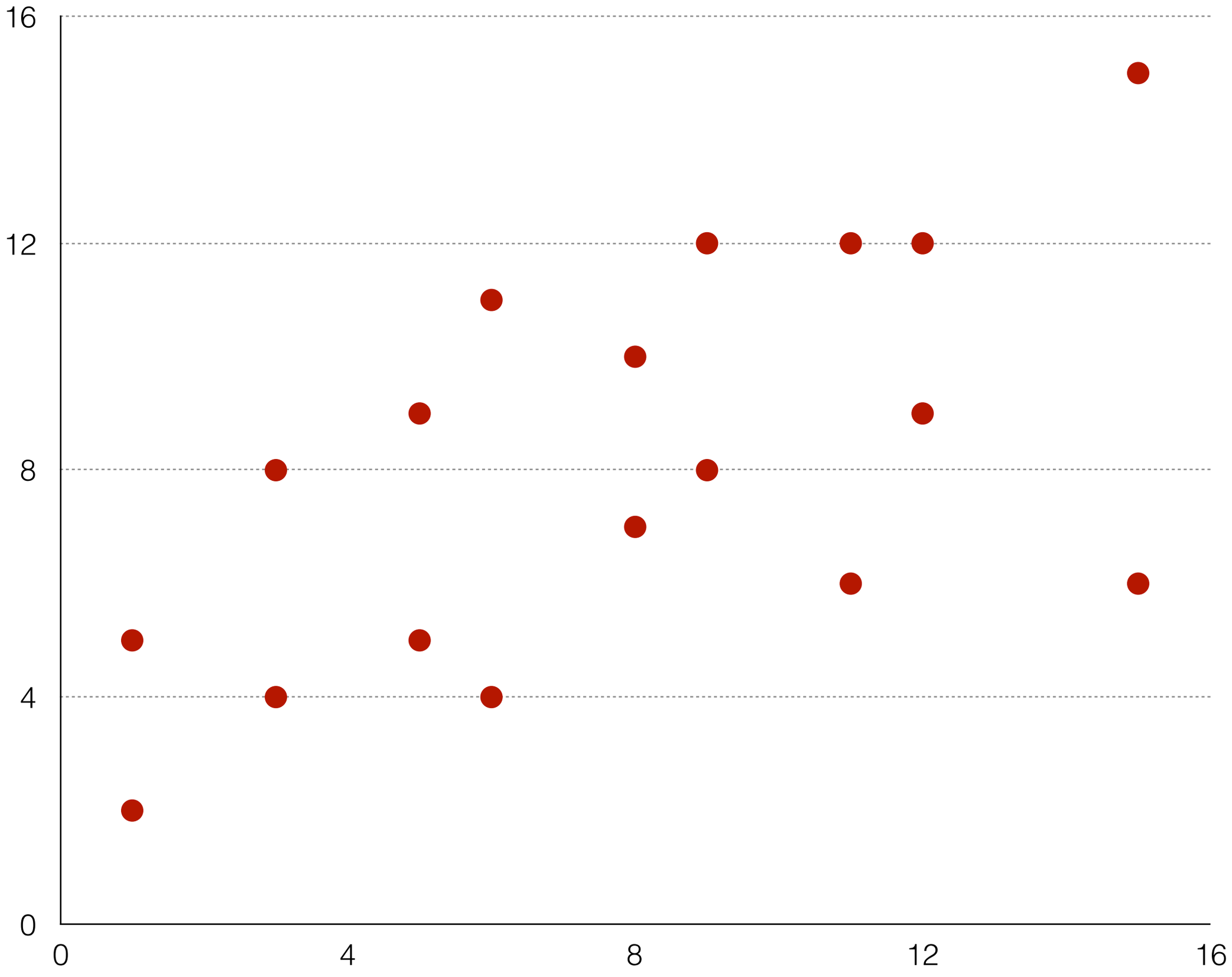
equation of a line

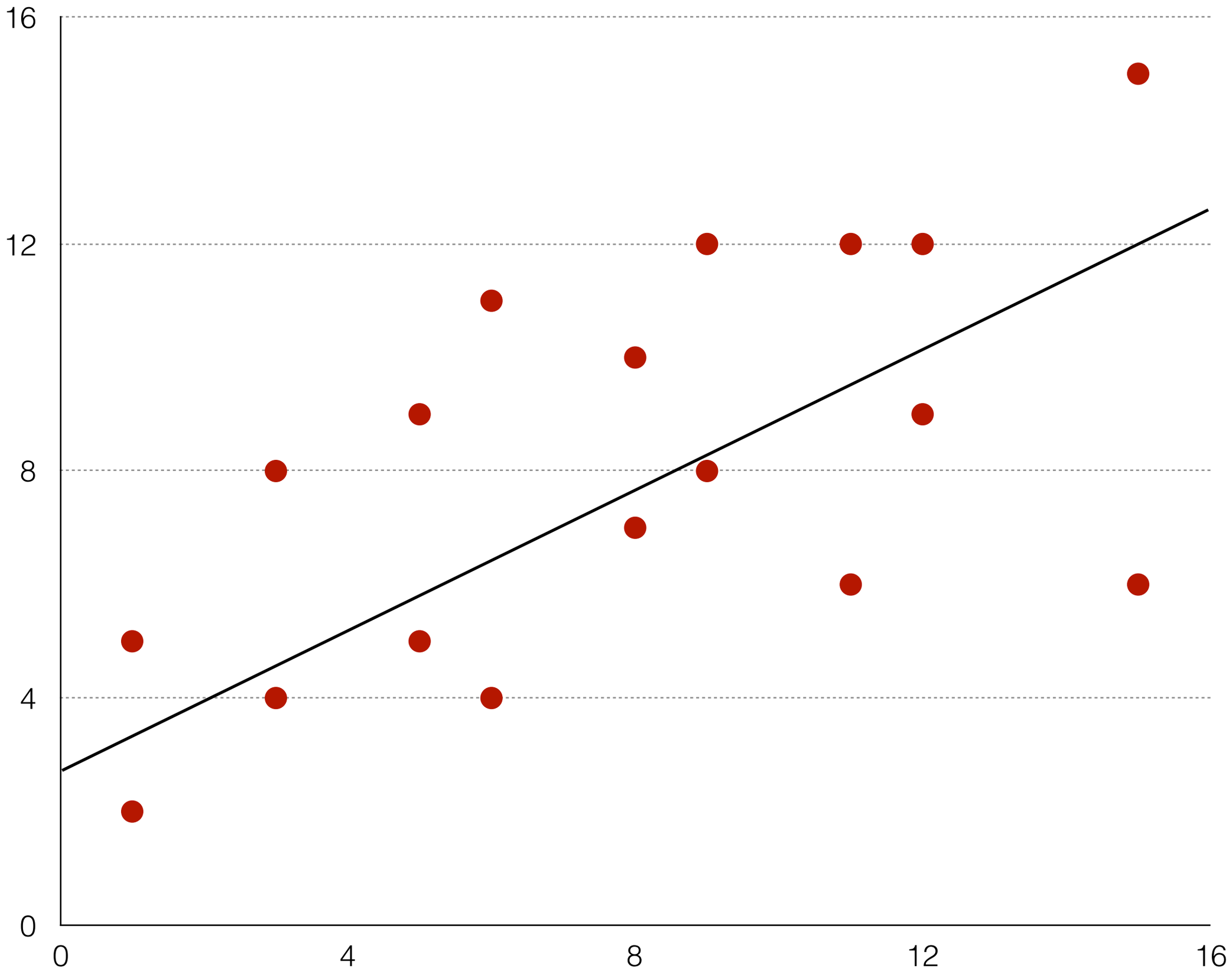
$$y = mx + b$$

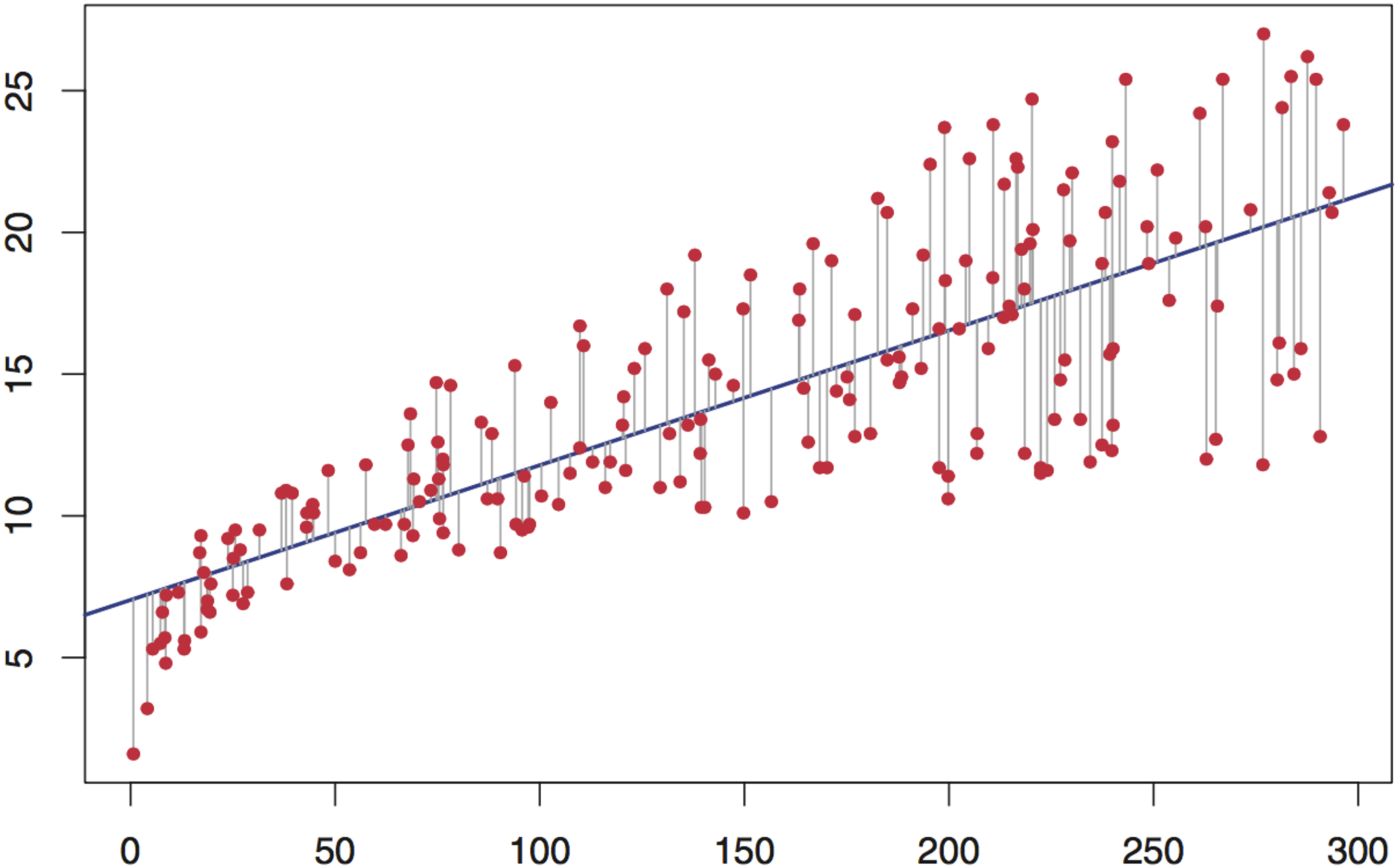
Simple Linear Regression

Input	learned coefficients (weights)	Output
x_1	β_0, β_1	y

$$y = \beta_0 + \beta_1 x_1$$


16
12
8
4
0
0
4
8
12
16


16
12
8
4
0
0
4
8
12
16


25
20
15
10
5
0
50
100
150
200
250
300

Multiple Linear Regression

$$LOC = 227.63 + 9.51x_1 + 2.7x_2 - 7.08x_3$$

x_1 = hour pair programming

x_2 = gender (m = 0; f = 1)

x_3 = number of social accounts

Multiple Linear Regression

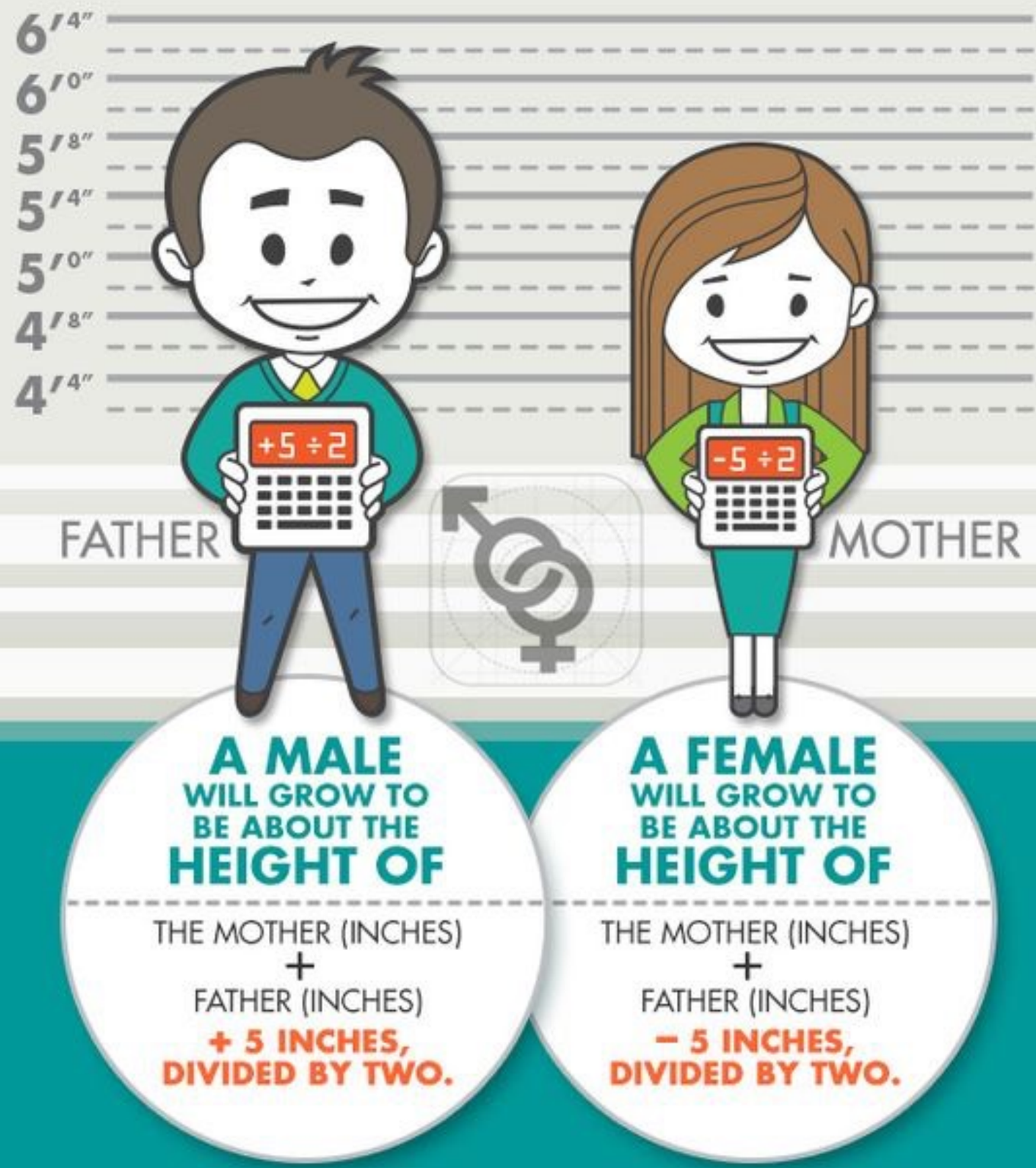
$$\text{Apps sold} = 46.55 + 35.03x_1 + 27.11x_2 + 52.48x_3$$

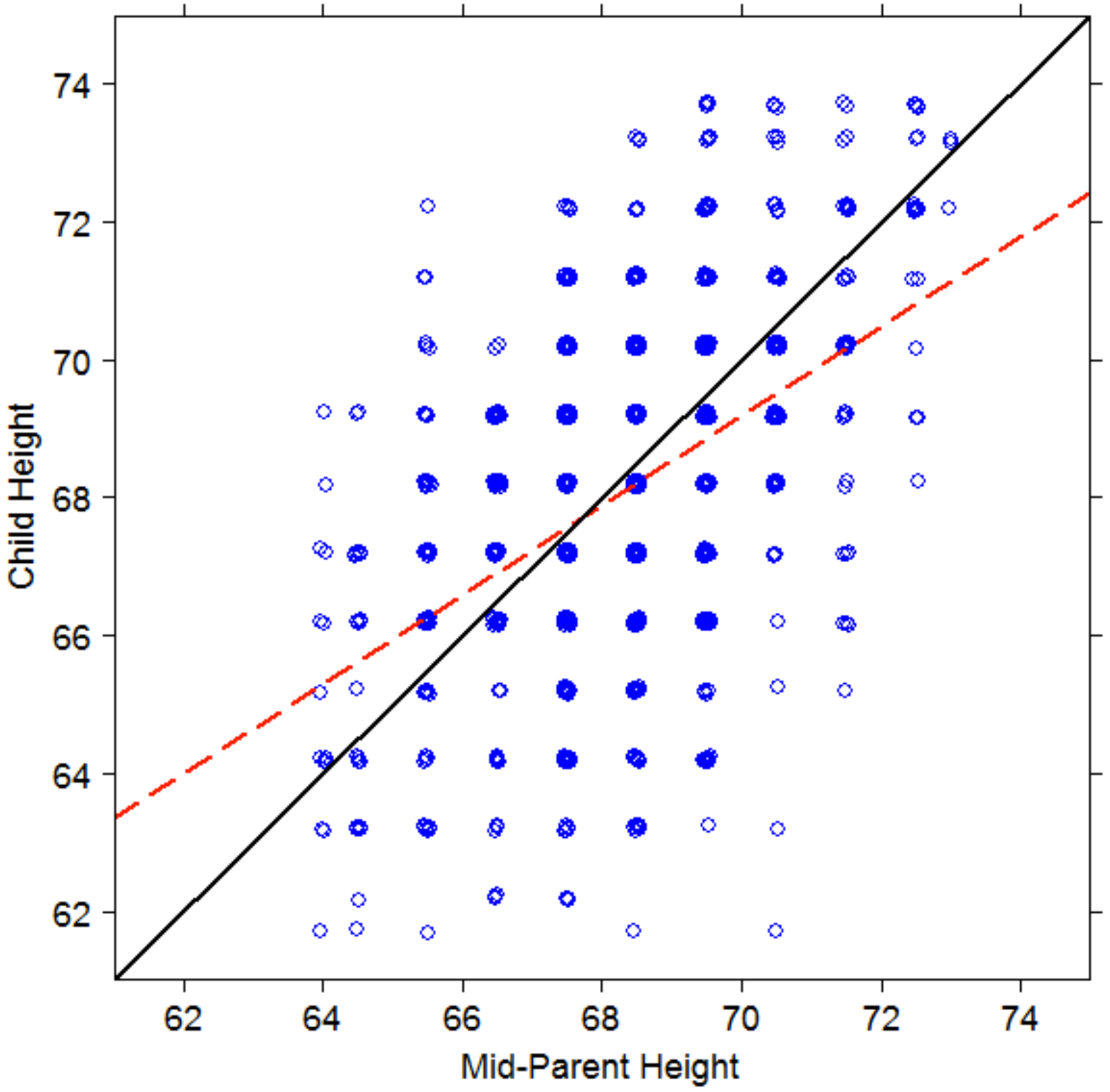
x_1 = per $100 advertising

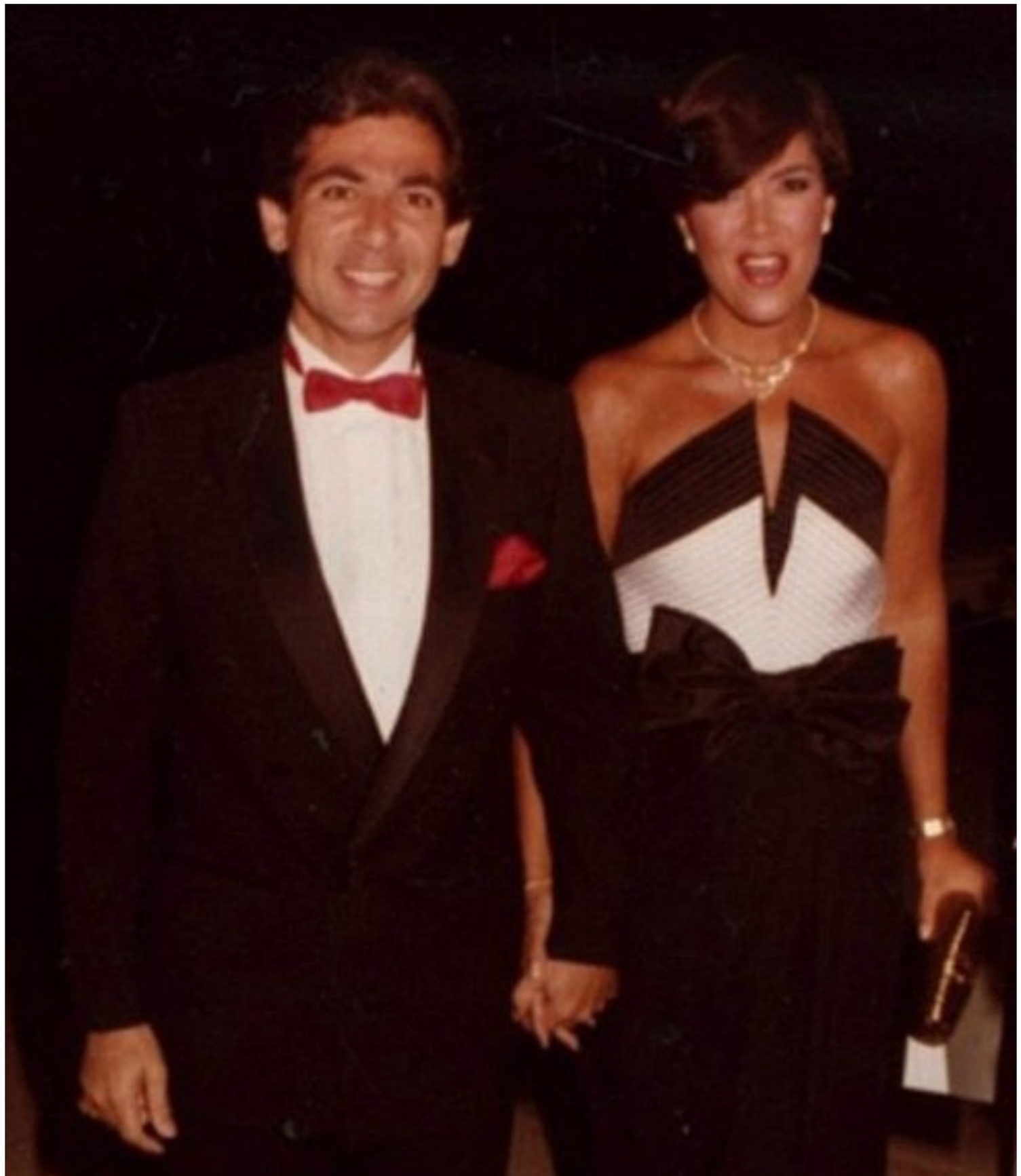
x_2 = per public talk

x_3 = per targeted podcast

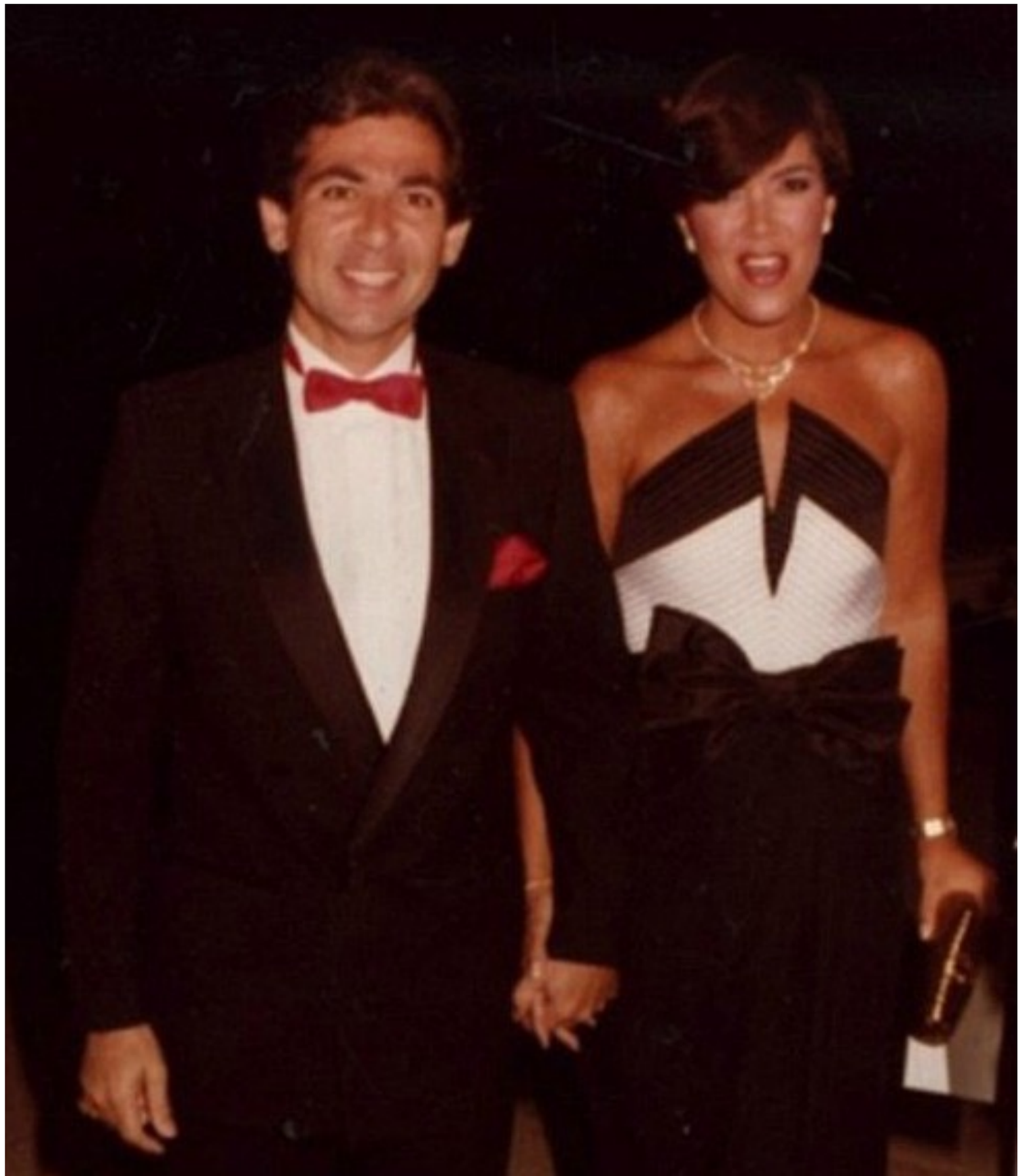
HEIGHT


6′4″
6′0″
5′8″
5′4″
5′0″
4′8″
4′4″
+5 ÷2
-5 ÷2
FATHER
MOTHER
A MALE
WILL GROW TO
BE ABOUT THE
HEIGHT OF
THE MOTHER (INCHES)
+
FATHER (INCHES)
+ 5 INCHES,
DIVIDED BY TWO.
A FEMALE
WILL GROW TO
BE ABOUT THE
HEIGHT OF
THE MOTHER (INCHES)
+
FATHER (INCHES)
− 5 INCHES,
DIVIDED BY TWO.

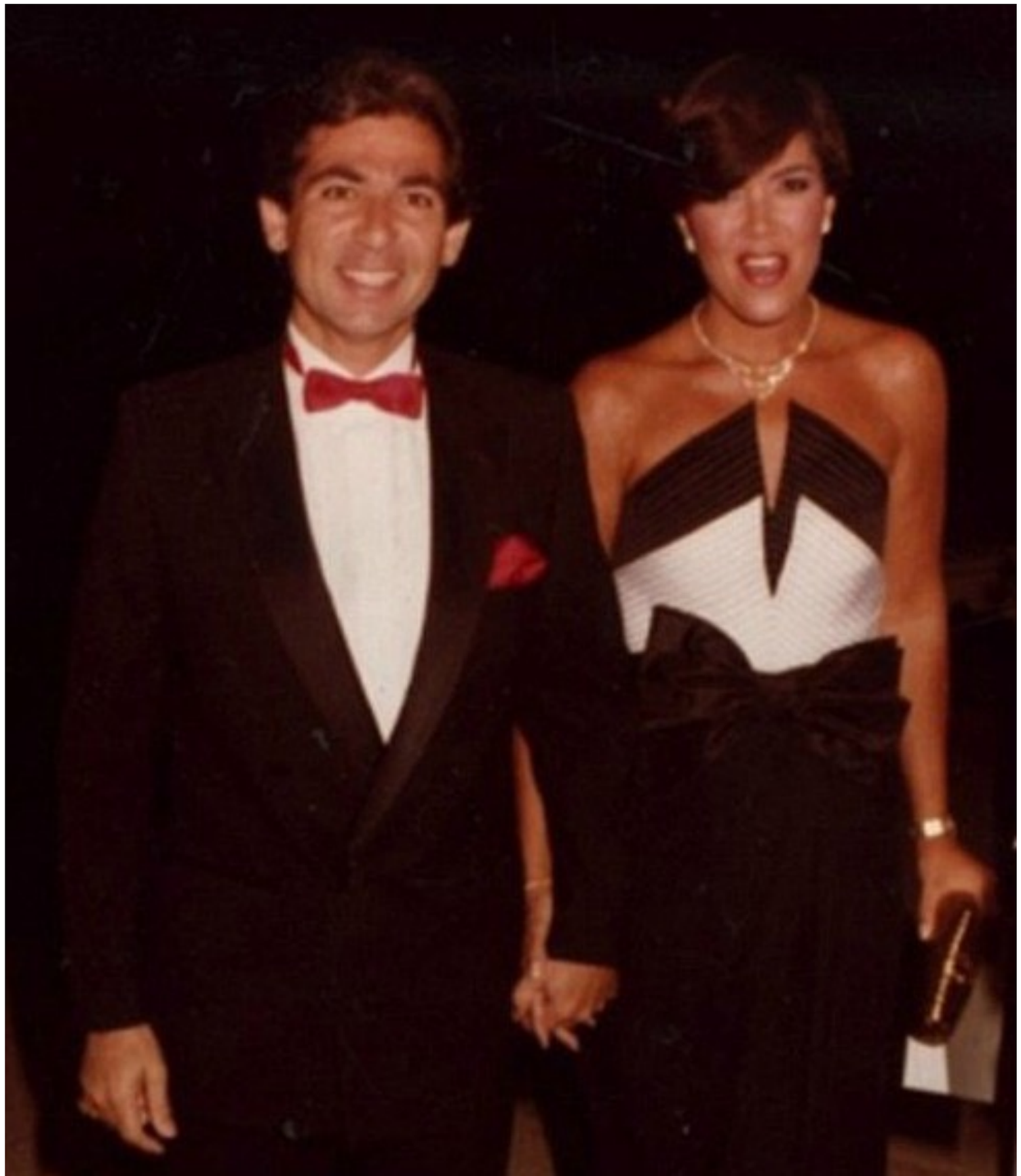

74
72
70
68
66
64
62
Child Height
62
64
66
68
70
72
74
Mid-Parent Height



5’7 —



5’7

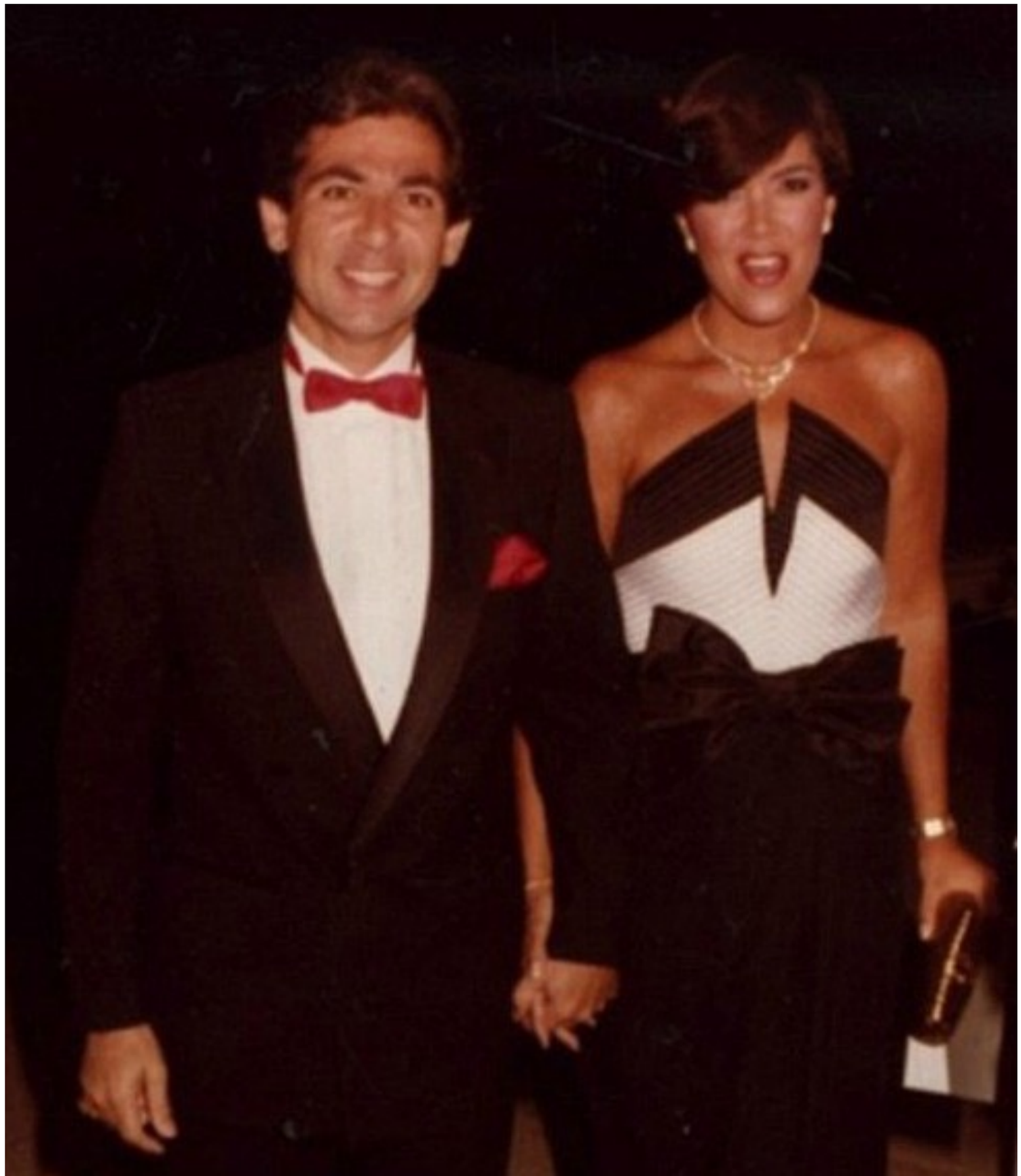


5’6

Mid-parent Height

5’6.5

5’7



5’6

5’0

5’3

5’10

5’0

5’3

5’10

Average

5’4 1/3

6’1

5’3

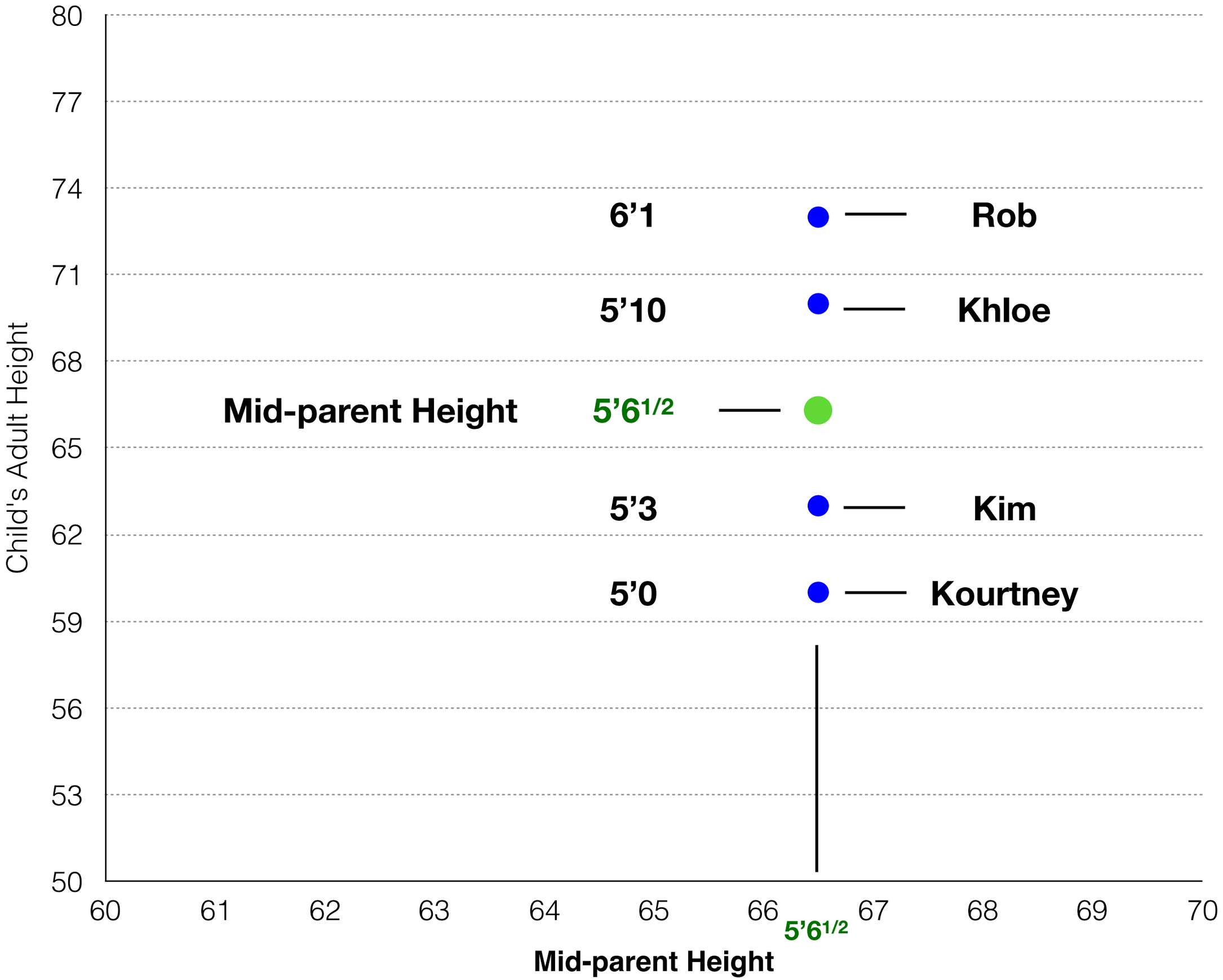
5’0

5’10


Nicole
by O·P·I
KARDASHIAN KHAOS
LAS VEGAS
mirage
Vegas Starts Here

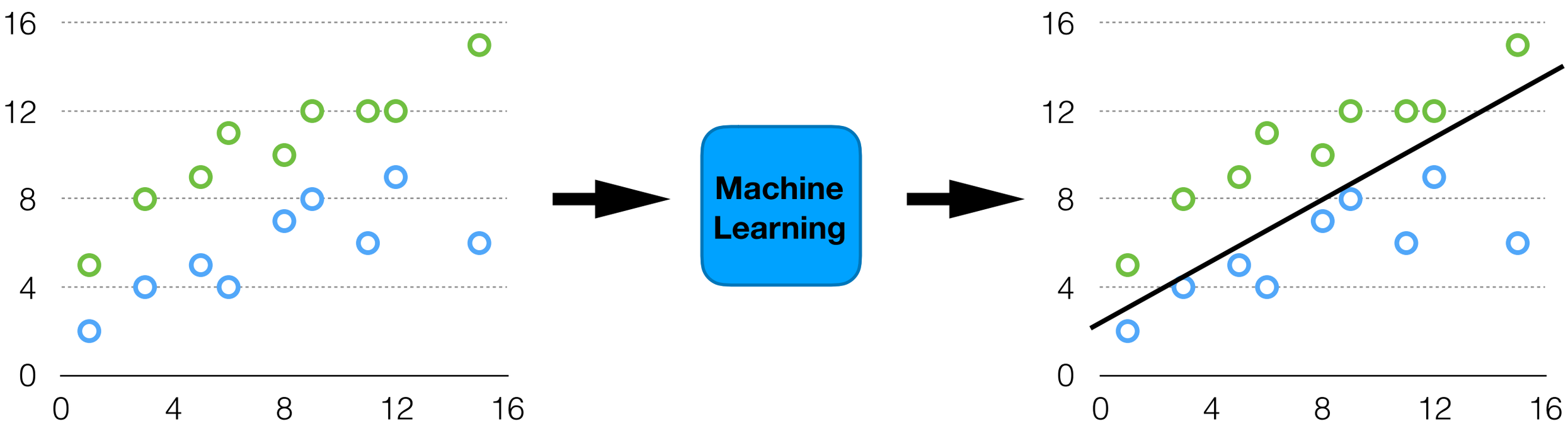
Mid-parent Height

5’6.5


Child's Adult Height
80
77
74
71
68
65
62
59
56
53
50
6’1
Rob
5’10
Khloe
Mid-parent Height
5’6 1/2
5’3
Kim
5’0
Kourtney
60
61
62
63
64
65
66
67
68
69
70
5’6 1/2
Mid-parent Height

CLASSIFICATION

What is a classifier?



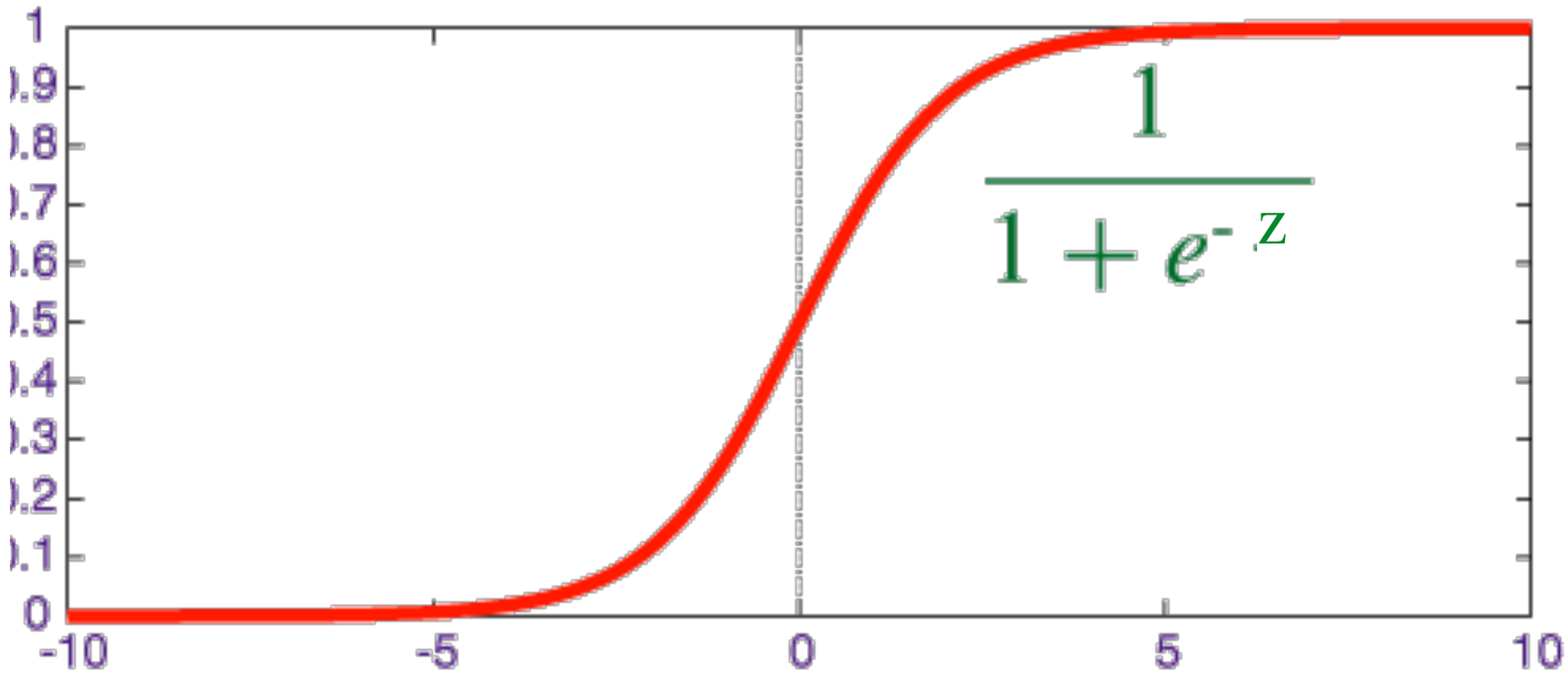
Logistic Regression

Logistic Regression

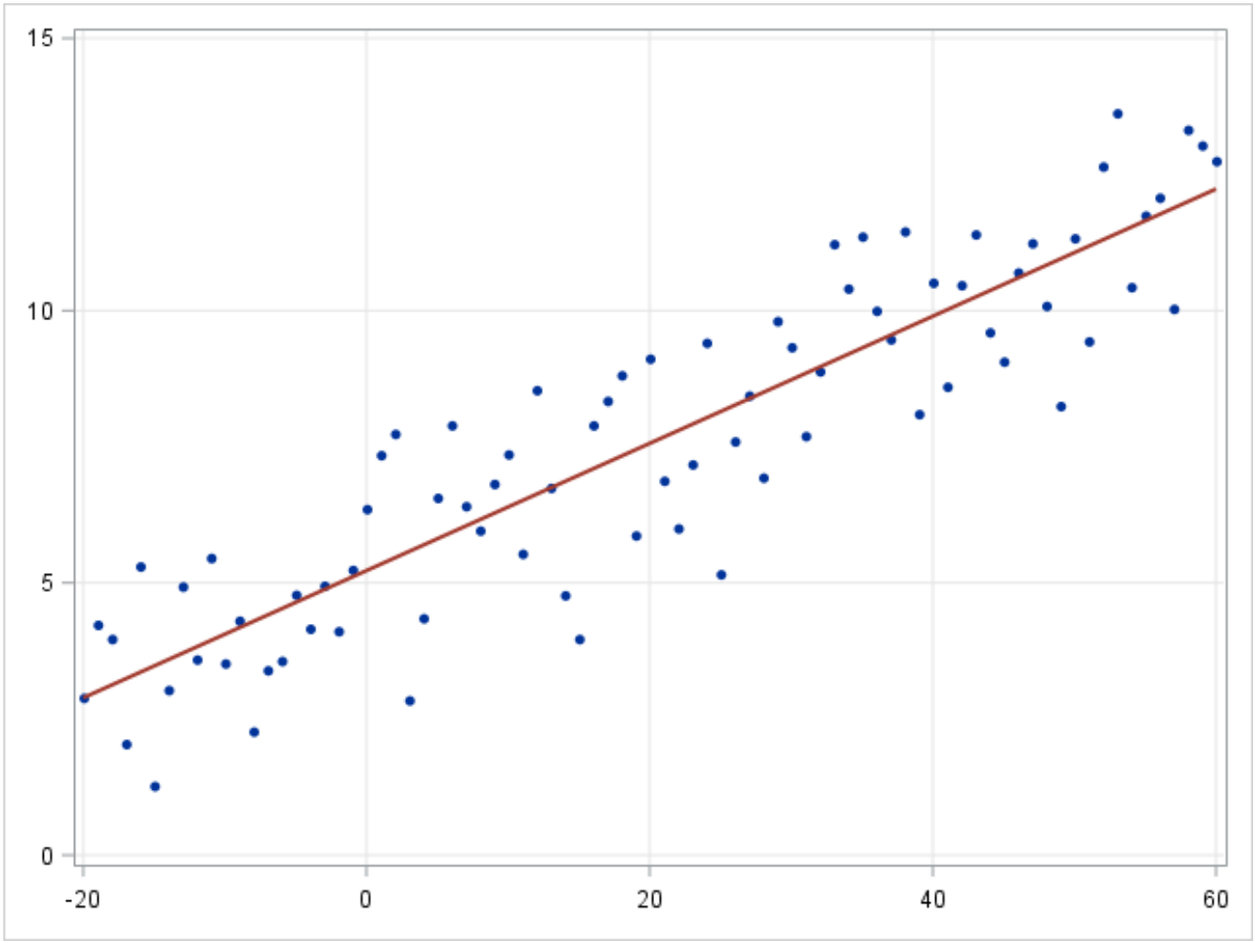
intermediate step

$$z = \beta_0 + \beta_1 x + \beta_2 x + \beta_3 x$$

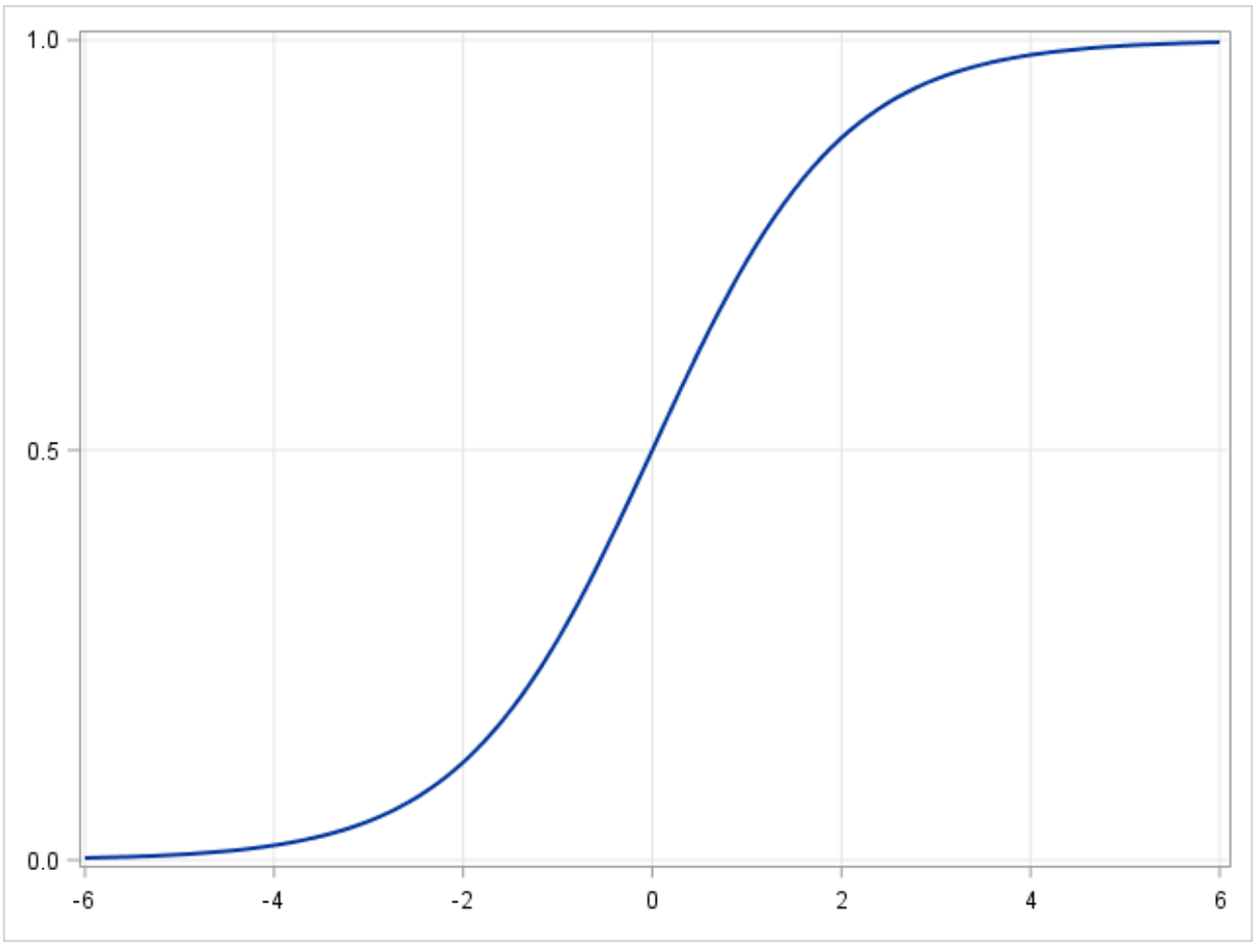
Logistic (Sigmoid) Function



Linear Regression



Logistic Regression



Logistic Regression

$$\hat{f}(X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2)}}$$

positive class = 1

negative class = 0

if probability >= 0.5: predict 1

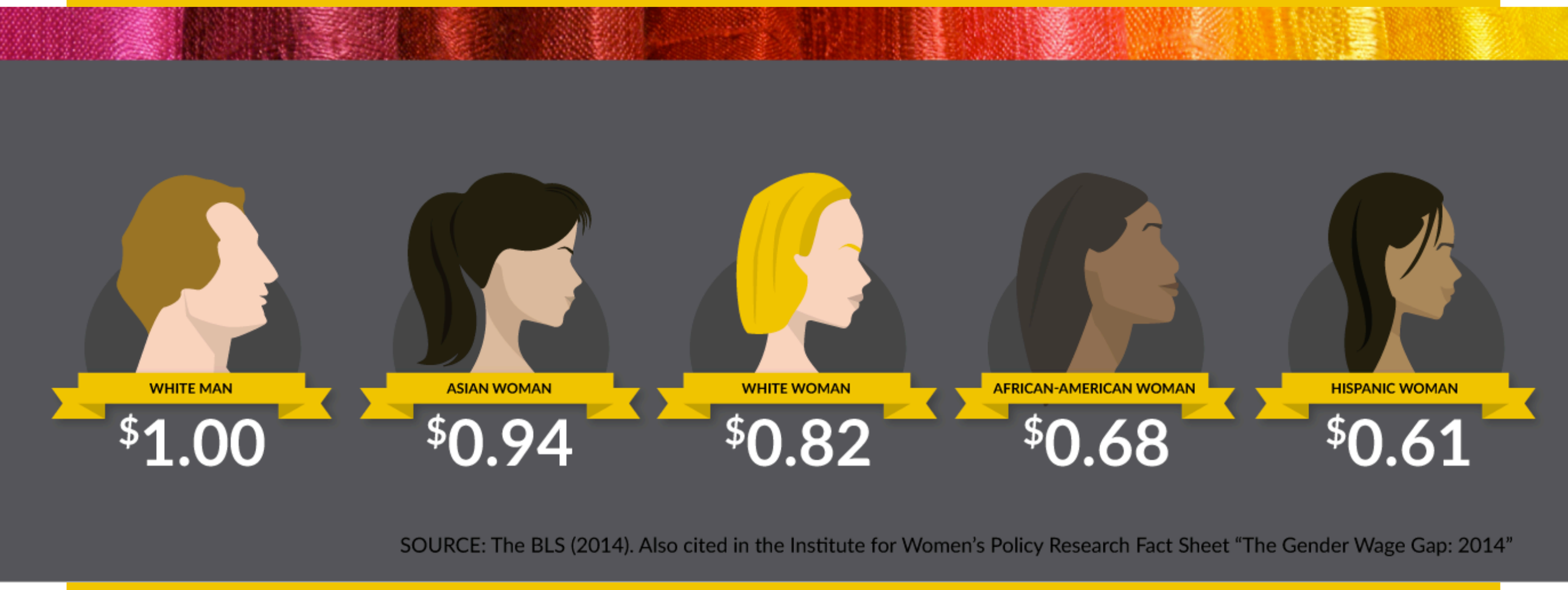
else: predict 0

U.S. Pay Gap

U.S. pay gap:
All full-time working men vs. women

SHUTTERSTOCK / CNNMONEY

U.S. pay gap:
All full-time working men and women



SOURCE: The BLS (2014). Also cited in the Institute for Women's Policy Research Fact Sheet "The Gender Wage Gap: 2014"

The highest paid occupations for women

Chief executives

Pharmacists

Lawyers

Computer and information systems managers

Physicians and surgeons

Nurse practitioners

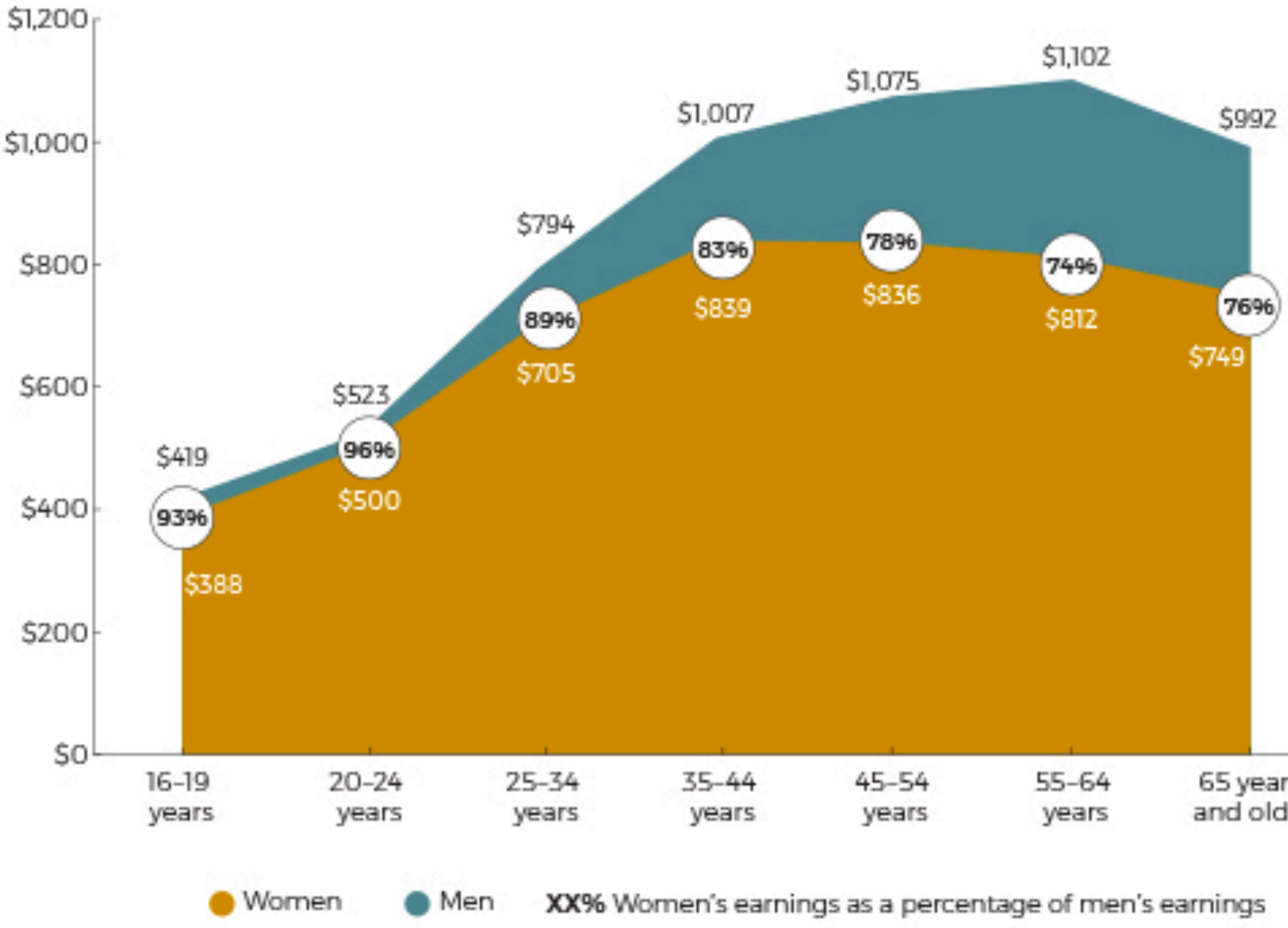
Engineers

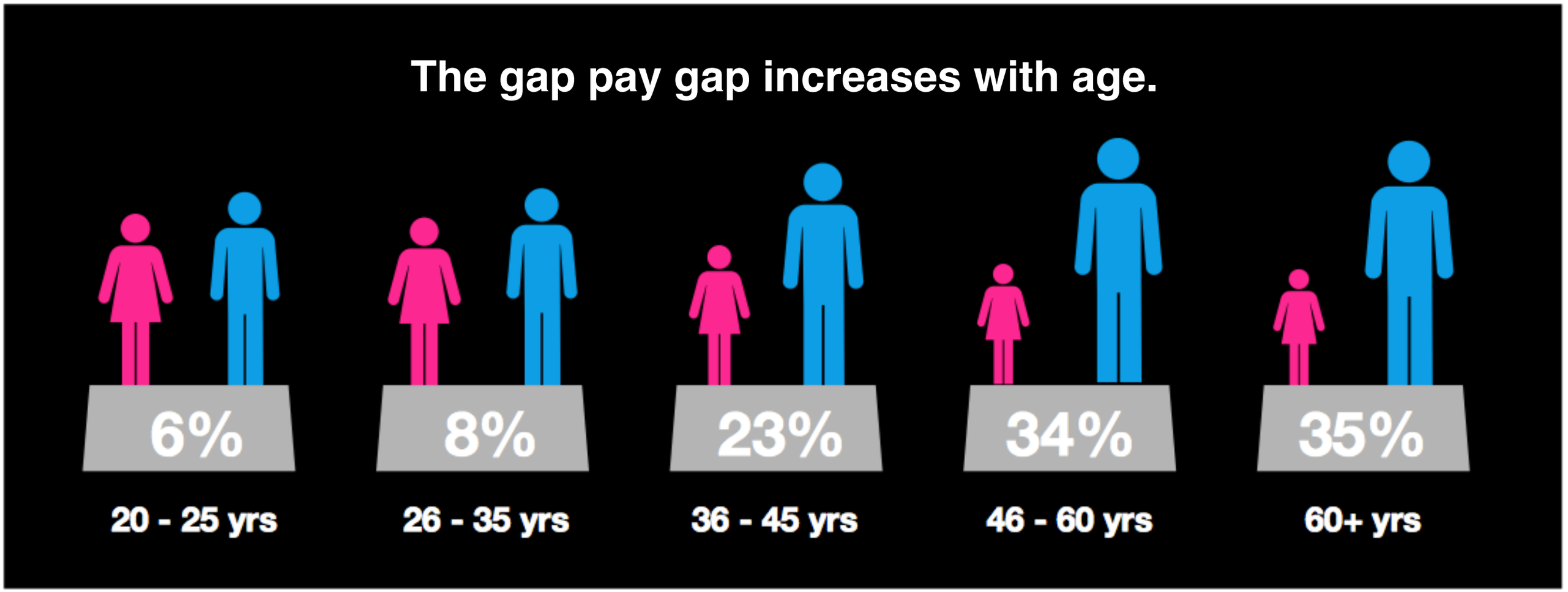
Software developers

Management analysts

Operations research analysts

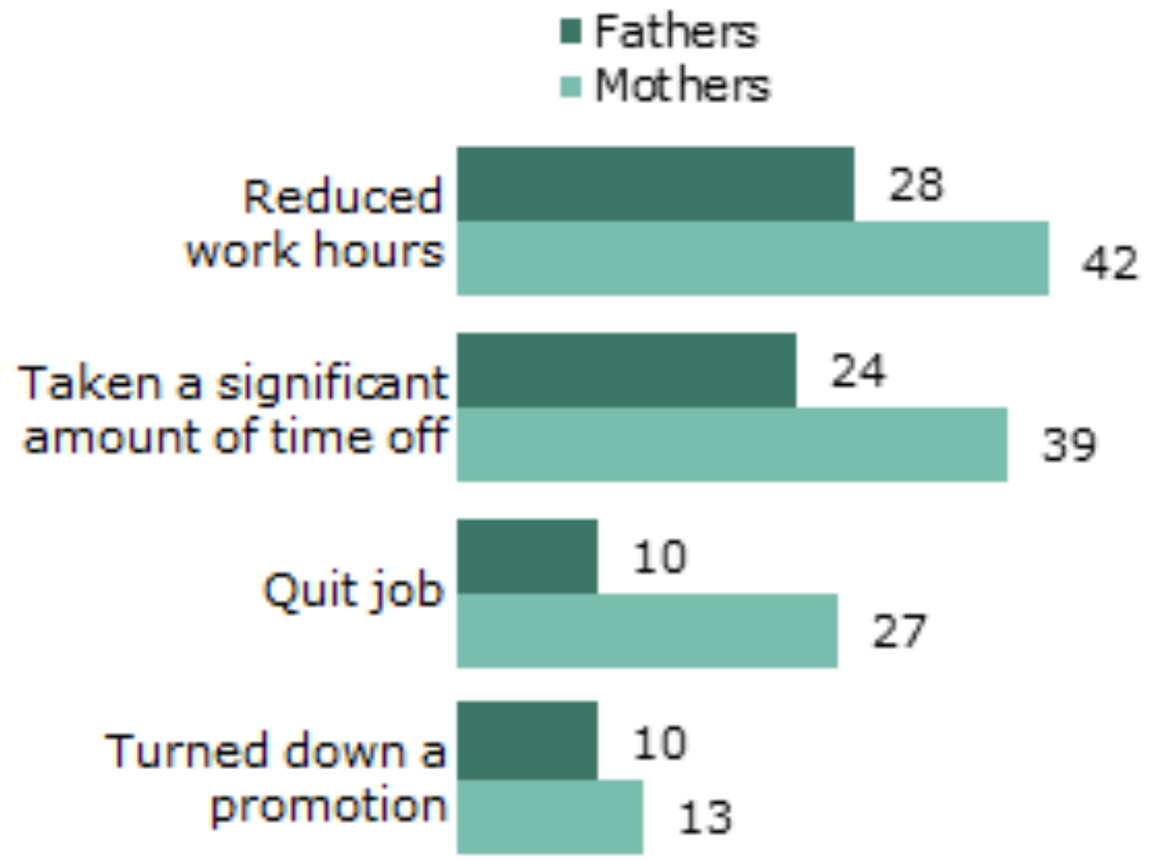
Median Weekly Earnings, by Age and Gender, 2016




The gap pay gap increases with age.
6%
8%
23%
34%
35%
20 - 25 yrs
26 - 35 yrs
36 - 45 yrs
46 - 60 yrs
60+ yrs

Mothers, More than Fathers, Experience Career Interruptions

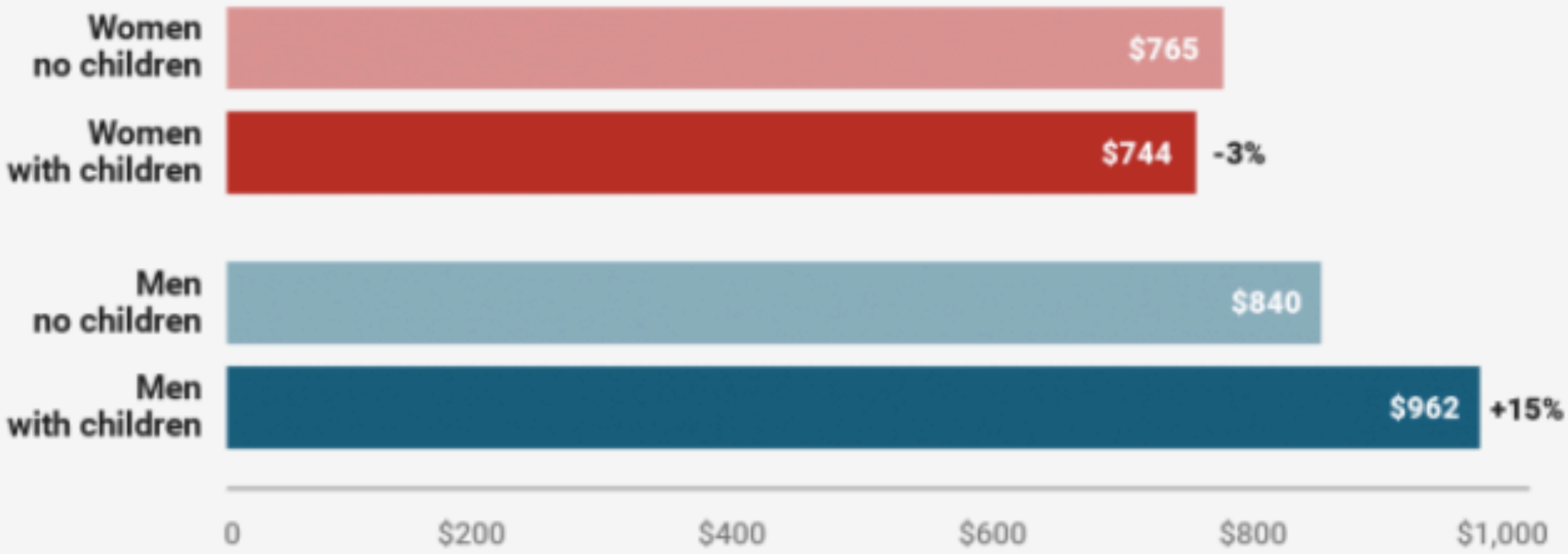
% saying they have ... in order to care for a child or family member



Notes: Based on those who have ever worked. "Fathers" and "mothers" include those with children of any age, including adult children (n=1,254).

PEW RESEARCH CENTER Q56a-d

WEEKLY EARNINGS FOR WOMEN WITH/WITHOUT CHILDREN



Data

Race 1	Race 2	Race 3	Race 4	Top Paying	Age	Children	Pay Gap
1	3	0	1	0	49	1	1
0	2	0	0	1	43	0	0
0	2	1	0	0	37	0	0
0	4	0	1	0	28	1	1
1	4	0	1	0	25	1	0
1	2	0	1	0	23	0	0
0	3	1	0	0	48	0	1
1	2	0	0	1	33	0	1
1	4	0	1	0	39	1	1
0	4	0	1	0	47	1	0

Pay Gap Decision Boundary

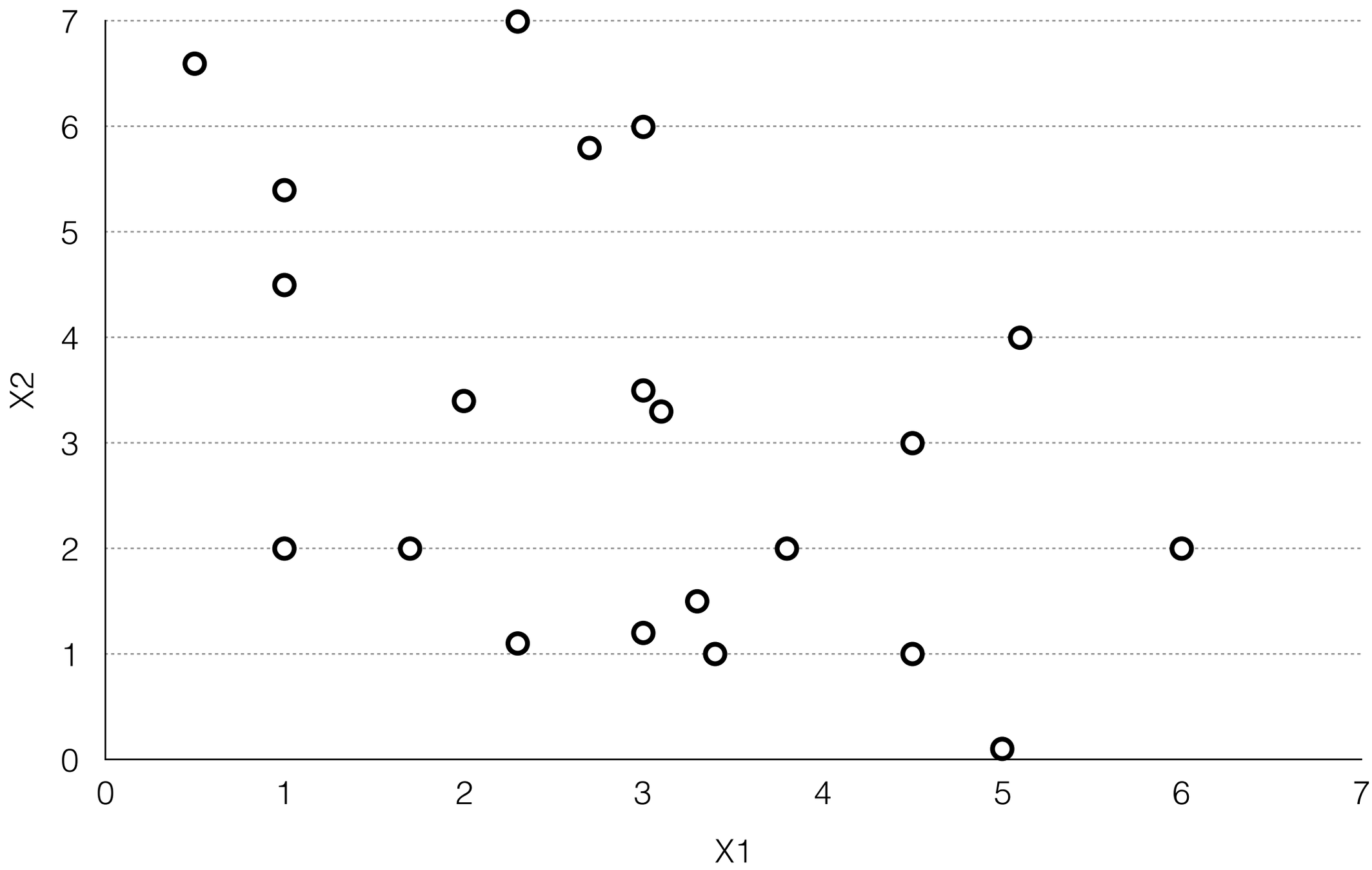
predict **1** when:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 \mathbf{>= 0}$$

predict **0** when:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 \mathbf{< 0}$$

Pay Gap Data



Decision Boundary

$$\text{Pay Gap} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Decision Boundary

$$\text{Pay Gap} = -6 + 1x_1 + 1x_2$$

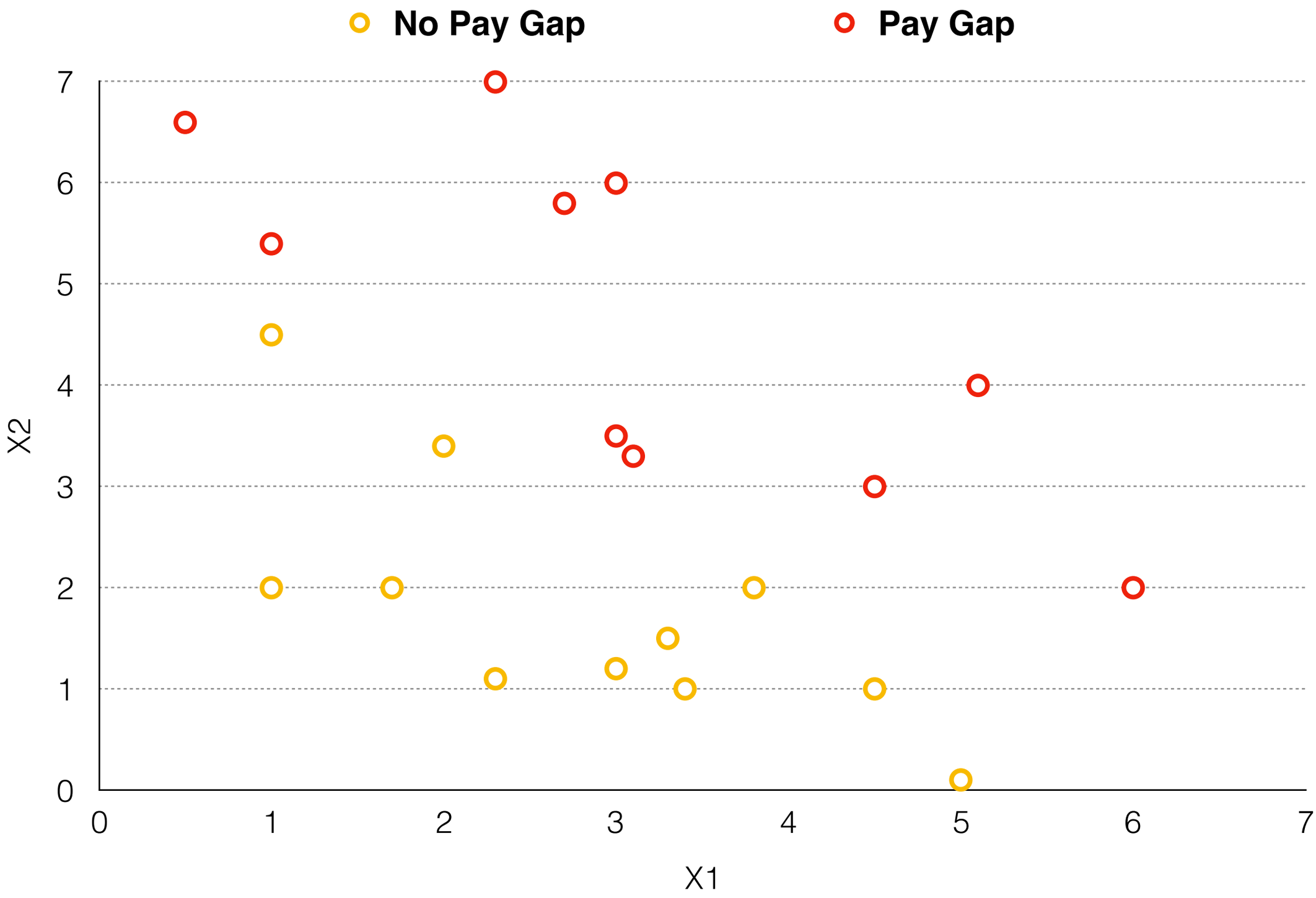
Decision Boundary

$$\text{Pay Gap} = -6 + 1x_1 + 1x_2$$

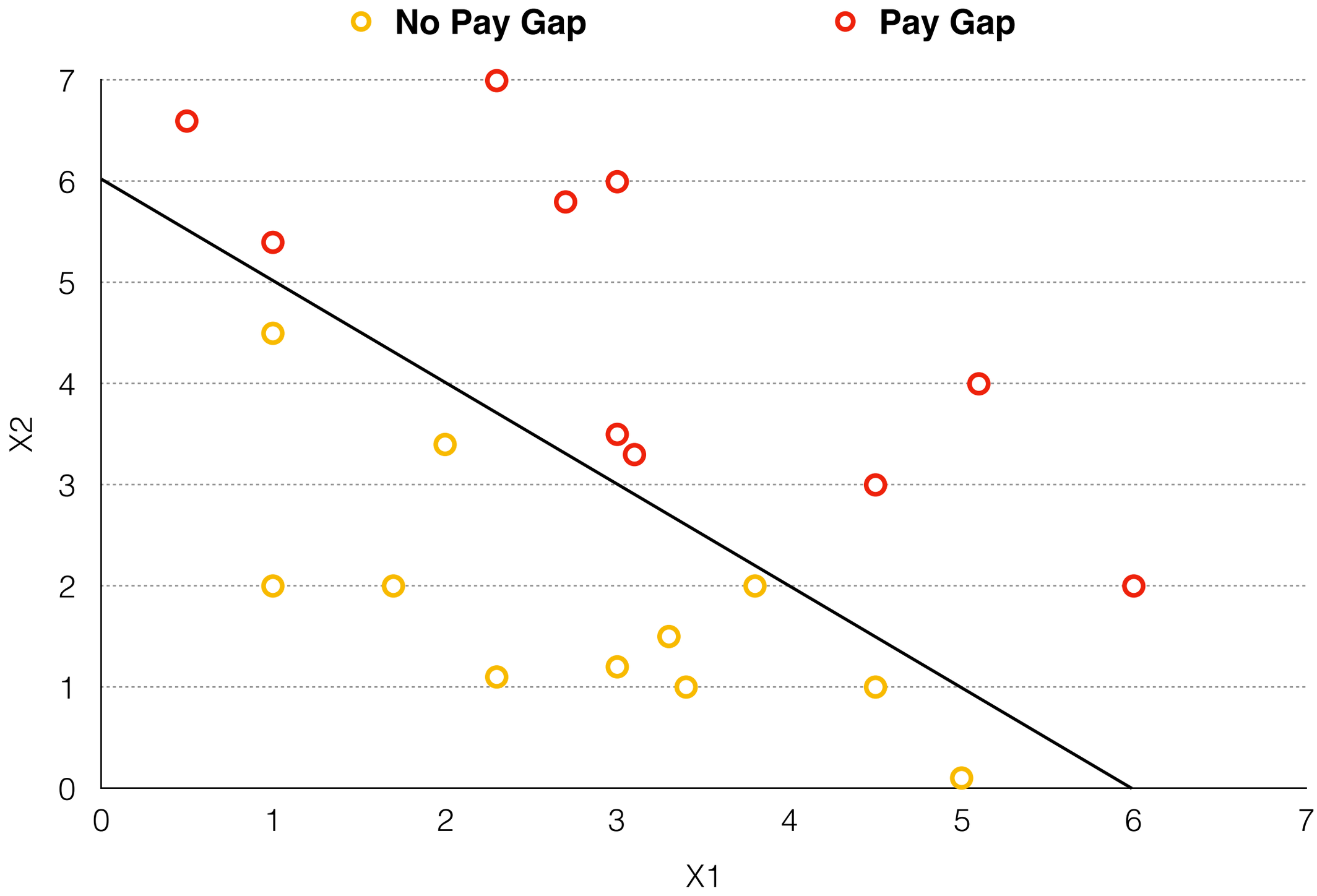
predict 1 when $-6 + 1x_1 + 1x_2 >= 0$

predict 0 when $-6 + 1x_1 + 1x_2 < 0$

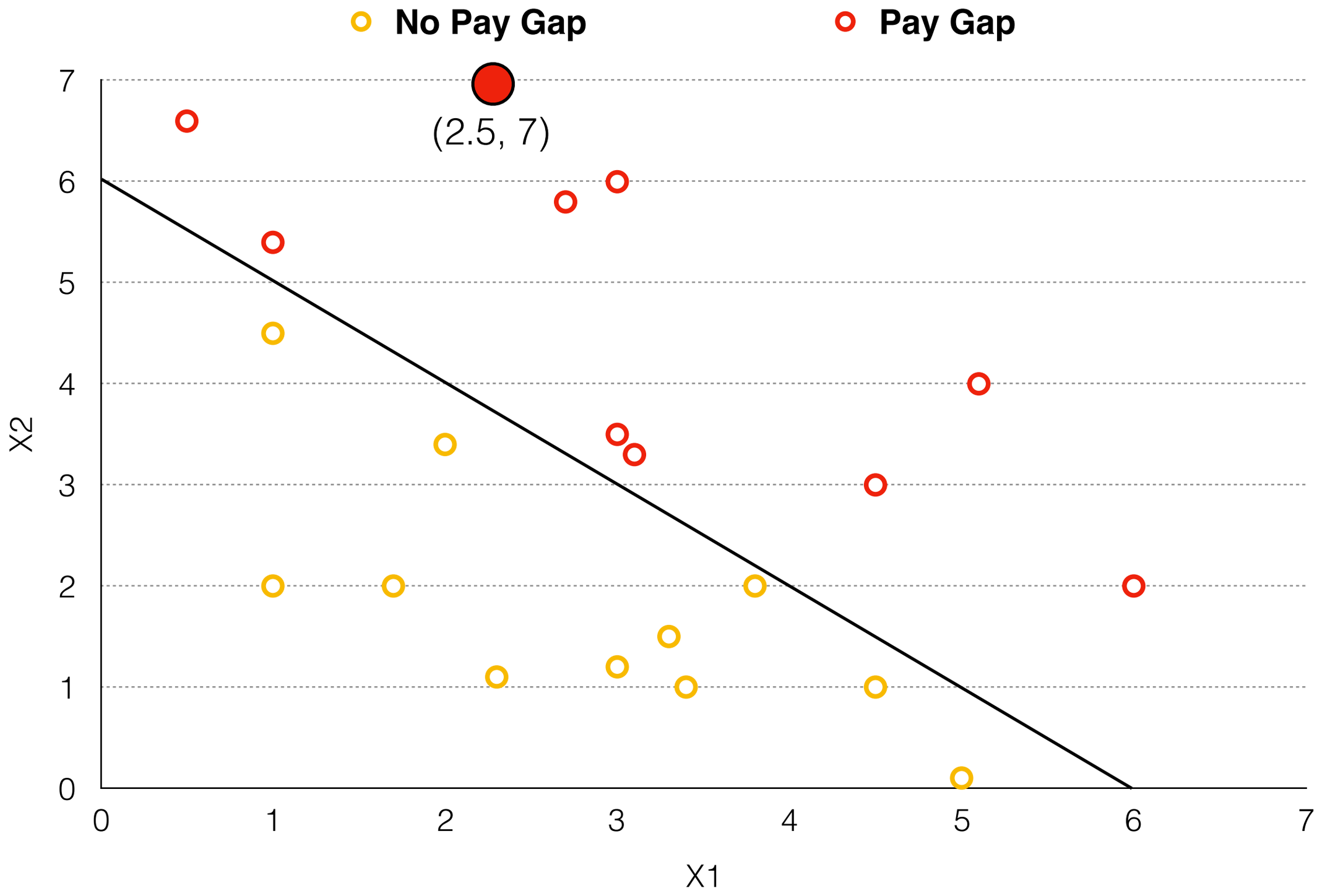
Pay Gap Decision Boundary



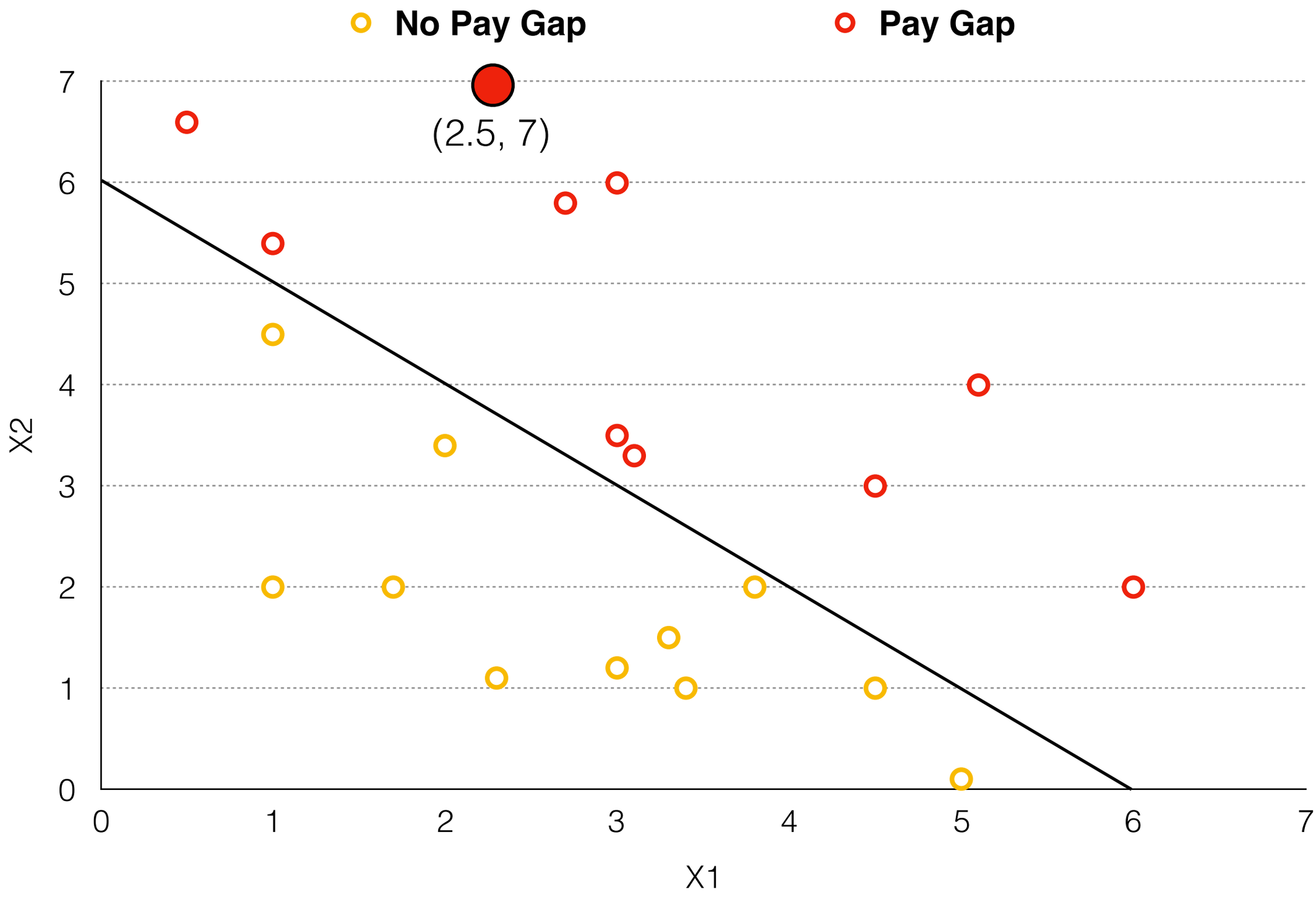
$$\hat{f}(X) = \frac{1}{1+e^{-(-6 + 1x_1 + 1x_2)}}$$



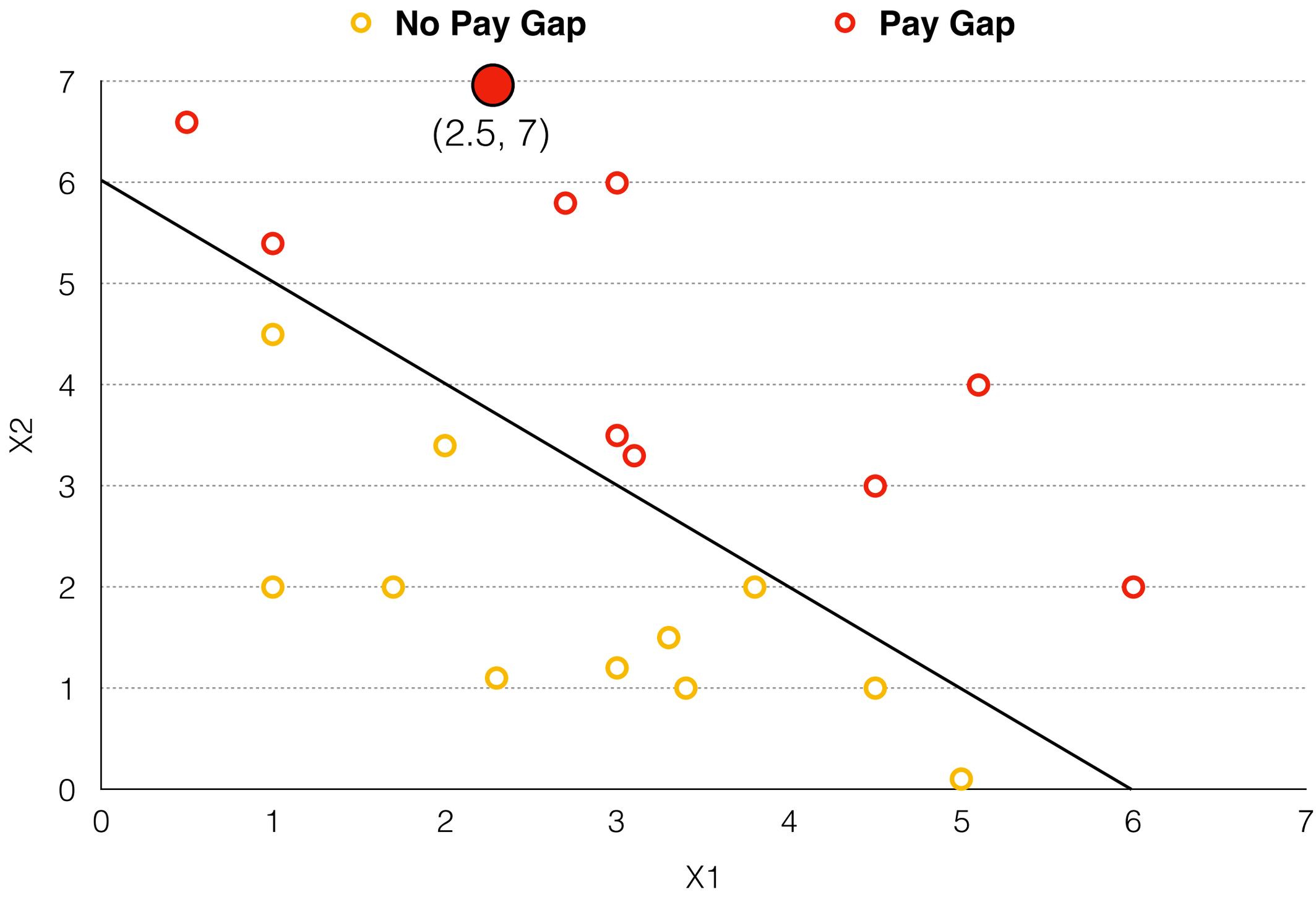
$$\hat{f}(X) = \frac{1}{1+e^{-(-6 + 1x_1 + 1x_2)}}$$



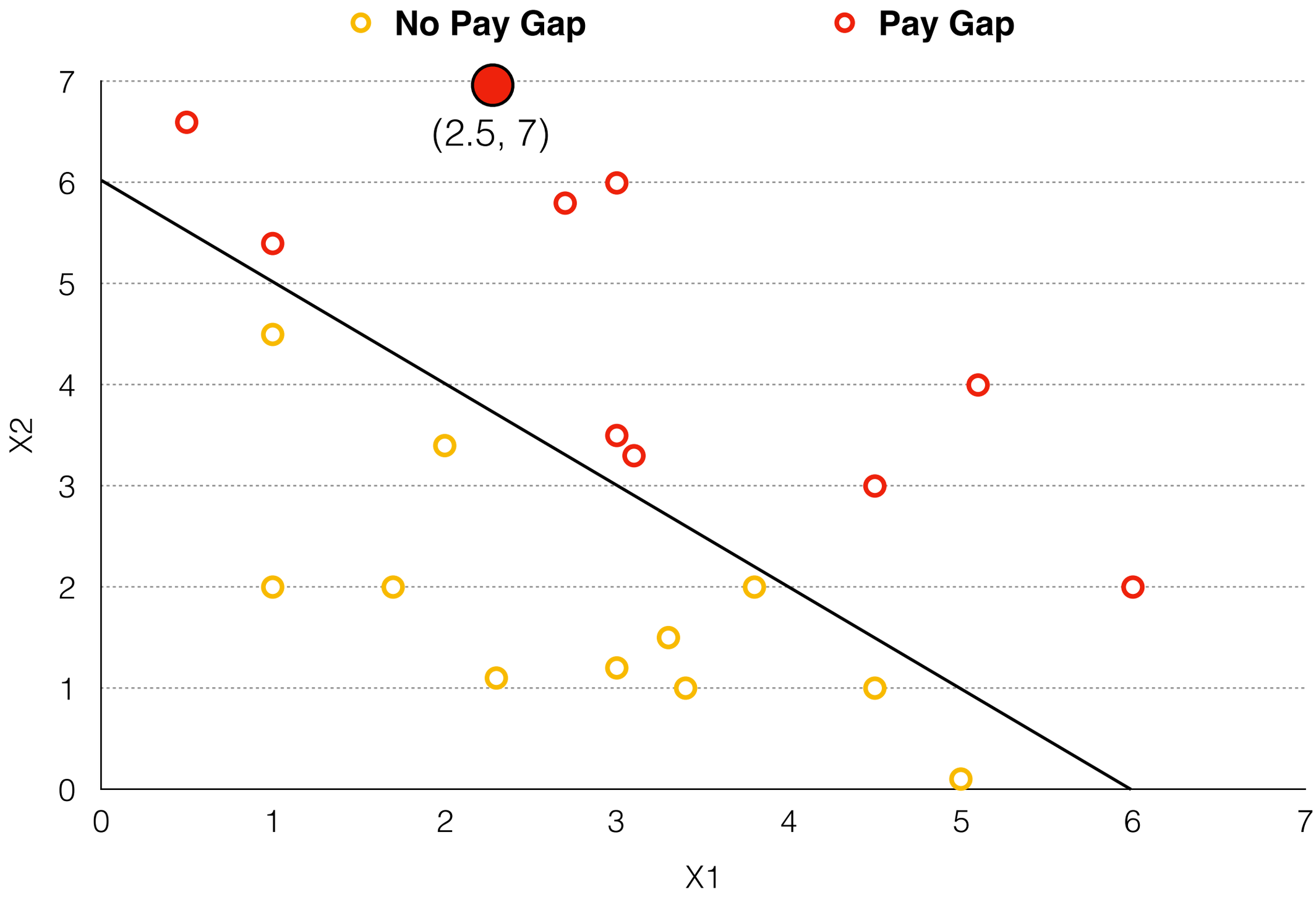
$$\hat{f}(X) = \frac{1}{1+e^{-(-6 + 2.5 + 7)}}$$



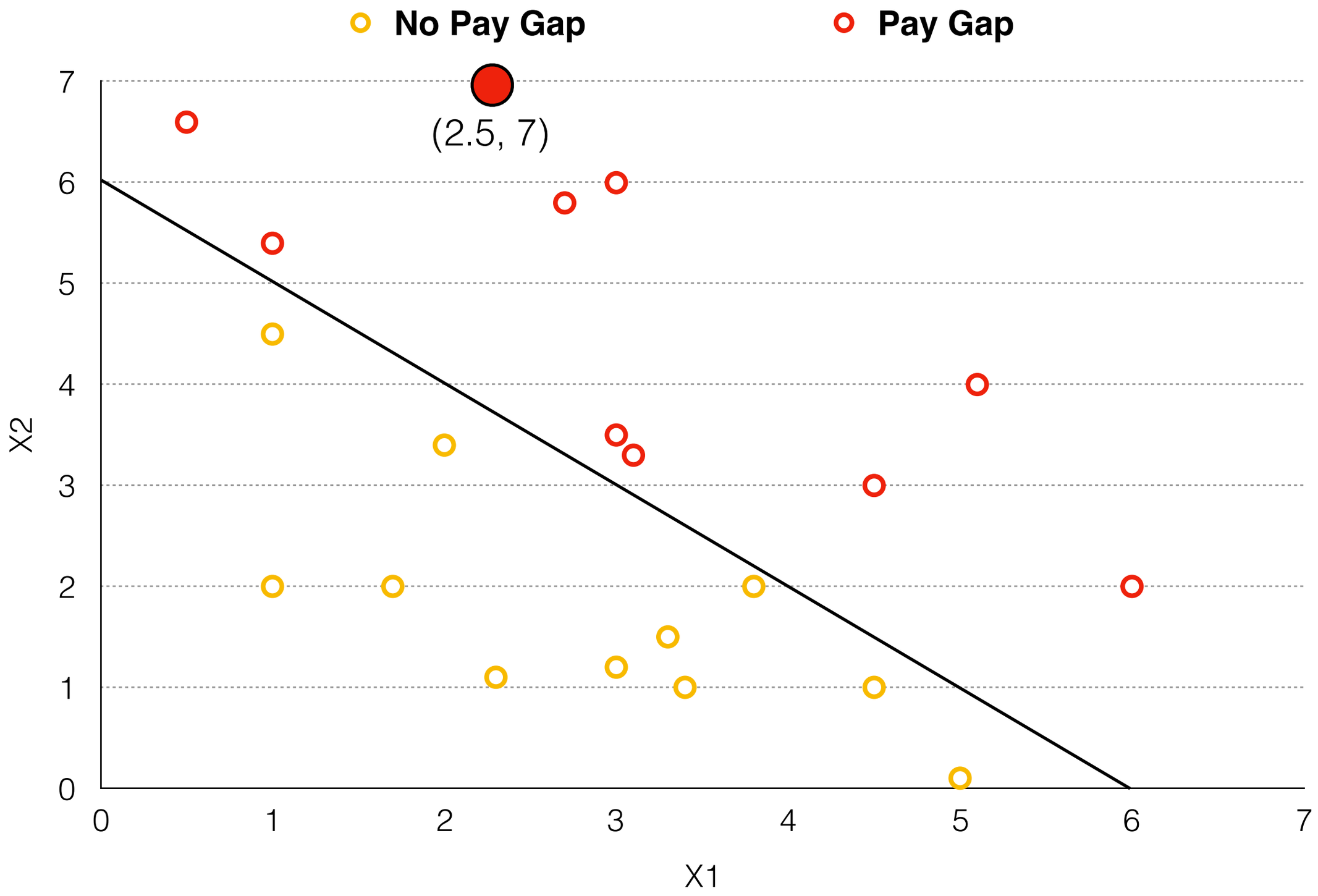
$$\hat{f}(X) = \frac{1}{1+e^{-(-6+9.5)}}$$



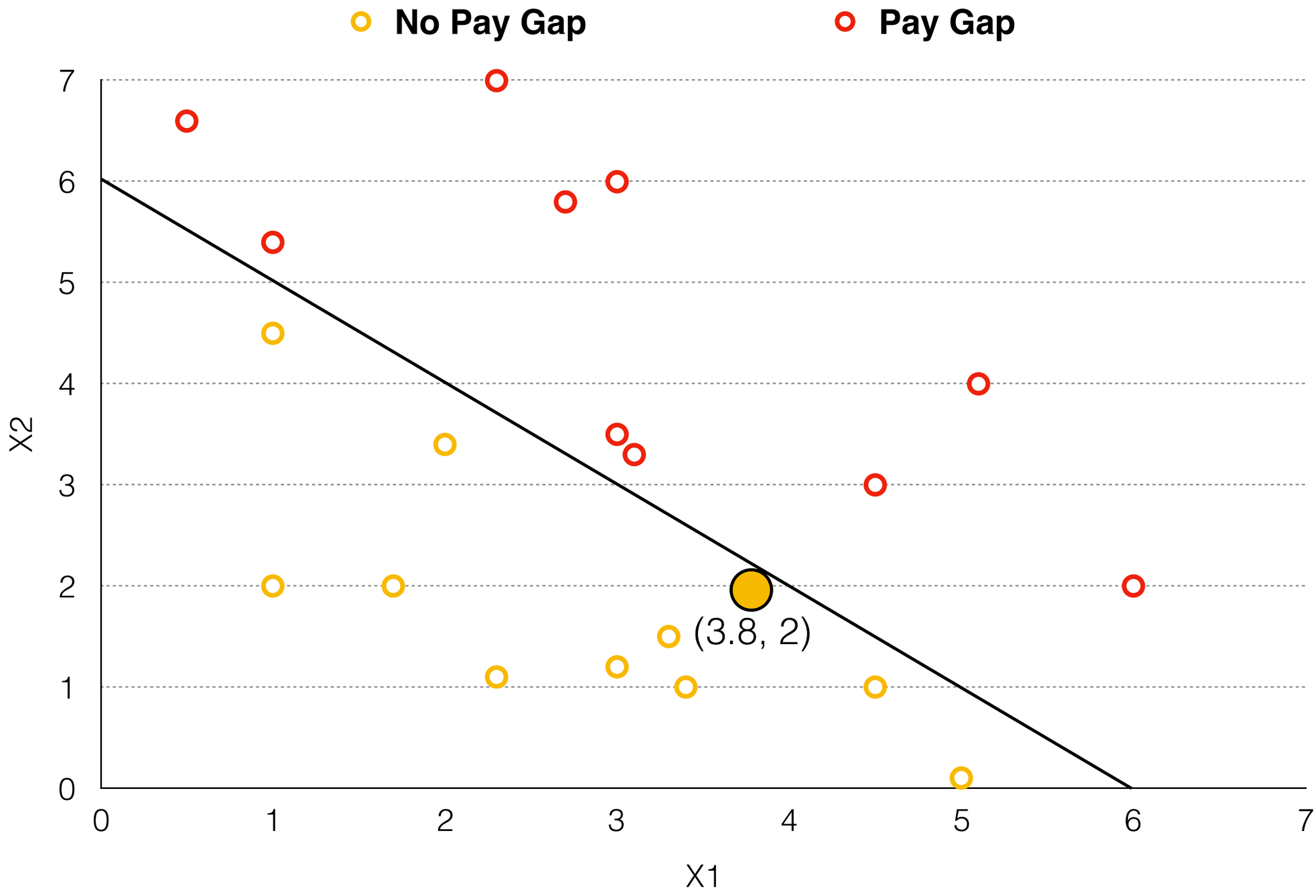
$$\hat{f}(X) = \frac{1}{1+e^{-(3.5)}}$$



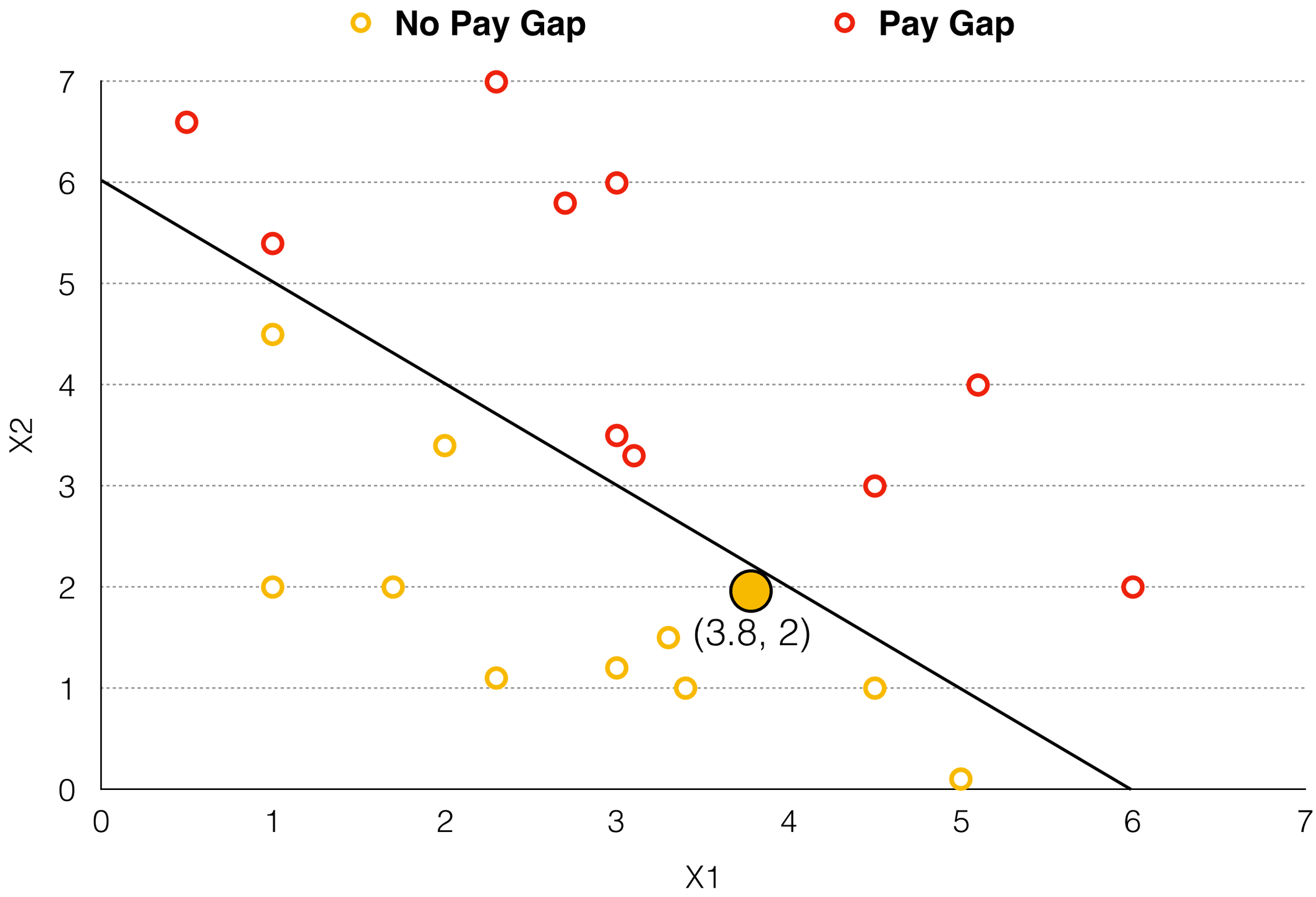
$$\hat{f}(X) = \frac{1}{1+e^{-(3.5)}} = \textbf{.97} \text{ (probability)}$$



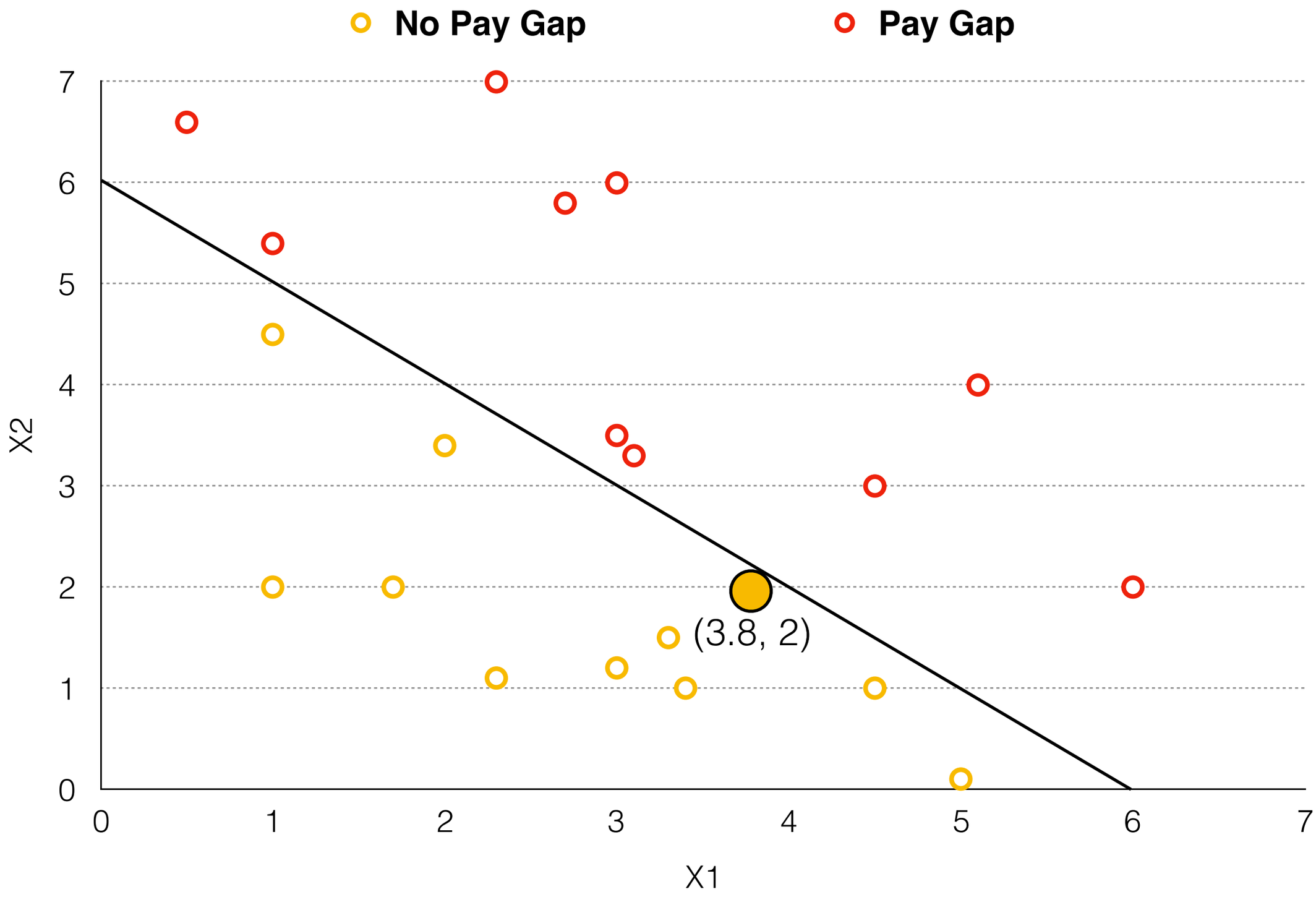
$$\hat{f}(X) = \frac{1}{1+e^{-(-6 + 1x_1 + 1x_2)}}$$



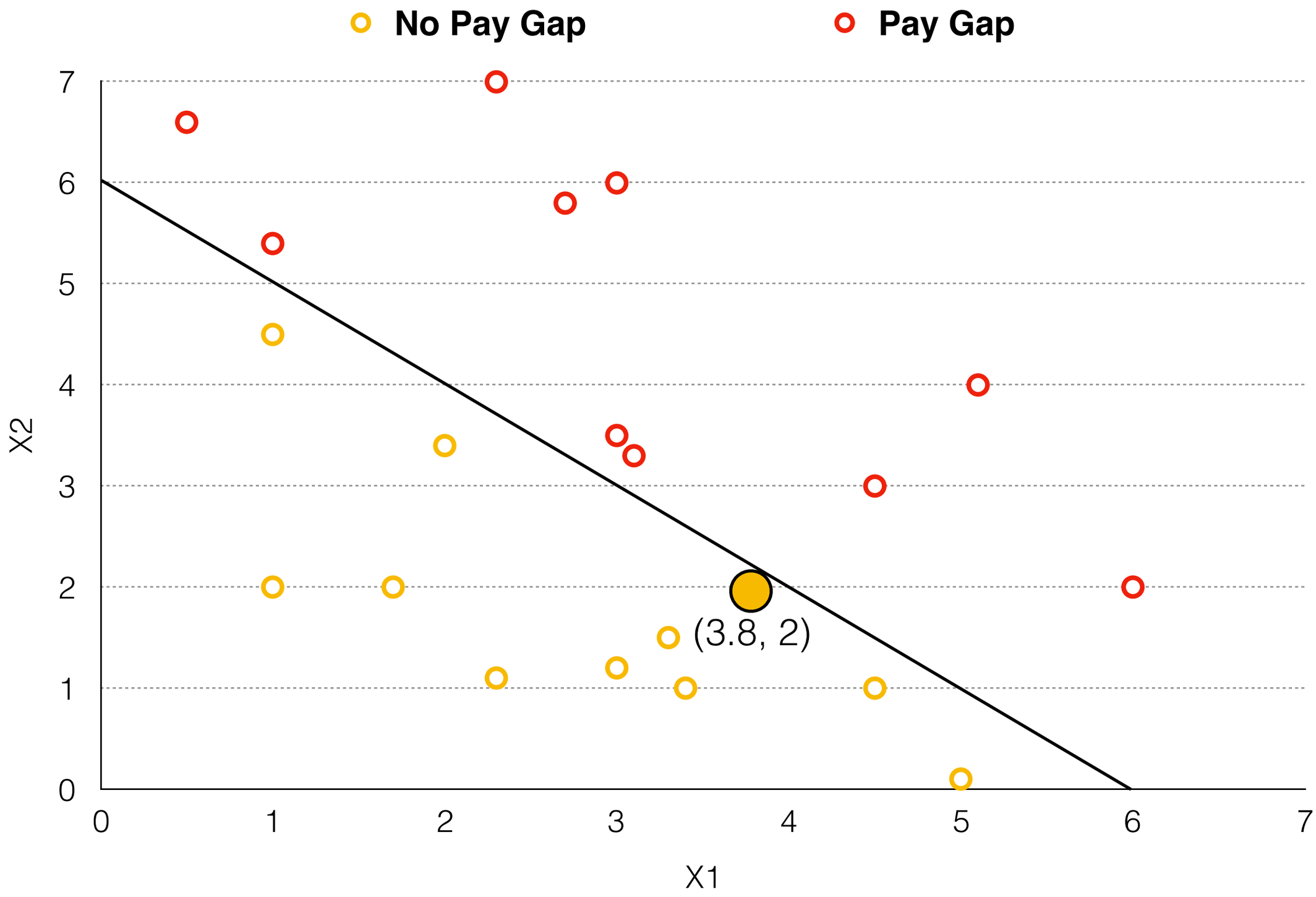
$$\hat{f}(X) = \frac{1}{1+e^{-(-6+3.8+2)}}$$



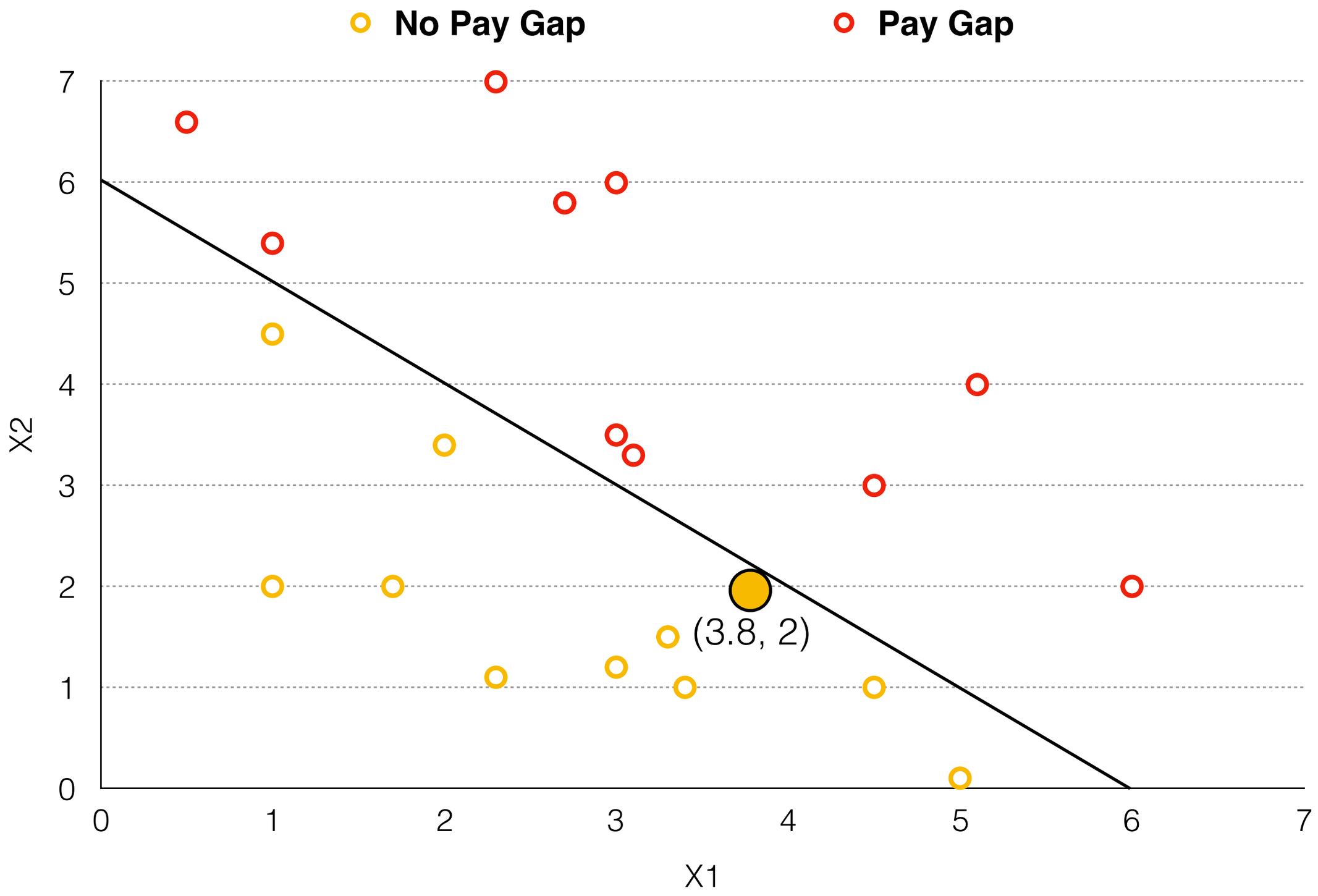
$$\hat{f}(X) = \frac{1}{1+e^{-(-6+5.8)}}$$



$$\hat{f}(X) = \frac{1}{1+e^{-(-0.2)}}$$



$$\hat{f}(X) = \frac{1}{1+e^{-(-0.2)}} = \textbf{.45} \text{ (probability)}$$



Support Vector Machine

Support Vector Machine

$$\beta_0 + \beta_1 x + \beta_2 x + \beta_3 x$$

Support Vector Machine

predict 1 when $\beta_0 + \beta_1 x_1 + \beta_2 x_2 >= 0$

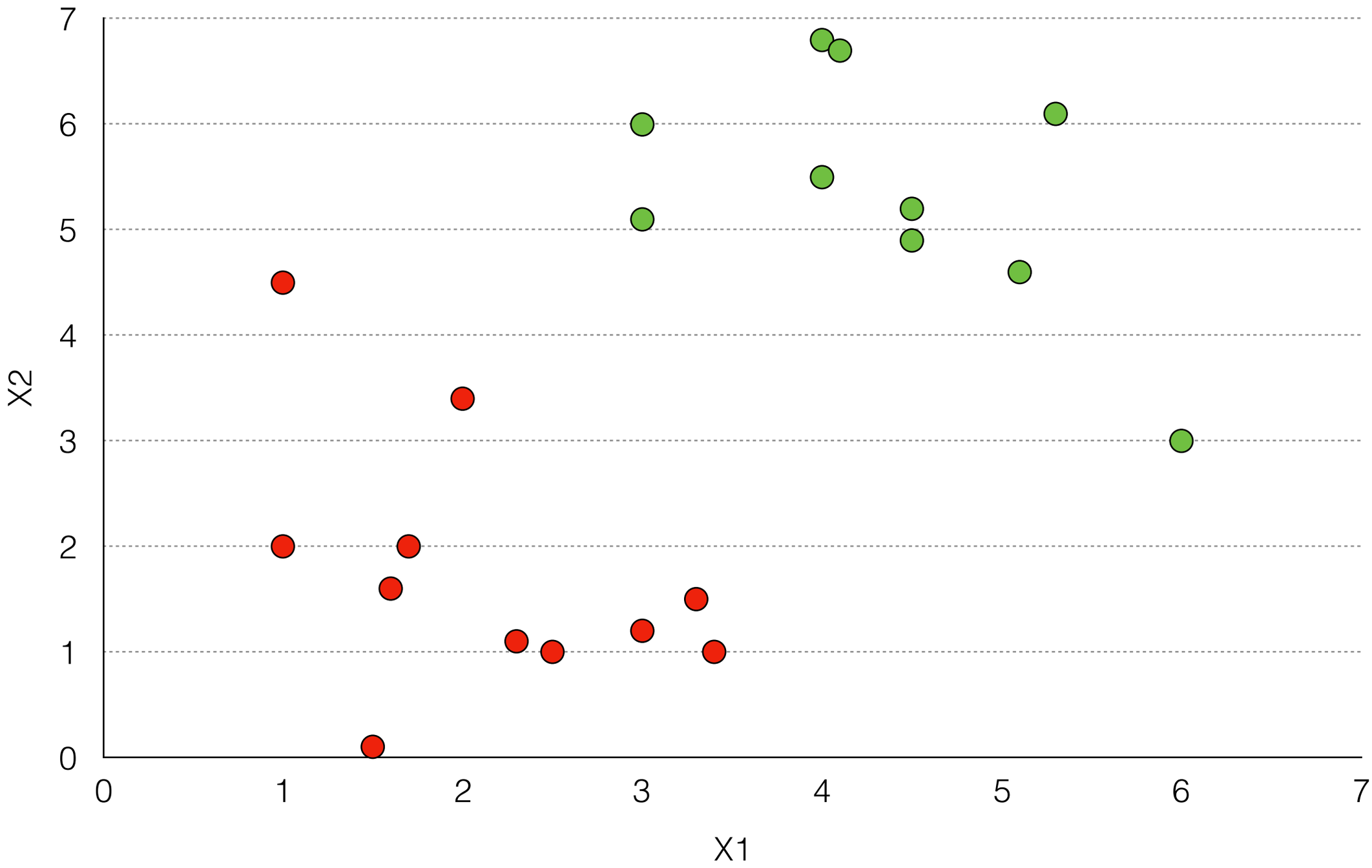
predict 0 when $\beta_0 + \beta_1 x_1 + \beta_2 x_2 < 0$

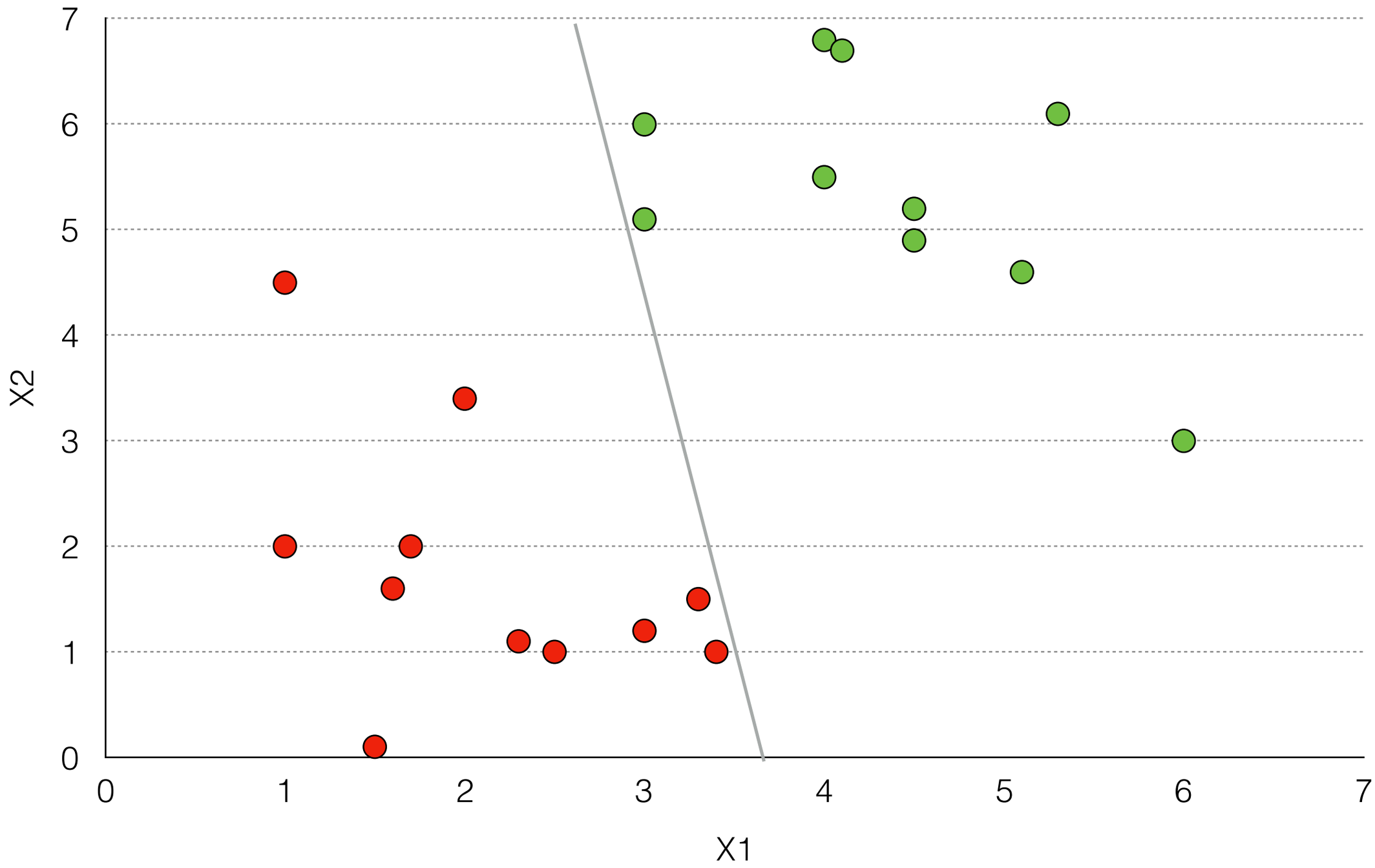
Support Vector Machine

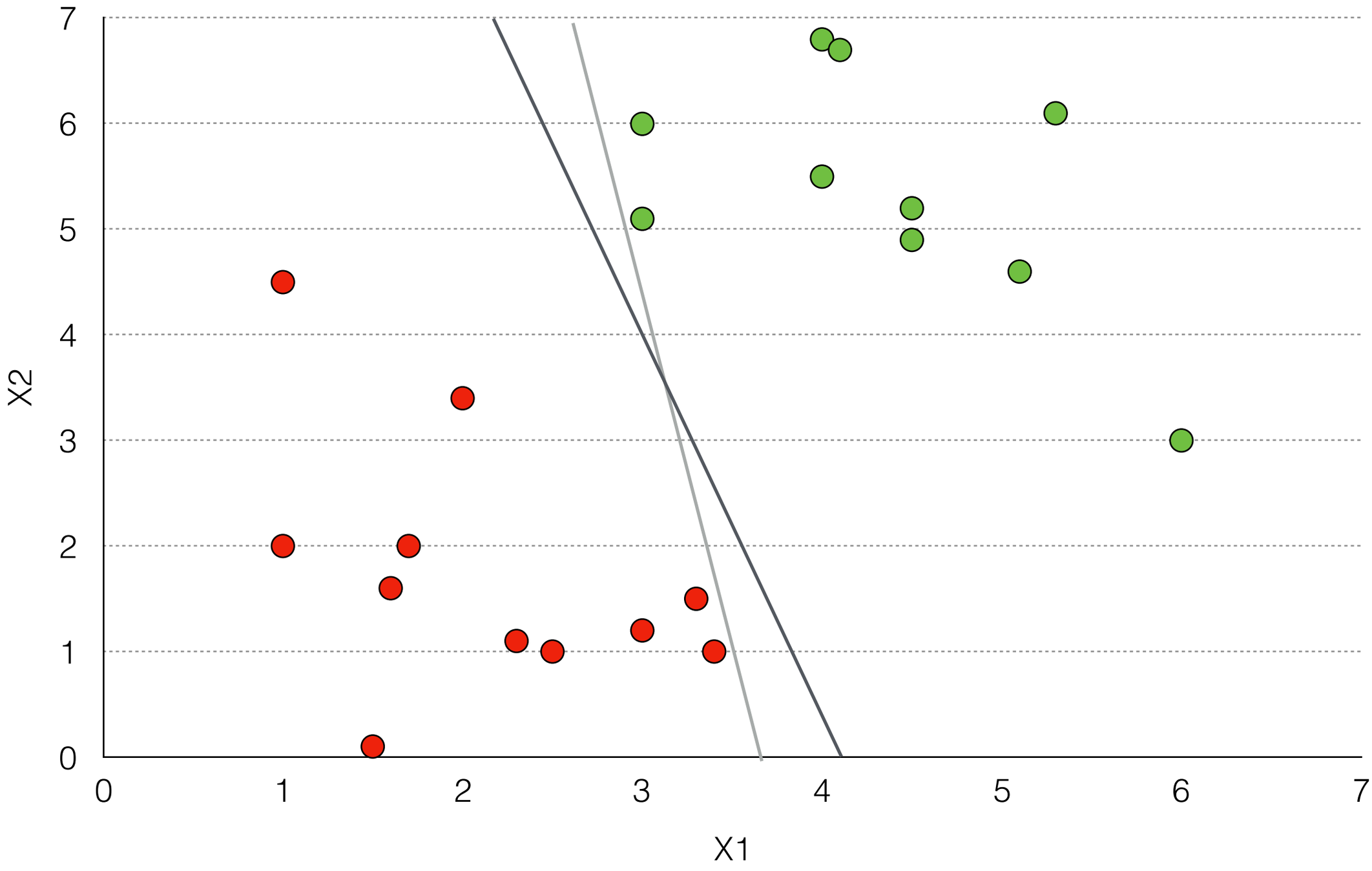
Large Margin Classifier

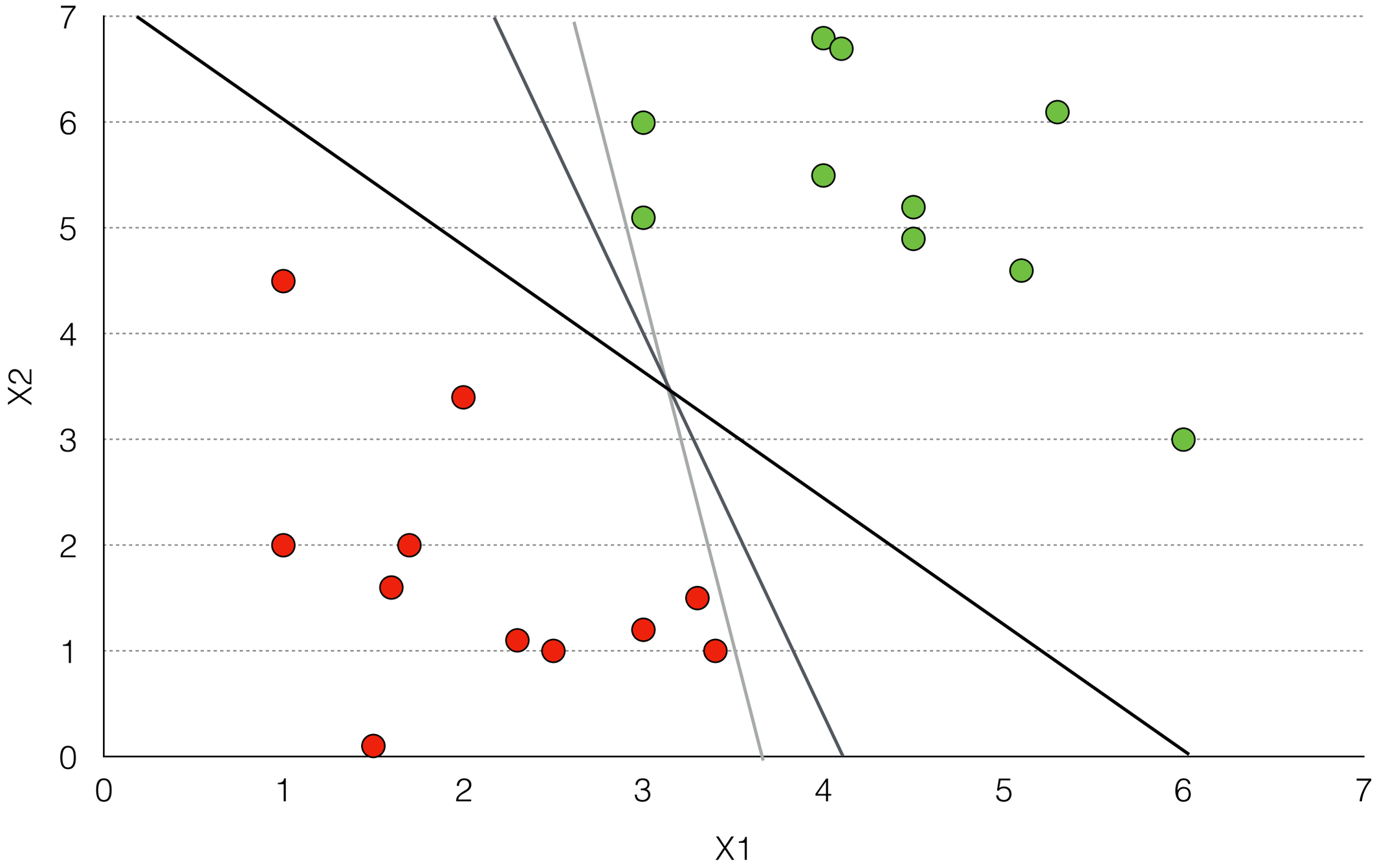
predict 1 when $\beta_0 + \beta_1 x_1 + \beta_2 x_2 >= 1$

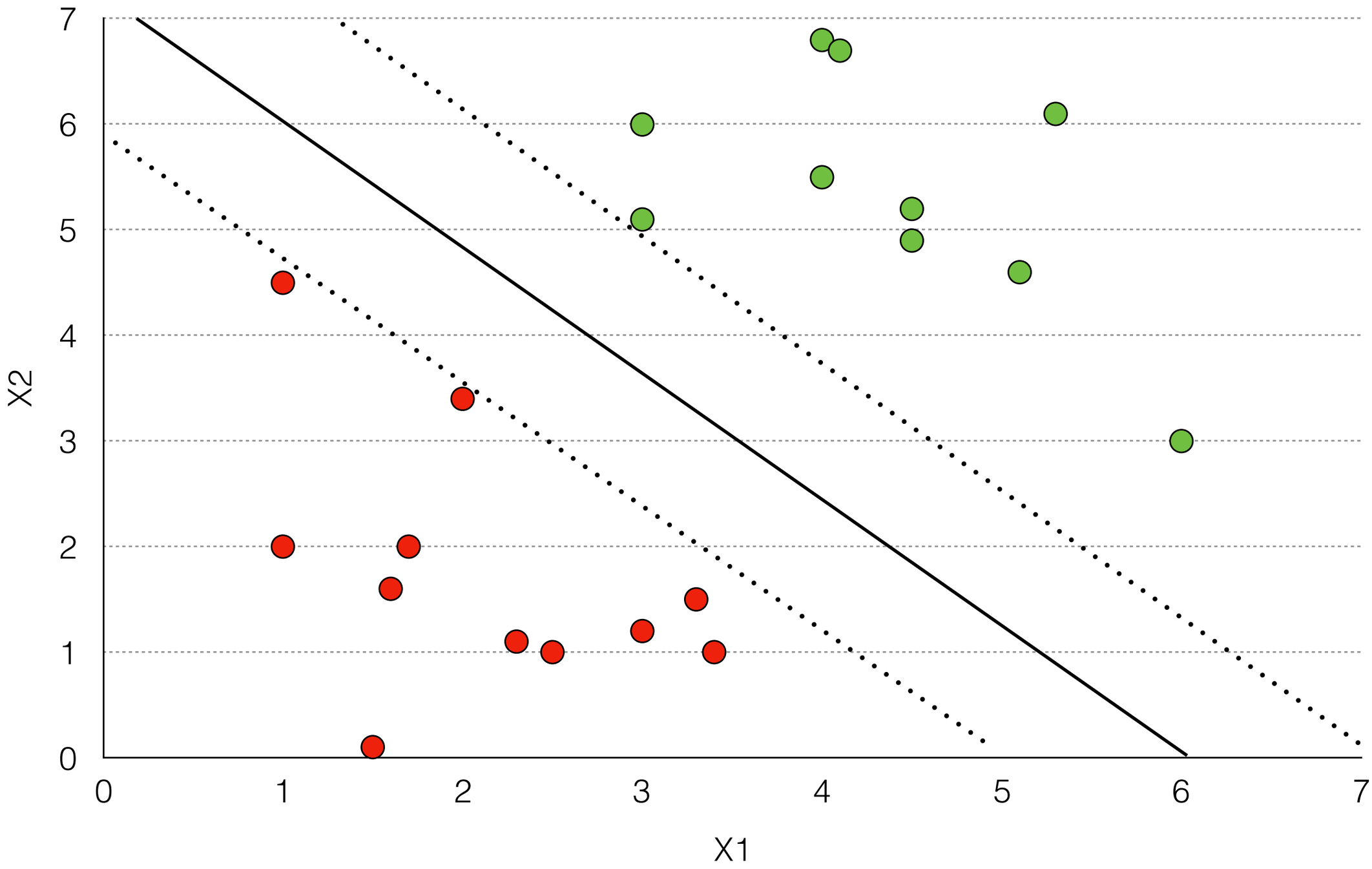
predict 0 when $\beta_0 + \beta_1 x_1 + \beta_2 x_2 < -1$


7
6
5
4
3
2
1
0
X2
0
1
2
3
4
5
6
7
X1


7
6
5
4
3
2
1
0
X2
0
1
2
3
4
5
6
7
X1


7
6
5
4
3
2
1
0
X2
0
1
2
3
4
5
6
7
X1


7
6
5
4
3
2
1
0
X2
0
1
2
3
4
5
6
7
X1


7
6
5
4
3
2
1
0
X2
0
1
2
3
4
5
6
7
X1

PLAYLIST

Discover Weekly

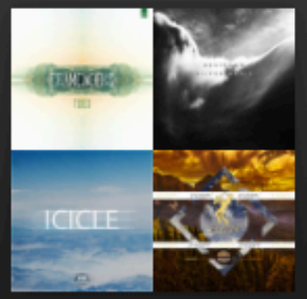
Your weekly mixtape of fresh music. Enjoy new discoveries and deep cuts chosen just for you. Updated every Monday, so save your favourites!

Created by: **Spotify** • 30 songs, 2 hr 48 min

PLAY FOLLOWING ••• FOLLOWER 1

Filter Download

	SONG	ARTIST	ALBUM		
+	The Sky out of Your Window	Melorman	Waves	4 days ago	3:21
+	You Have Love	Axel Thesleff	You Have Love	4 days ago	7:21
+	You're Still In It	Chihei Hatakeyama	You're Still In It	4 days ago	18:26
+	Morning Mountain	Essay	Morning Mountain	4 days ago	6:42



PLAYLIST

LIKED

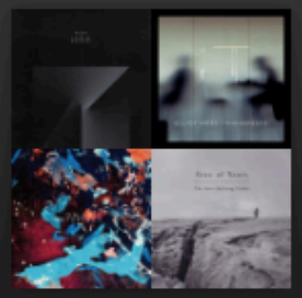
Created by: **hisbiz** • 334 songs, 24 hr 17 min

PLAY

Filter

Download

	TITLE	ARTIST
+	Breathing Light	Frameworks
+	Superior	Silver Maple
+	Icicle	AK
+	Jazzin	Flap Jack
+	Dusk	filous
+	The Way U Do	Shlohmo
+	Mirror Maru	Cashmere ...
+	Never Too Far	Sorrow
+	Mixed Signals - Synkro Remix	Frederic R...
+	Swarm	Boogrov, ...



PLAYLIST

REJECTED

Created by: **hisbiz** • 331 songs, 28 hr 44 min

PLAY

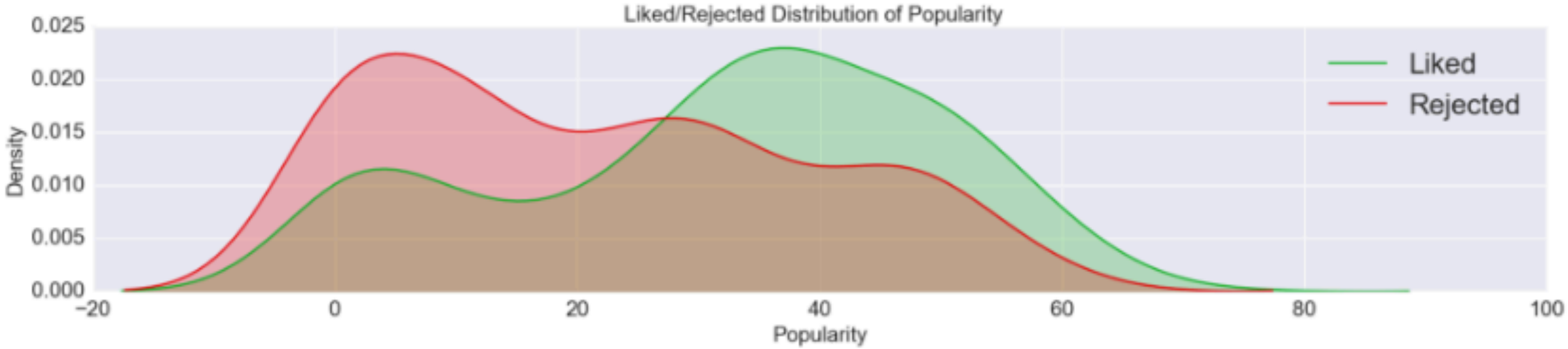
Filter

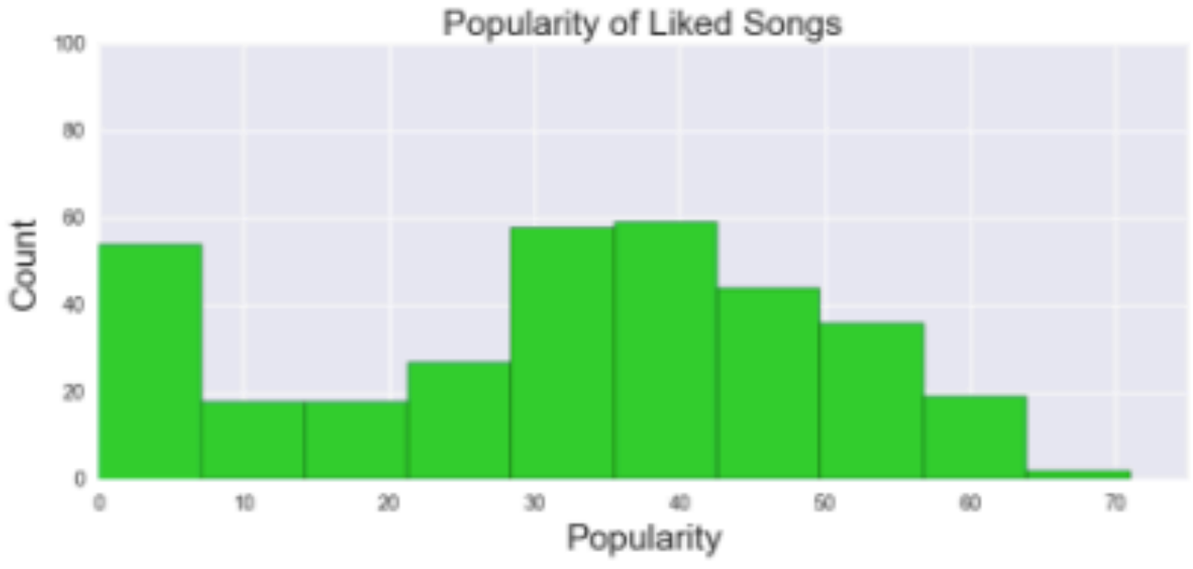
Download

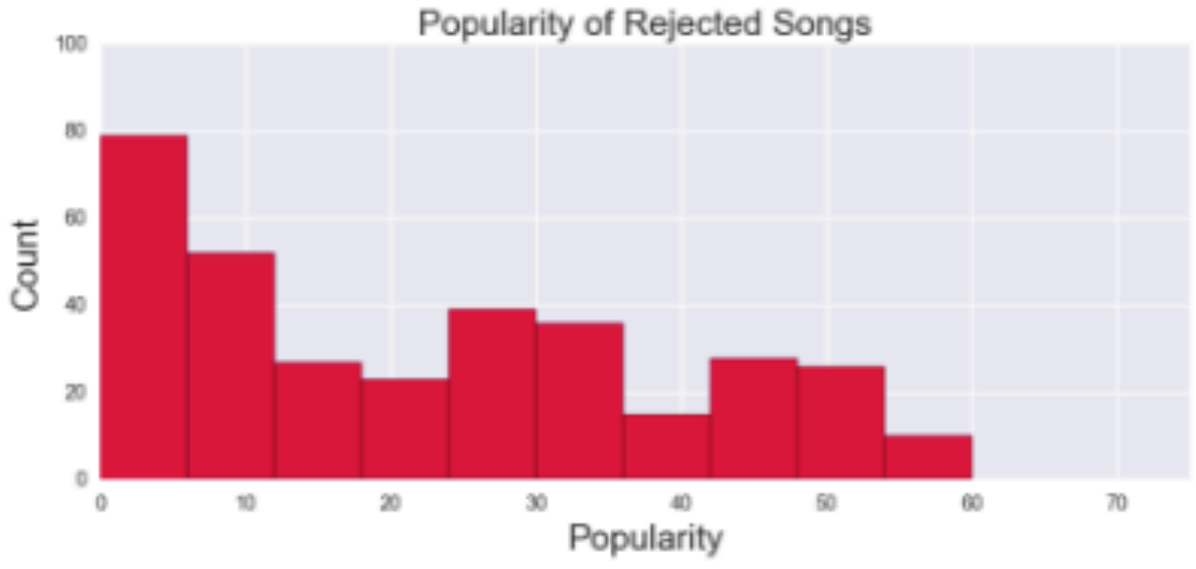
	TITLE	ARTIST
+	Frogs	Charles M...
+	Best Light	Elliot Moss
+	The Silence	Om Unit fe...
+	Passing Skies	Seas of Ye...
+	Urban Transition	Jimmy Wa...
+	Springflower	NkisOk
+	Where Did The Children Go	Deformer
+	Tactical Nuclear Penguin	Spenghead
+	Prometheus	Pythius
+	Inadequante	HE3Dless

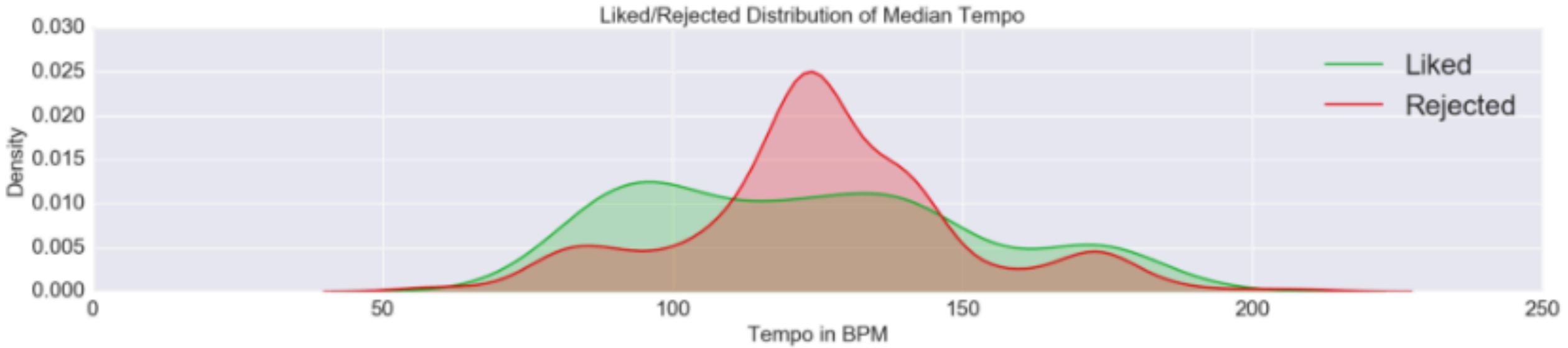
Select Features

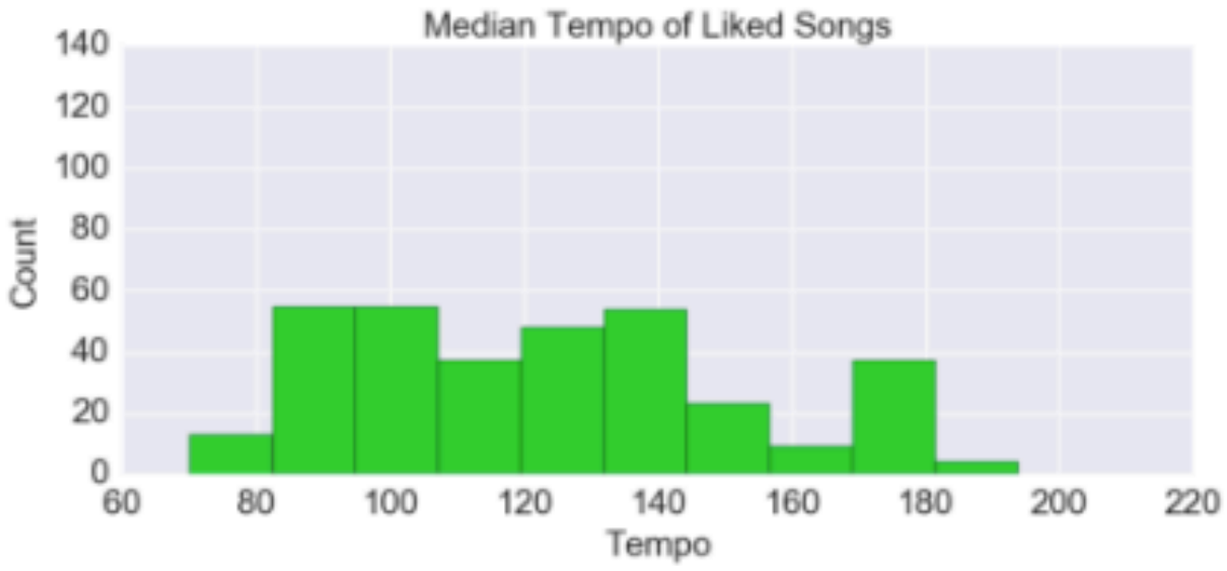
Song	Duration	Pitch	Timbre	Tempo	Popularity	Genre

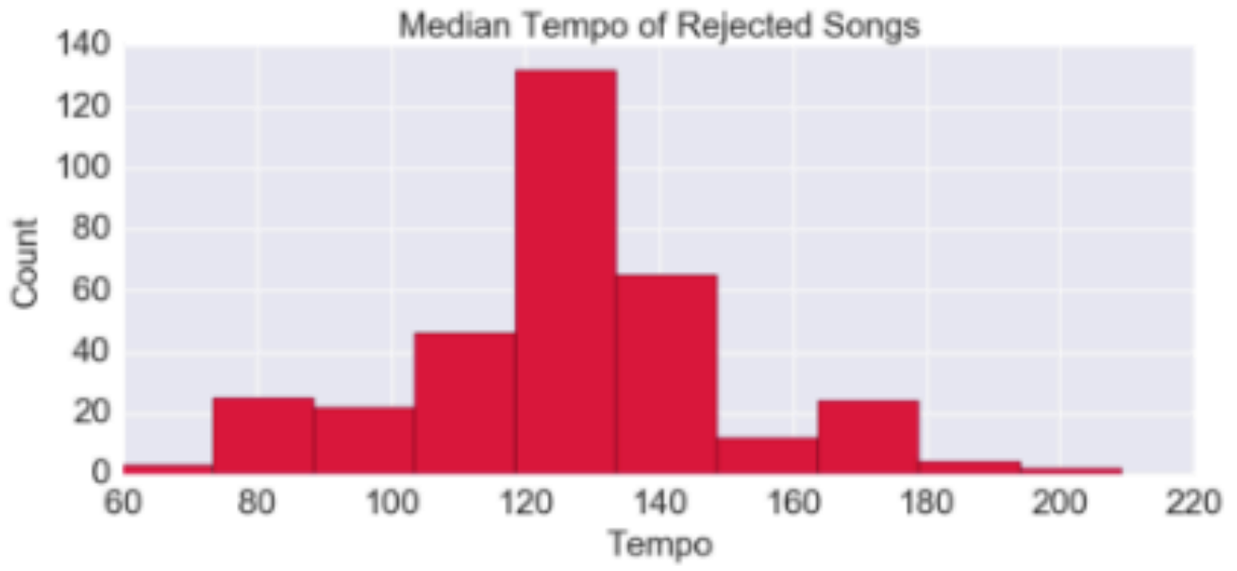

Liked/Rejected Distribution of Popularity
Liked
Rejected
Density
0.025
0.020
0.015
0.010
0.005
0.000
−20
0
20
40
60
80
100
Popularity

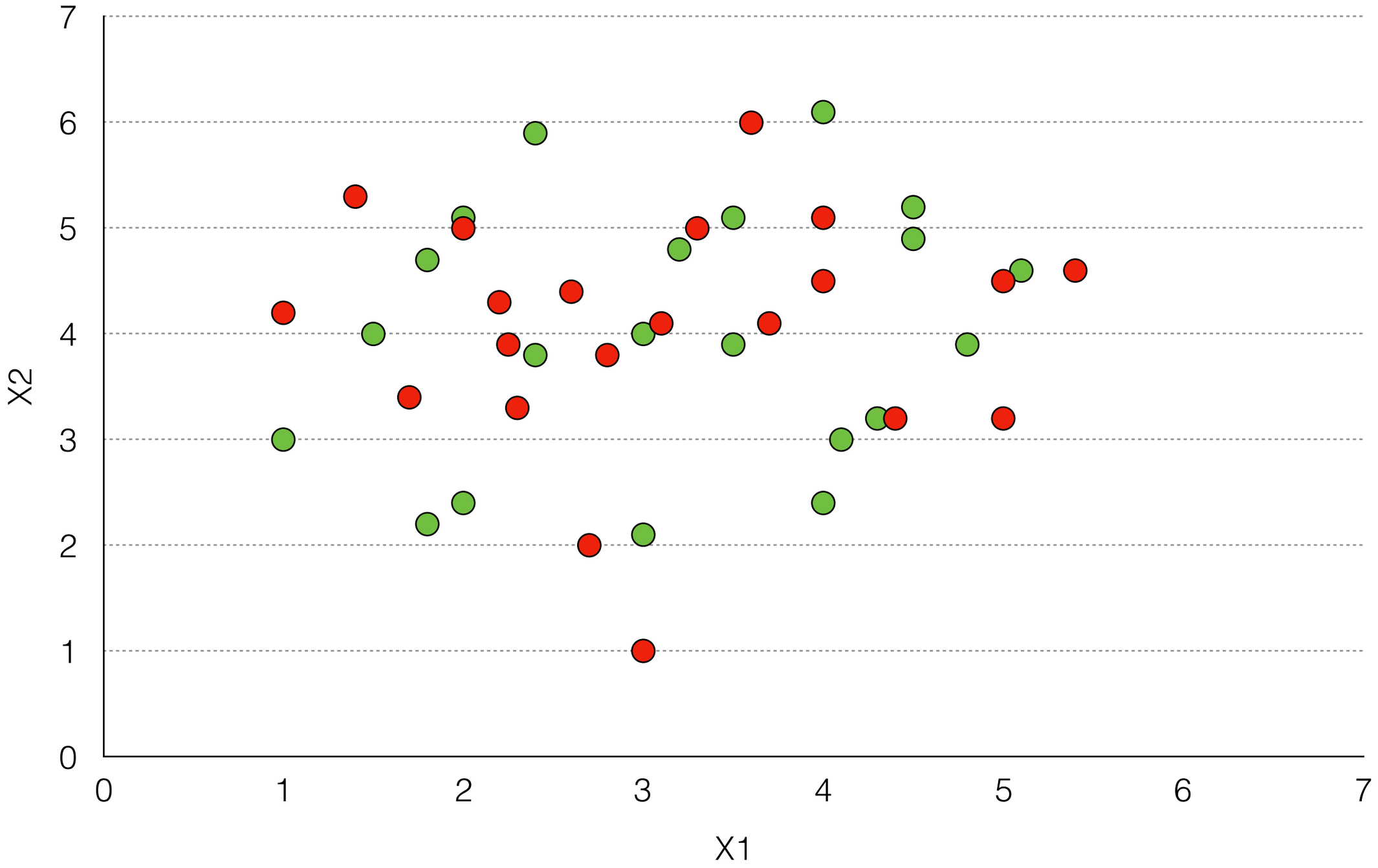

Popularity of Liked Songs
Count
0
20
40
60
80
100
0
10
20
30
40
50
60
70
Popularity


Popularity of Rejected Songs
Count
0
20
40
60
80
100
0
10
20
30
40
50
60
70
Popularity


Liked/Rejected Distribution of Median Tempo
Density
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
Tempo in BPM
Liked
Rejected


Median Tempo of Liked Songs
Count
140
120
100
80
60
40
20
0
60
80
100
120
140
160
180
200
220
Tempo


Median Tempo of Rejected Songs
Count
140
120
100
80
60
40
20
0
60
80
100
120
140
160
180
200
220
Tempo

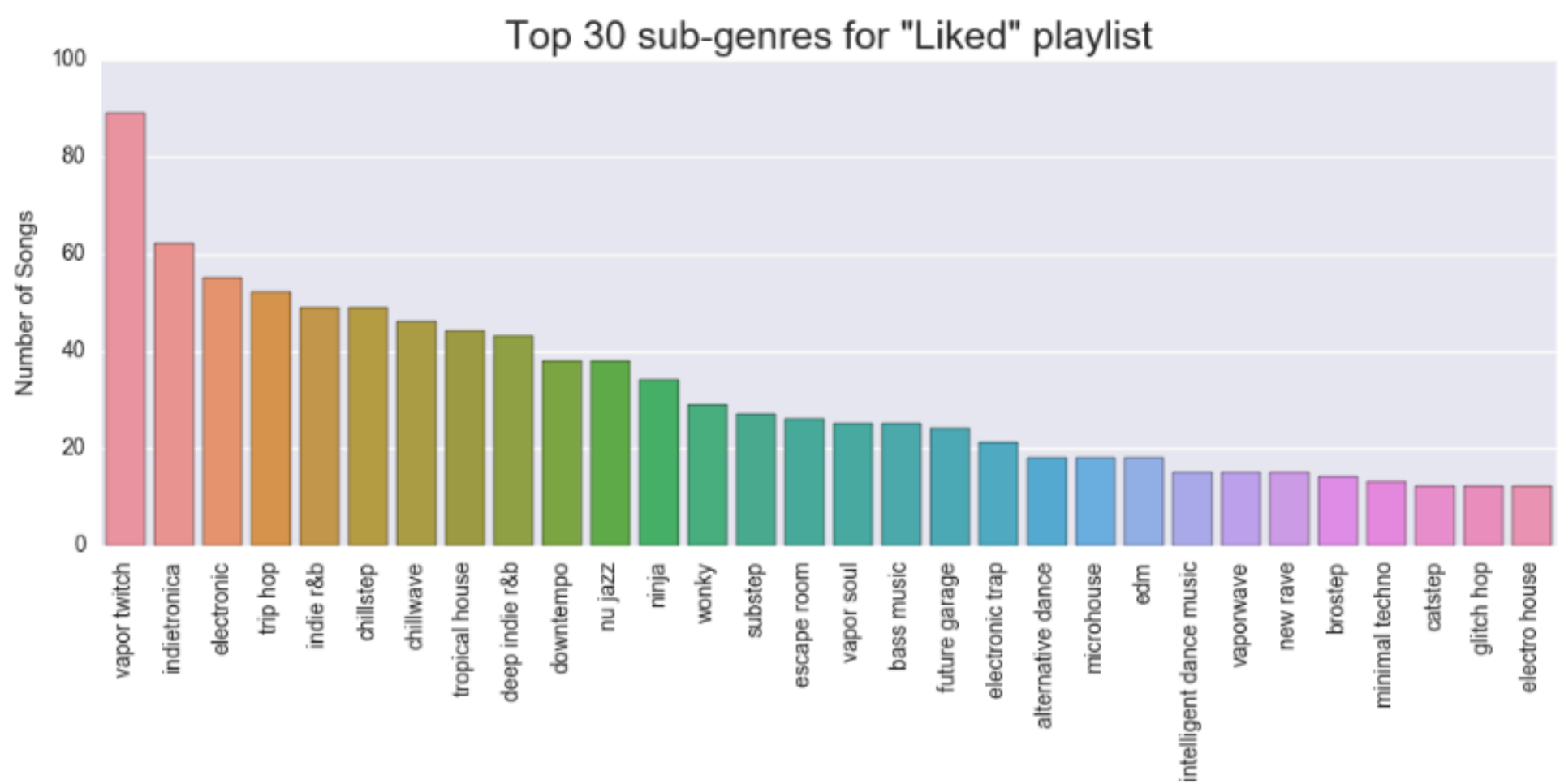

7
6
5
4
3
2
1
0
X2
0
1
2
3
4
5
6
7
X1

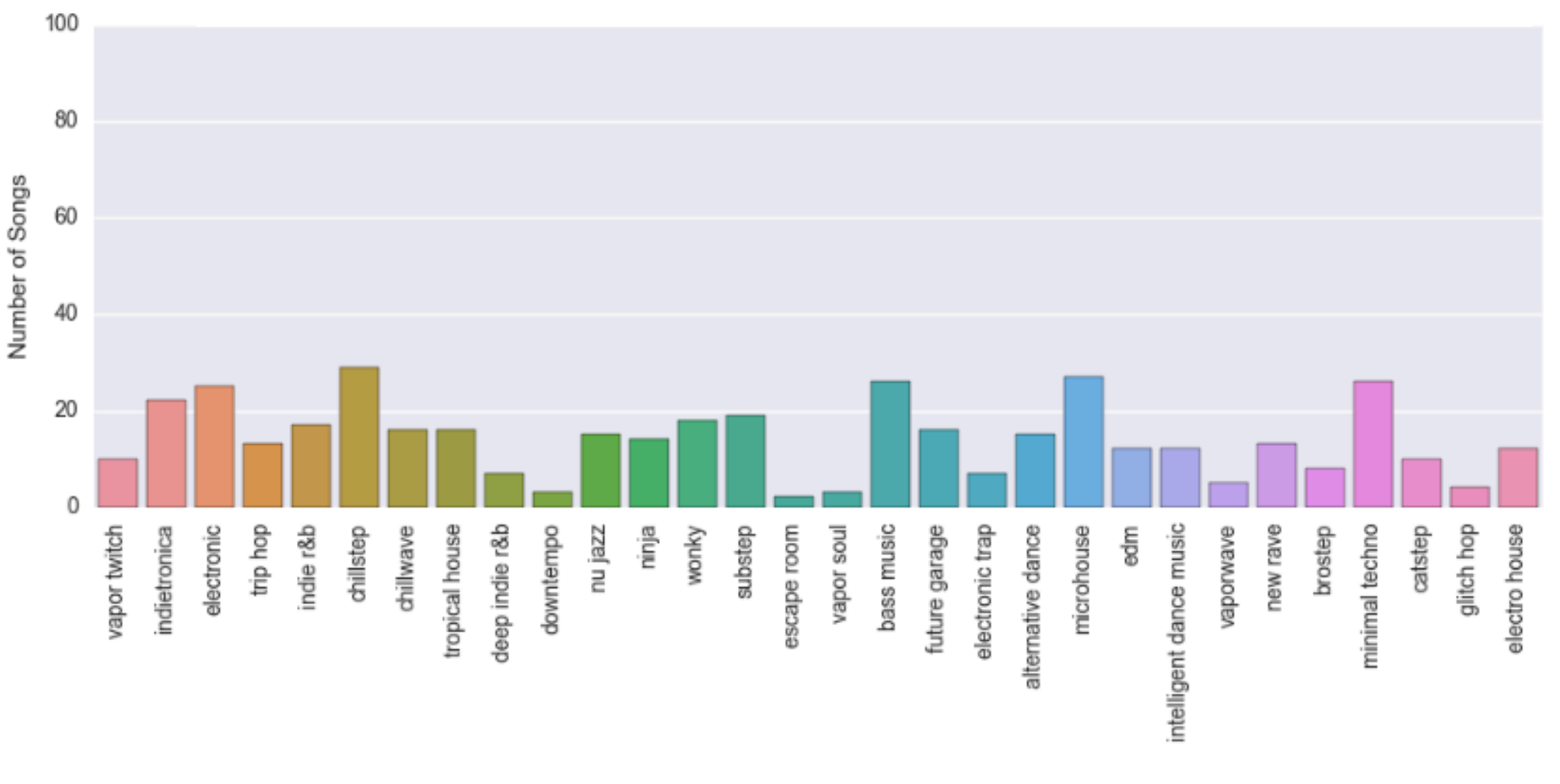
Feature Engineering

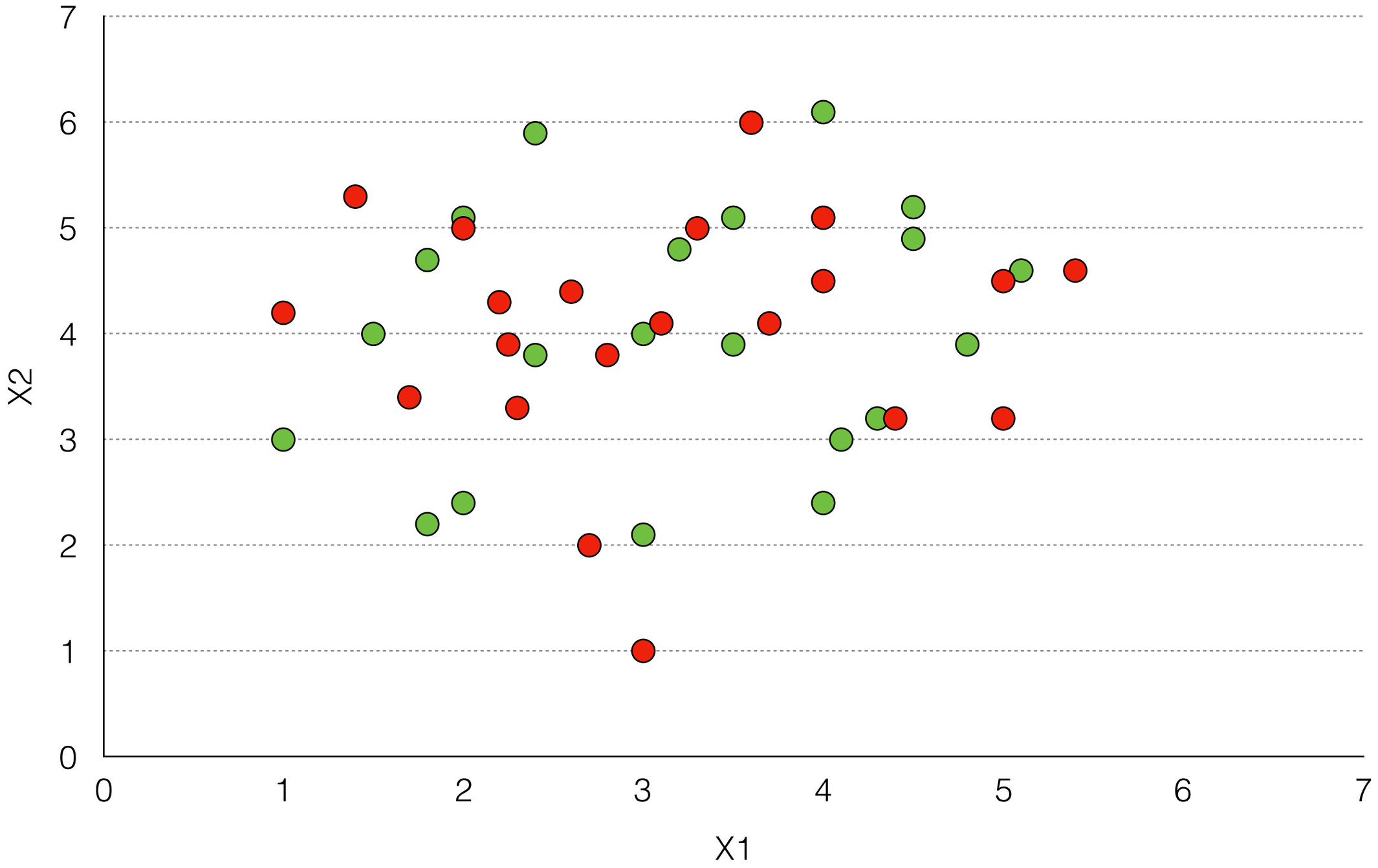
	Song	Artist Genres
10	Takeover	['neurostep', 'vapor twitch']
11	Kissed By A Kisser	['acid jazz', 'ninja', 'nu jazz', 'trip hop']
12	Parks On Fire	['chillstep']
13	Schwindelig - Original	['deep disco house', 'deep euro house', 'deep ...
14	I'll Be Your Reason	['bass trap', 'brostep', 'catstep', 'edm', 'el...
15	Sun Models (feat. Madelyn Grant)	['chillwave', 'edm', 'electronic trap', 'indie...
16	Unfold	['deep tropical house', 'downtempo', 'tropical...
17	Night - Lone Wolf Trait Remix	['bow pop', 'compositional ambient', 'minimal'...
18	Whyarntyou	['bass music', 'chillstep', 'future garage', '...
19	Feeling	['vapor twitch']
20	Ocelot	['chillstep', 'downtempo', 'electronic', 'nu j...

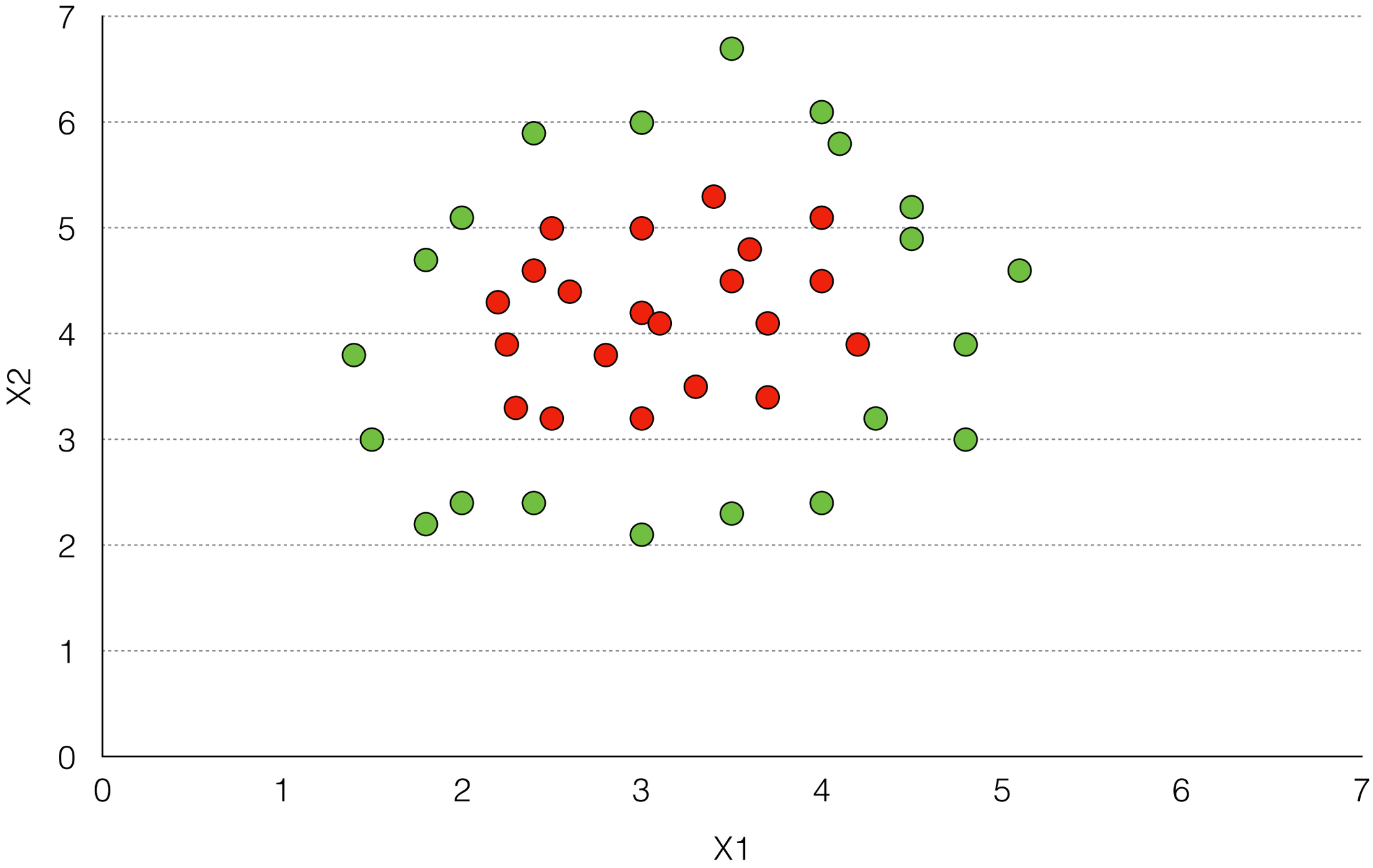
Top 30 Genres in My “Liked” and “Rejected” Playlists

	Liked Sub-Genres	Rejected Sub-Genres
1	(vapor twitch, 89)	(dubstep, 32)
2	(indietronica, 62)	(chillstep, 29)
3	(electronic, 55)	(microhouse, 27)
4	(trip hop, 52)	(bass music, 26)
5	(indie r&b, 49)	(minimal techno, 26)
6	(chillstep, 49)	(electronic, 25)
7	(chillwave, 46)	(house, 24)
8	(tropical house, 44)	(indietronica, 22)
9	(deep indie r&b, 43)	(tech house, 22)
10	(downtempo, 38)	(indie jazz, 21)
11	(nu jazz, 38)	(substep, 19)
12	(ninja, 34)	(fourth world, 18)
13	(wonky, 29)	(wonky, 18)
14	(substep, 27)	(indie r&b, 17)
15	(escape room, 26)	(future garage, 16)
16	(vapor soul, 25)	(tropical house, 16)
17	(bass music, 25)	(chillwave, 16)
18	(future garage, 24)	(alternative dance, 15)
19	(electronic trap, 21)	(nu jazz, 15)
20	(alternative dance, 18)	(ninja, 14)
21	(microhouse, 18)	(techno, 14)
22	(edm, 18)	(trip hop, 13)
23	(intelligent dance music, 15)	(new rave, 13)
24	(vaporwave, 15)	(compositional ambient, 13)
25	(new rave, 15)	(electro house, 12)
26	(brostep, 14)	(intelligent dance music, 12)
27	(minimal techno, 13)	(float house, 12)
28	(catstep, 12)	(minimal tech house, 12)
29	(glitch hop, 12)	(edm, 12)
30	(electro house, 12)	(deep melodic euro house, 11)


Top 30 sub-genres for "Liked" playlist
Number of Songs
0
20
40
60
80
100
vapor twitch
indietronica
electronic
trip hop
indie r&b
chillstep
chillwave
tropical house
deep indie r&b
downtempo
nu jazz
ninja
wonky
substep
escape room
vapor soul
bass music
future garage
electronic trap
alternative dance
microhouse
edm
intelligent dance music
vaporwave
new rave
brostep
minimal techno
catstep
glitch hop
electro house

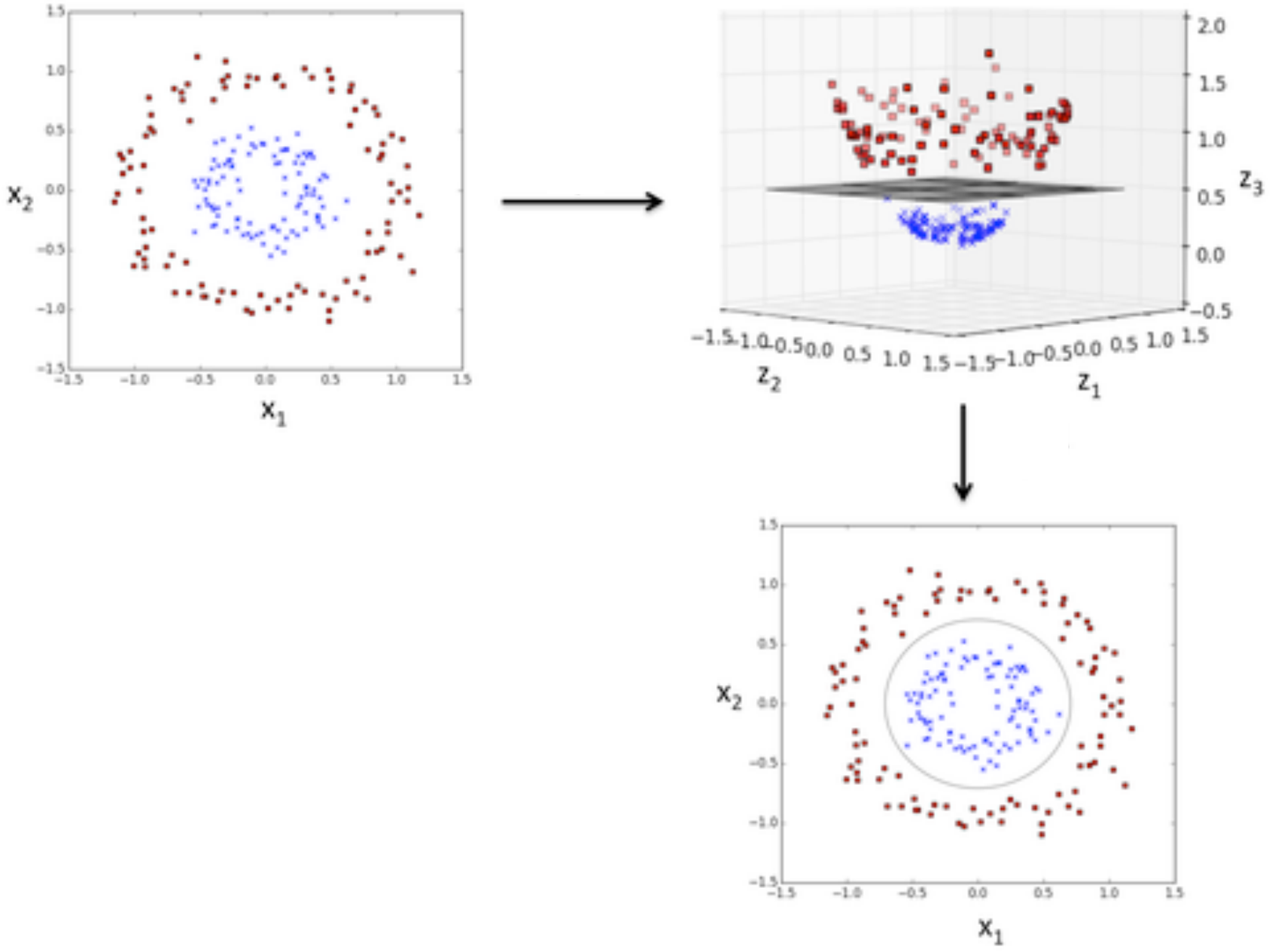

Number of Songs
0
20
40
60
80
100
vapor twitch
indietronica
electronic
trip hop
indie r&b
chillstep
chillwave
tropical house
deep indie r&b
downtempo
nu jazz
ninja
wonky
substep
escape room
vapor soul
bass music
future garage
electronic trap
alternative dance
microhouse
edm
intelligent dance music
vaporwave
new rave
brostep
minimal techno
catstep
glitch hop
electro house

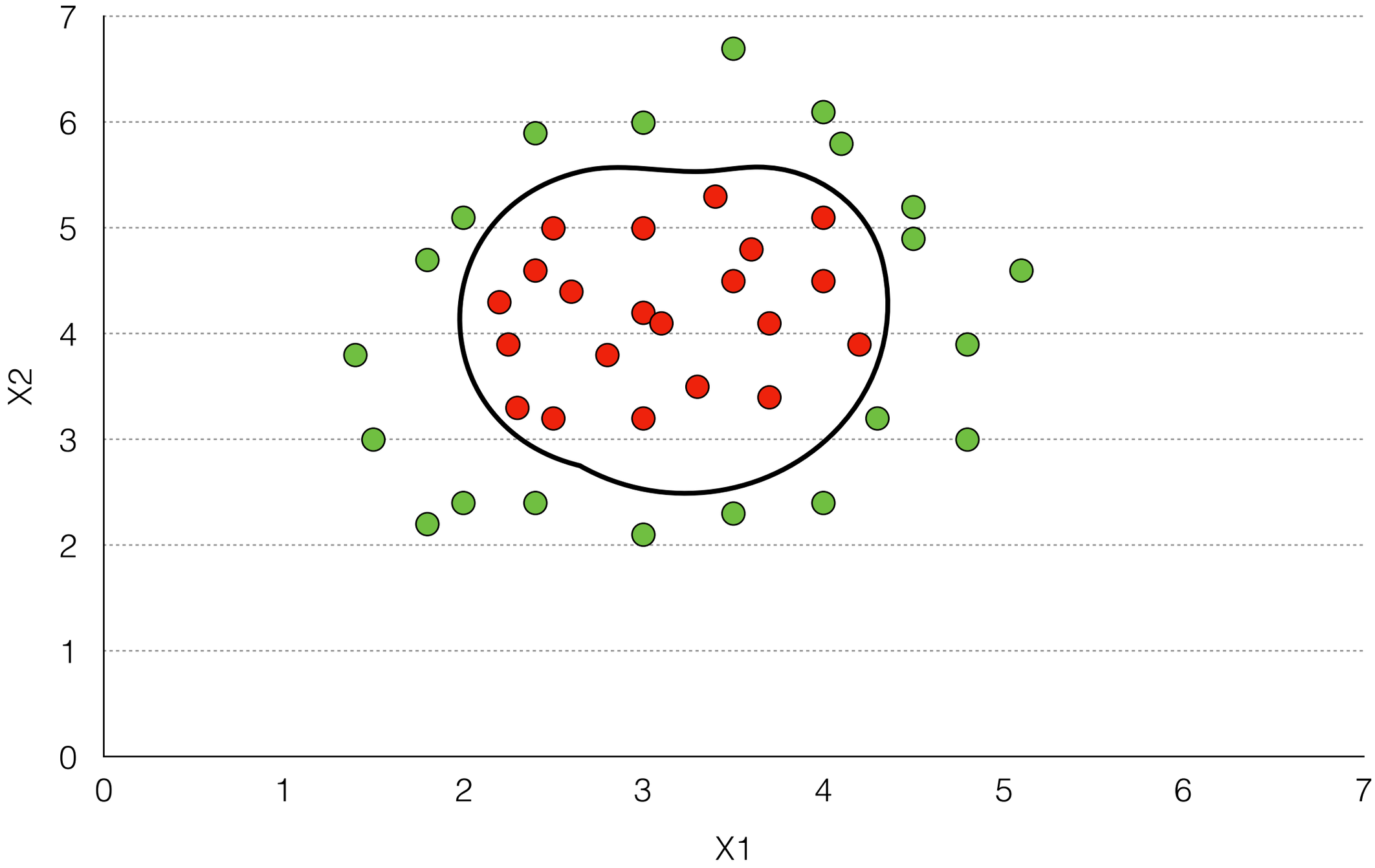

7
6
5
4
3
2
1
0
X2
0
1
2
3
4
5
6
7
X1

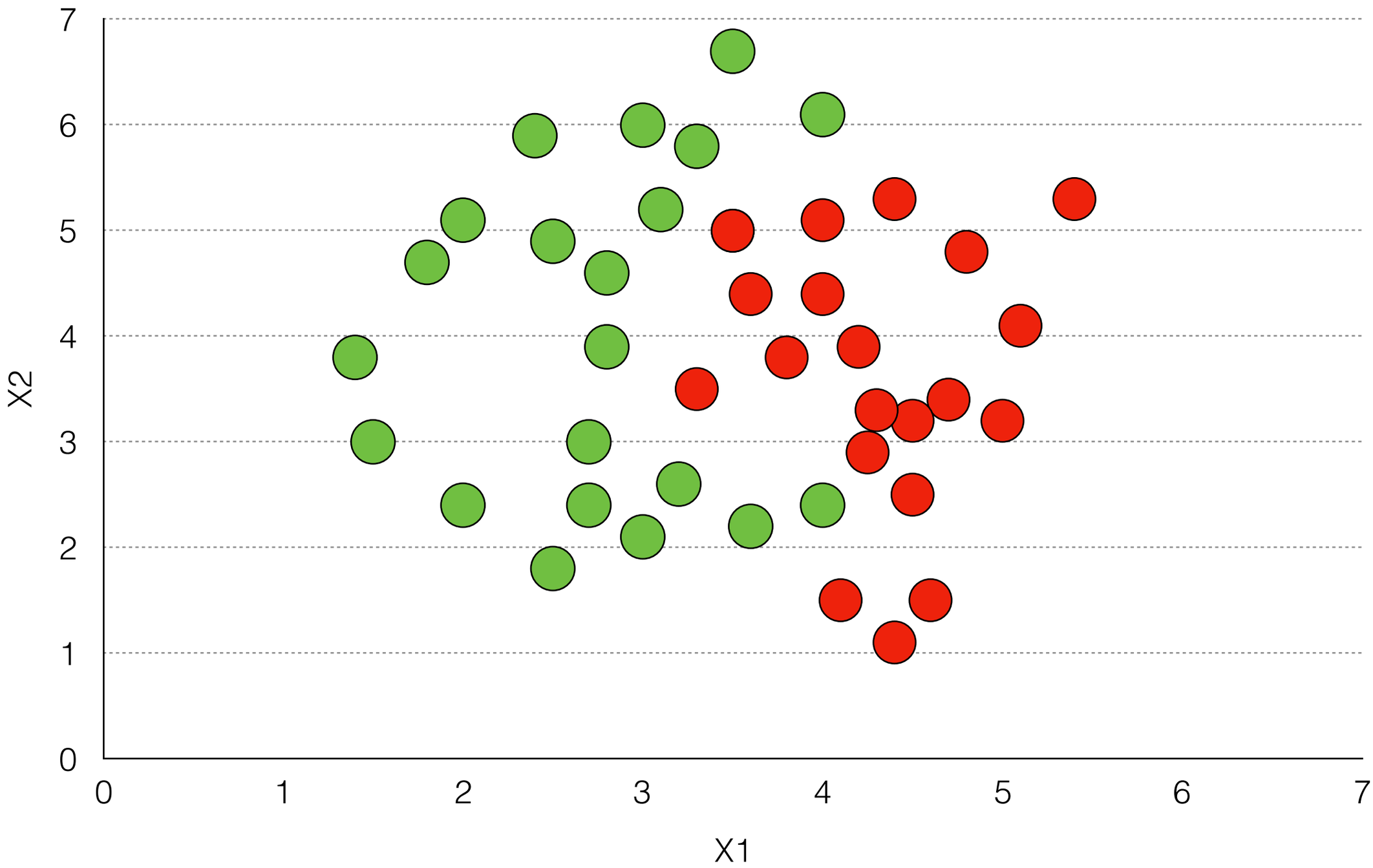

7
6
5
4
3
2
1
0
X2
0
1
2
3
4
5
6
7
X1

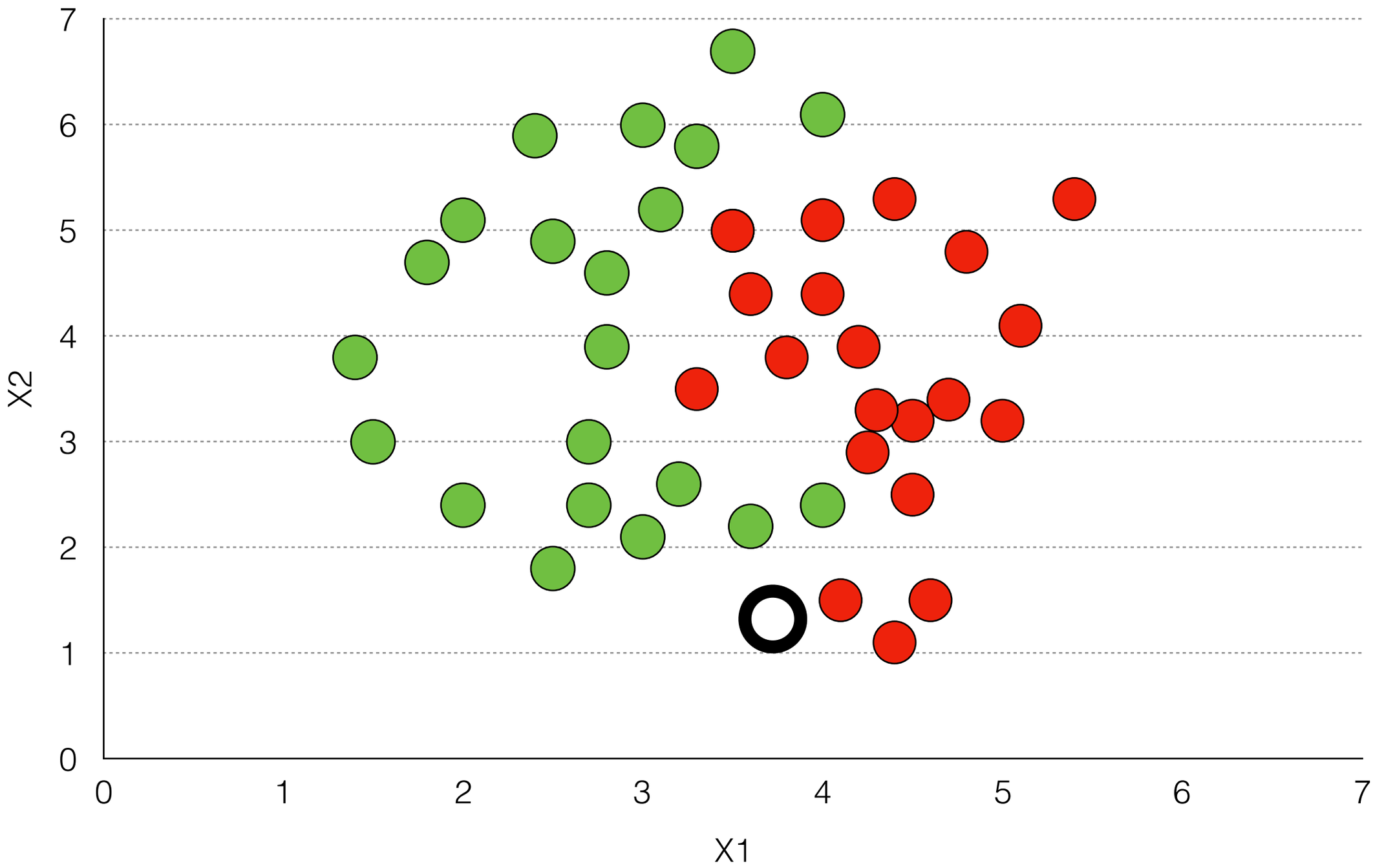
Kernel

non-linear classification

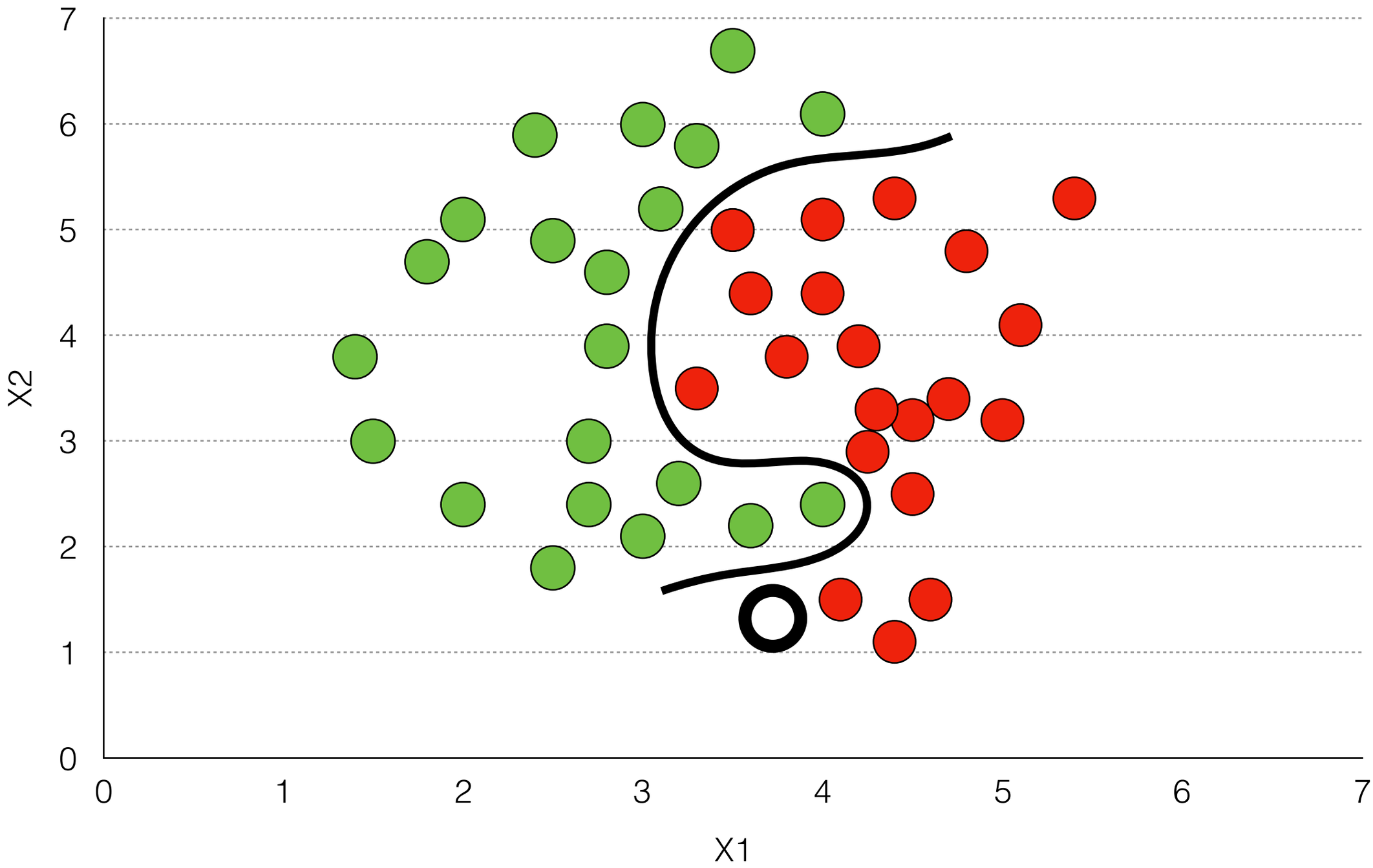

x_2
x_1
z_3
z_2
z_1
x_2
x_1


7
6
5
4
3
2
1
0
X2
0
1
2
3
4
5
6
7
X1

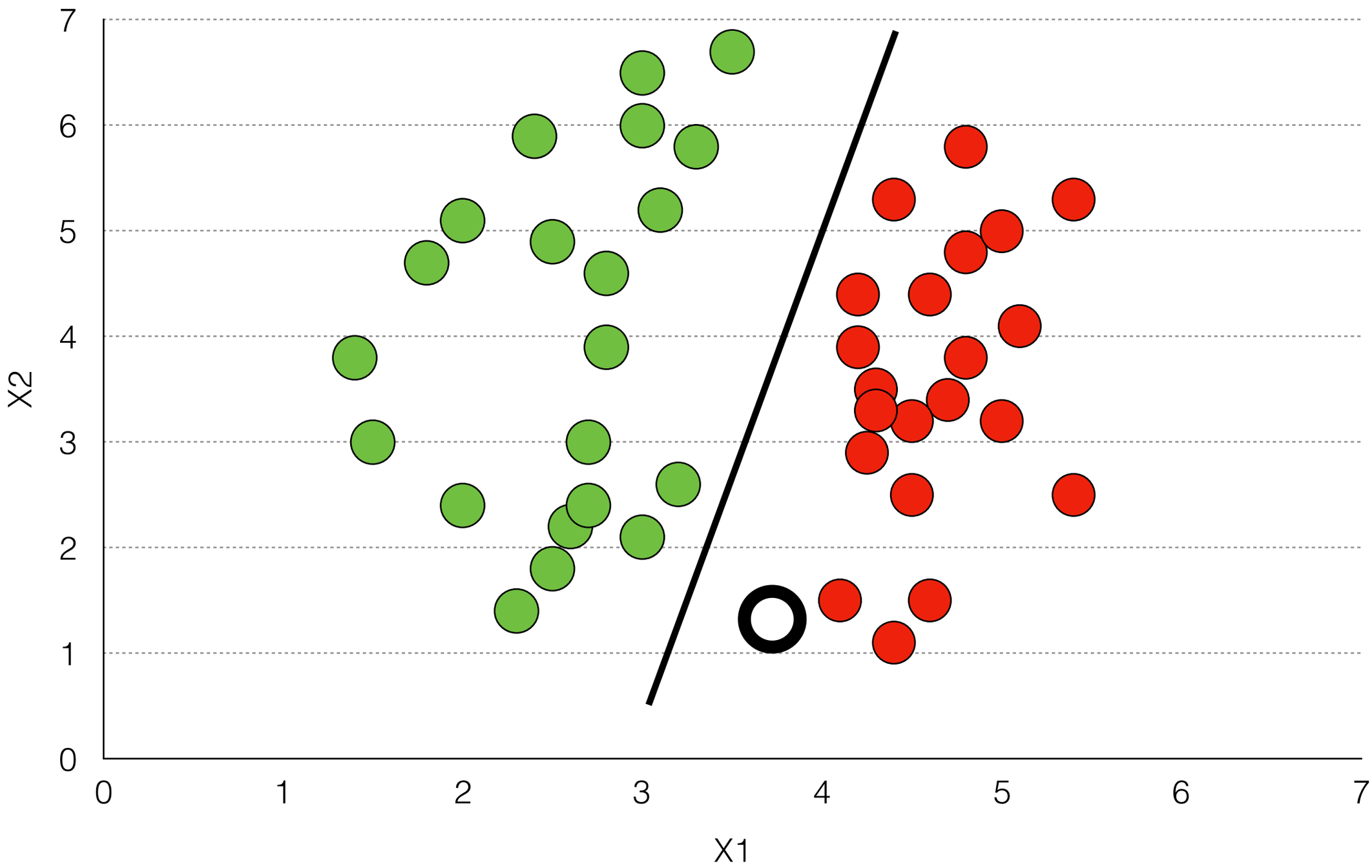

7
6
5
4
3
2
1
0
X2
0
1
2
3
4
5
6
7
X1


7
6
5
4
3
2
1
0
X2
0
1
2
3
4
5
6
7
X1

Input Space

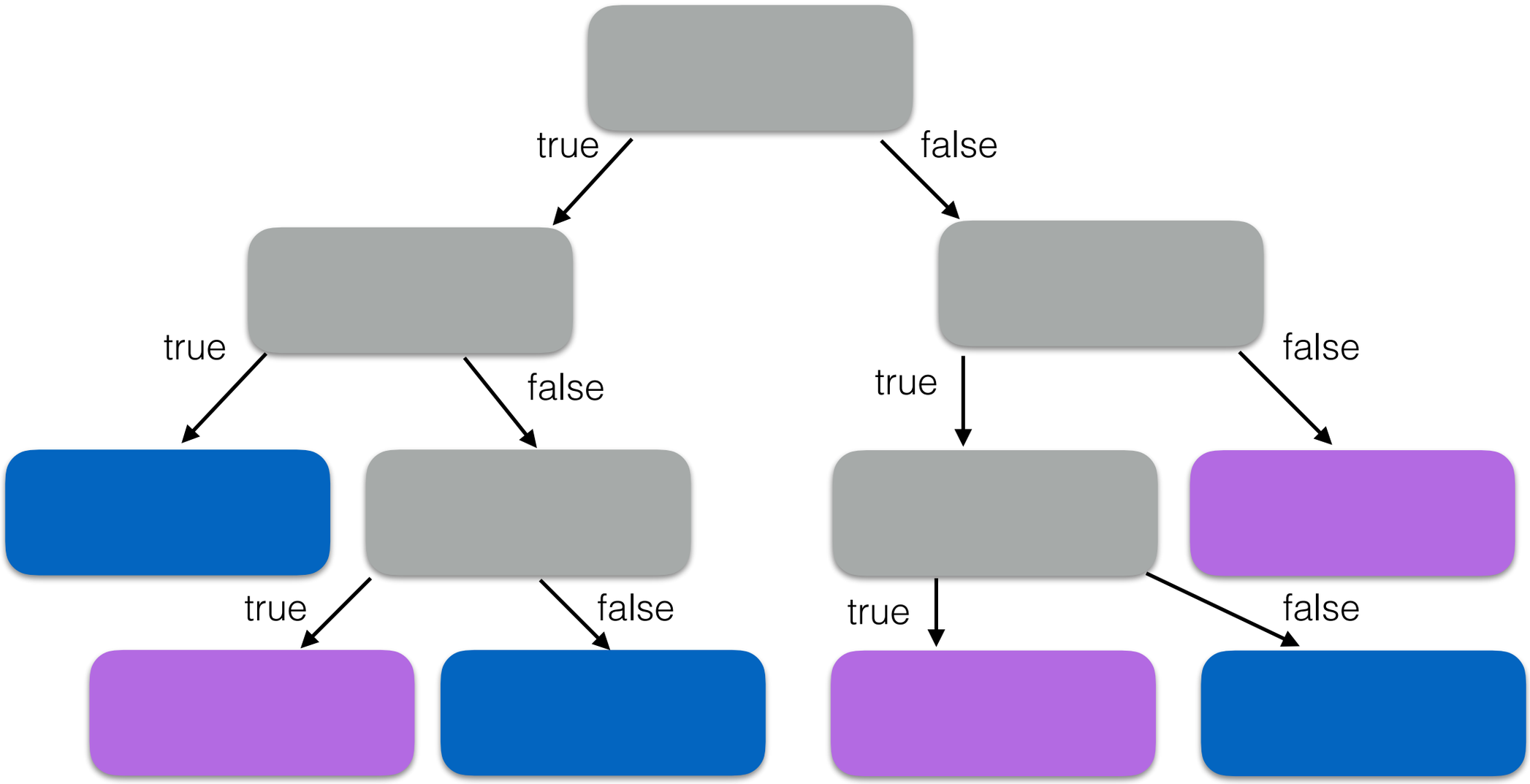


Feature Space



Decision Tree

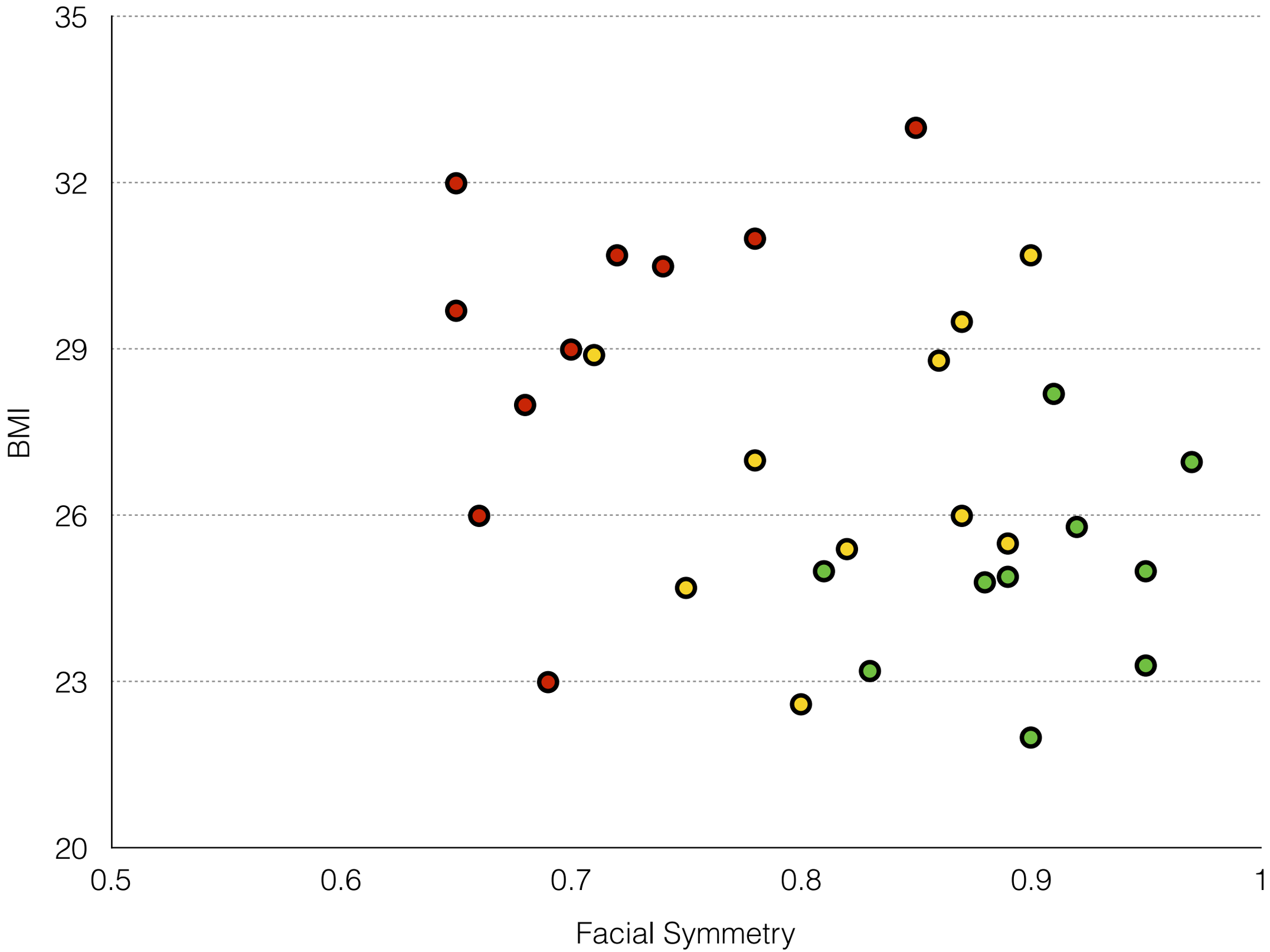
Decision Tree



Short-term Attractiveness

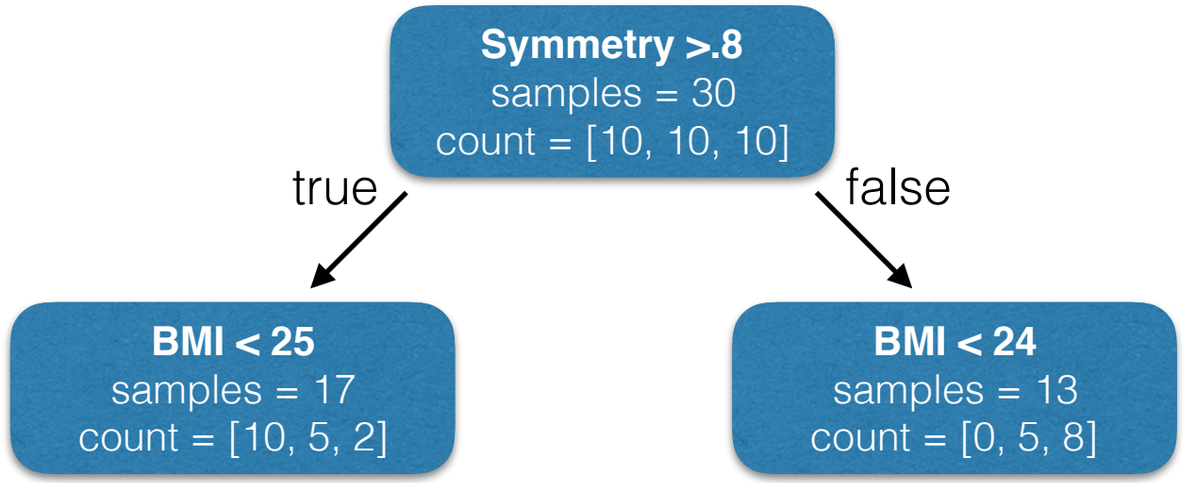
Short-term Attractiveness

Facial Symmetry	BMI	Waist-to-Hip	Well-Groomed
0.9	23.4	0.93	1
0.85	27.9	0.87	0
0.65	27.1	0.79	1
0.85	22.6	0.91	1
0.9	30.3	0.82	0
0.75	29.0	0.82	0
0.85	22.3	0.89	1
0.7	37.6	0.73	0
0.85	24.2	0.85	0

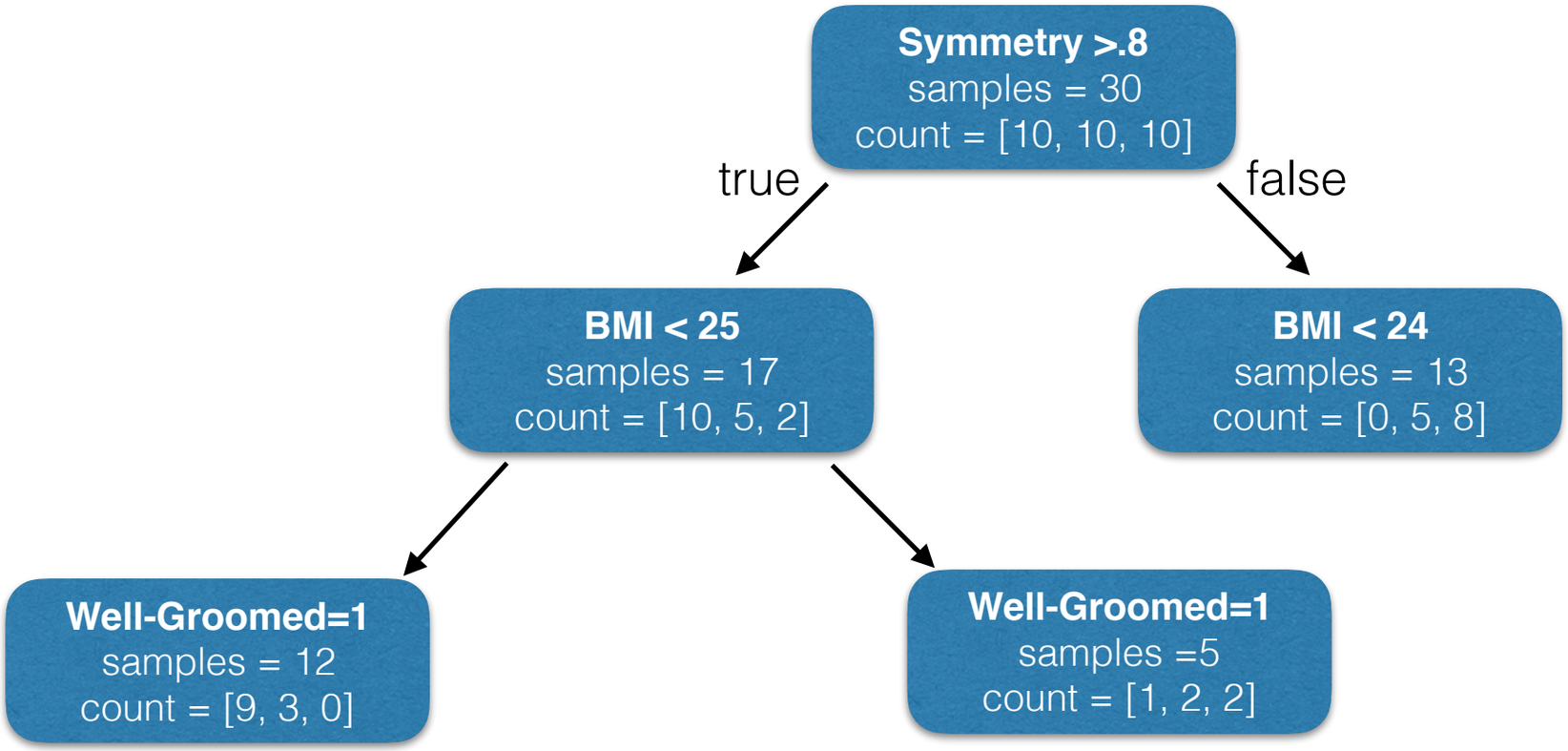

35
32
29
26
23
20
BMI
0.5
0.6
0.7
0.8
0.9
1
Facial Symmetry

Symmetry >.8
samples = 30
count = [10, 10, 10]

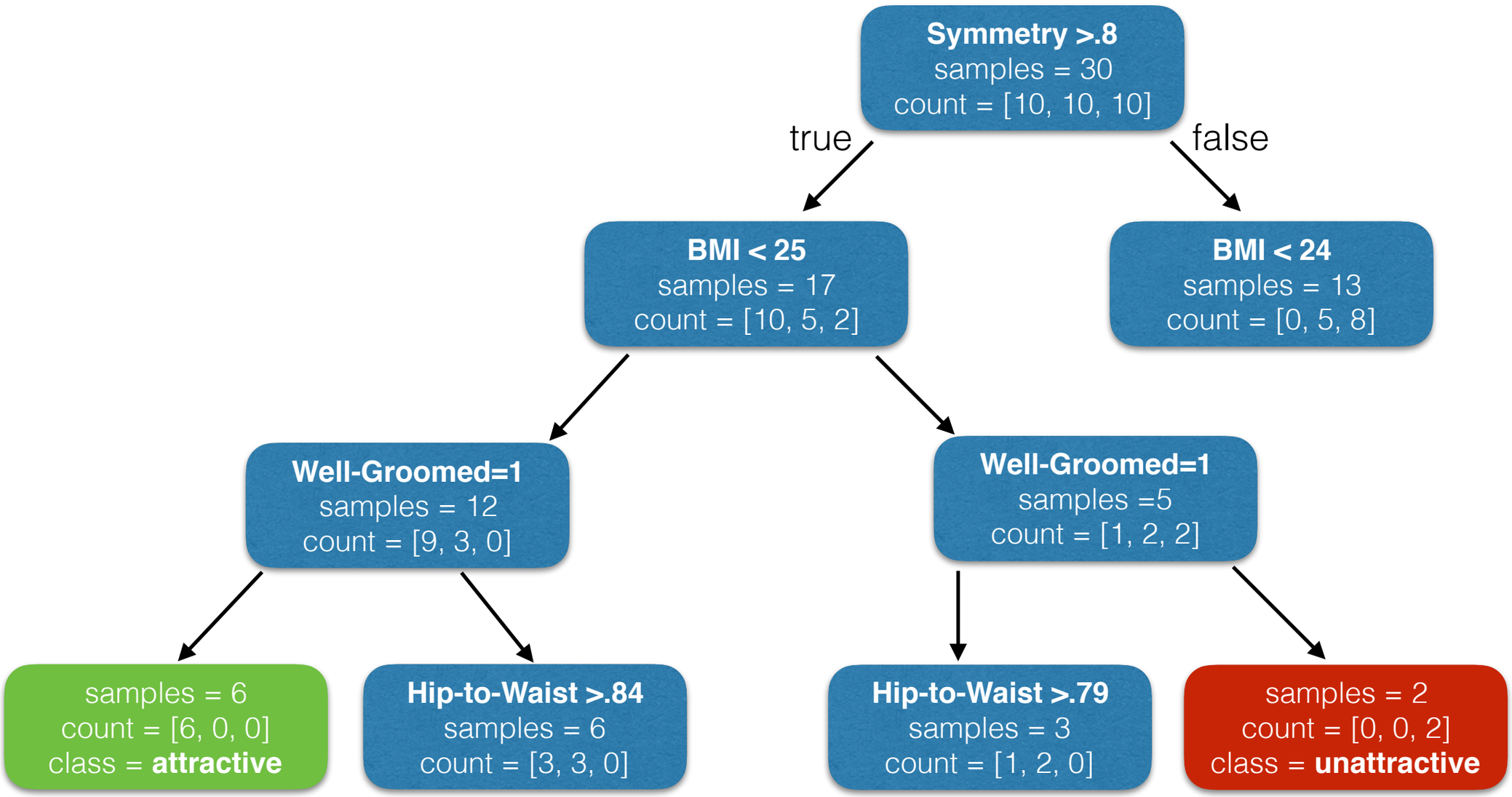
[att, ave, un]


Symmetry >.8
samples = 30
count = [10, 10, 10]
true
false
BMI < 25
samples = 17
count = [10, 5, 2]
BMI < 24
samples = 13
count = [0, 5, 8]

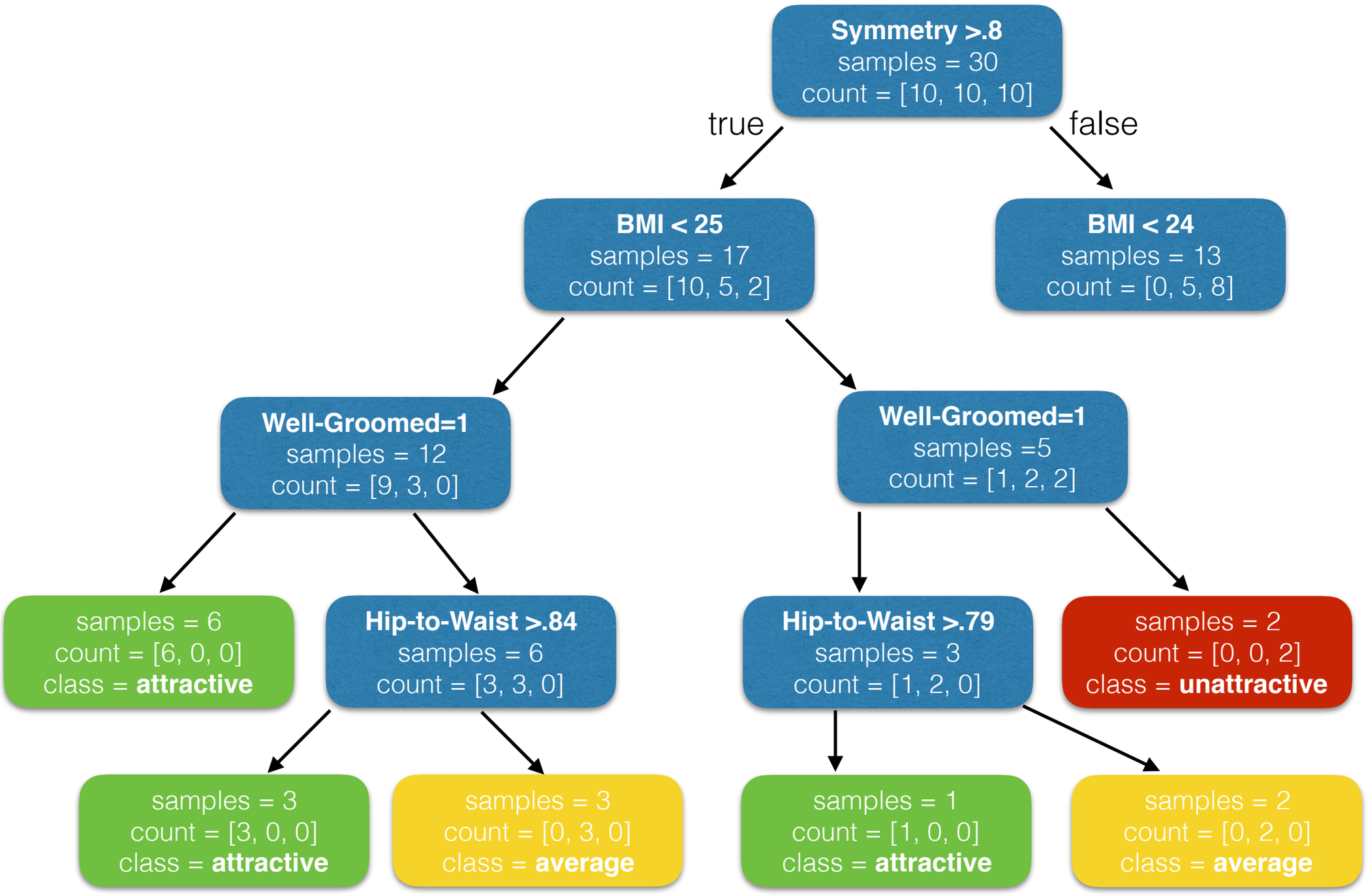
[att, ave, un]


Symmetry >.8
samples = 30
count = [10, 10, 10]
true
false
BMI < 25
samples = 17
count = [10, 5, 2]
BMI < 24
samples = 13
count = [0, 5, 8]
Well-Groomed=1
samples = 12
count = [9, 3, 0]
Well-Groomed=1
samples =5
count = [1, 2, 2]

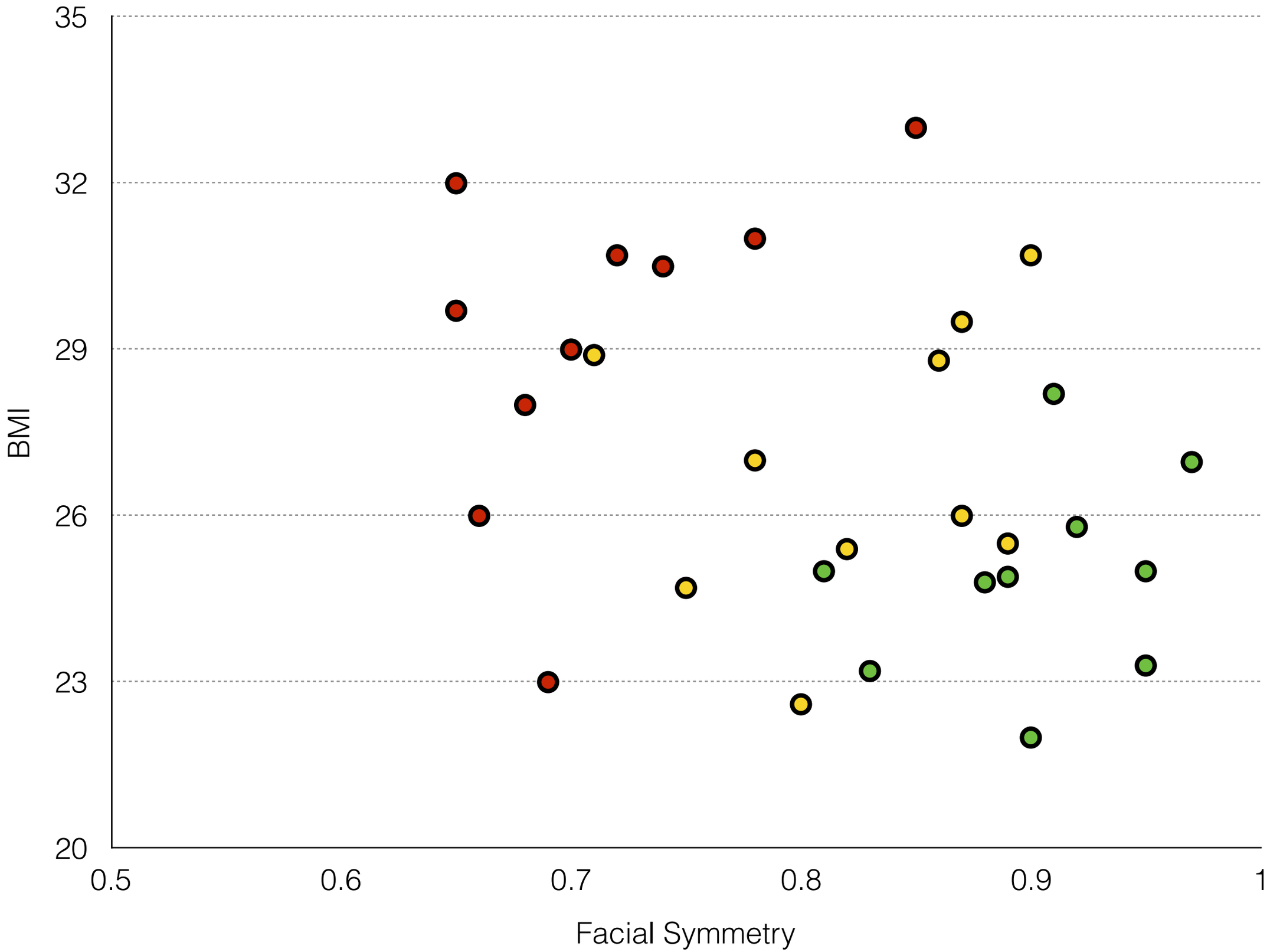
[att, ave, un]

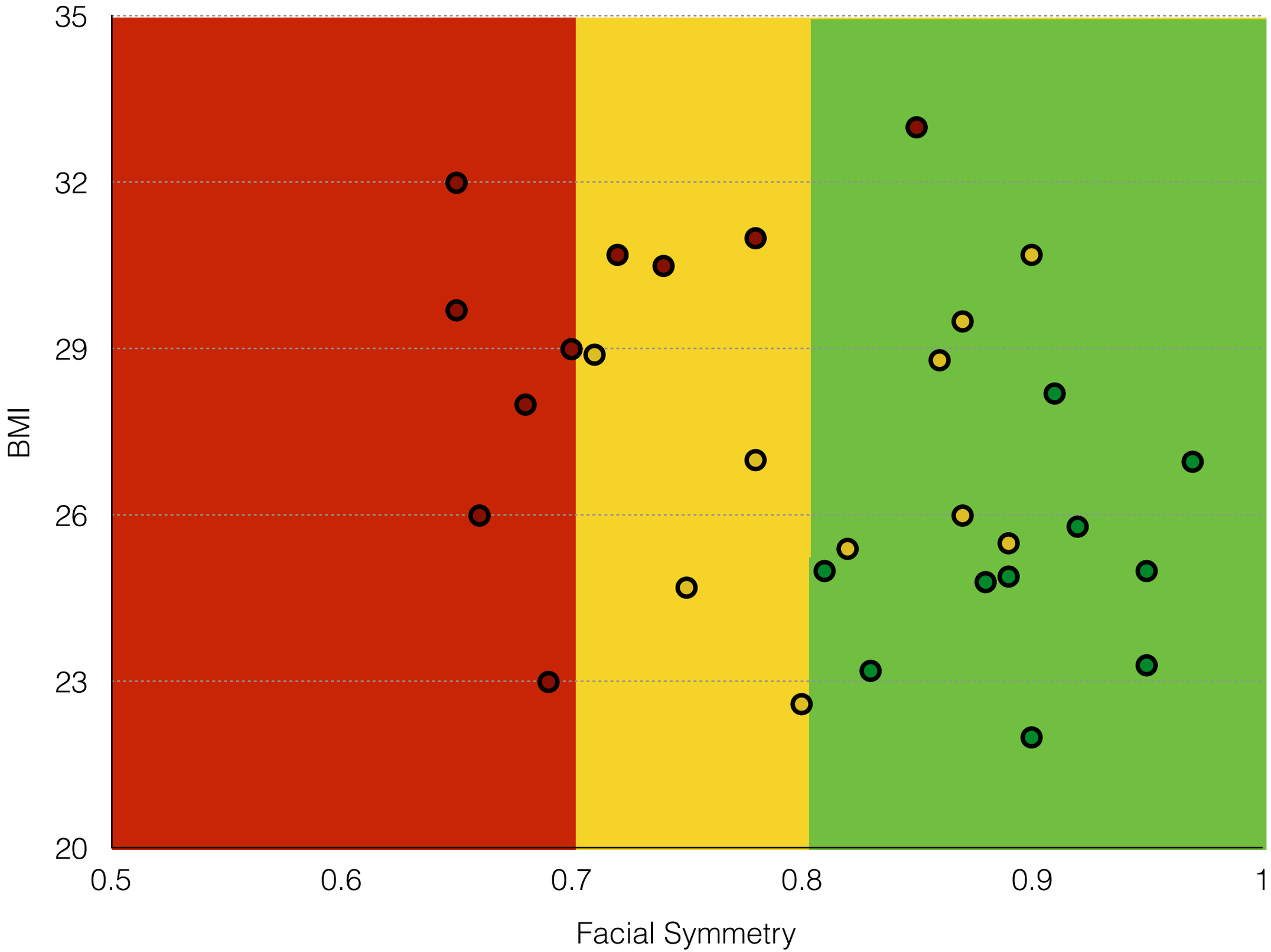

Symmetry >.8
samples = 30
count = [10, 10, 10]
true
false
BMI < 25
samples = 17
count = [10, 5, 2]
BMI < 24
samples = 13
count = [0, 5, 8]
Well-Groomed=1
samples = 12
count = [9, 3, 0]
Well-Groomed=1
samples =5
count = [1, 2, 2]
samples = 6
count = [6, 0, 0]
class = attractive
Hip-to-Waist >.84
samples = 6
count = [3, 3, 0]
Hip-to-Waist >.79
samples = 3
count = [1, 2, 0]
samples = 2
count = [0, 0, 2]
class = unattractive

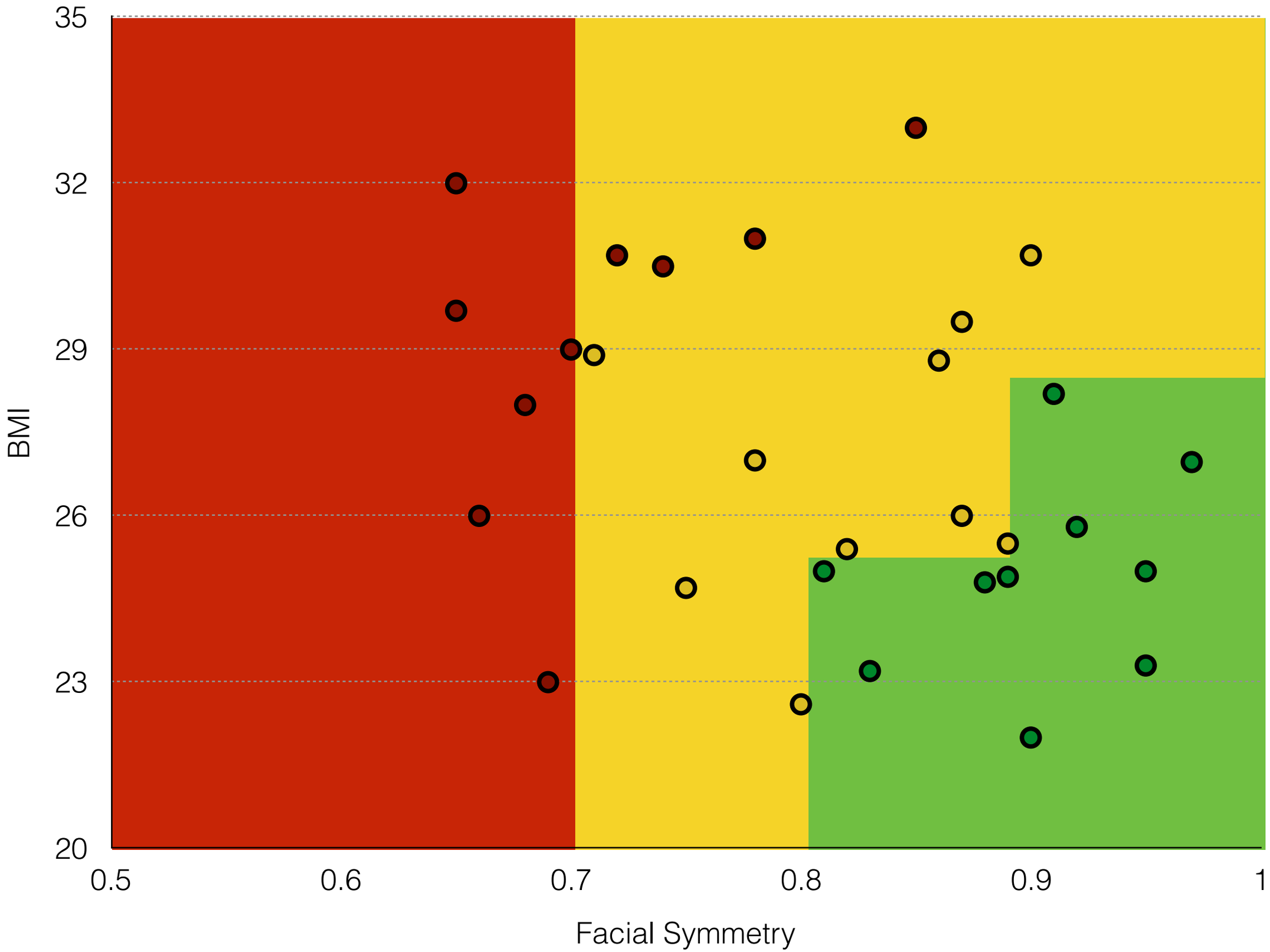
[att, ave, un]

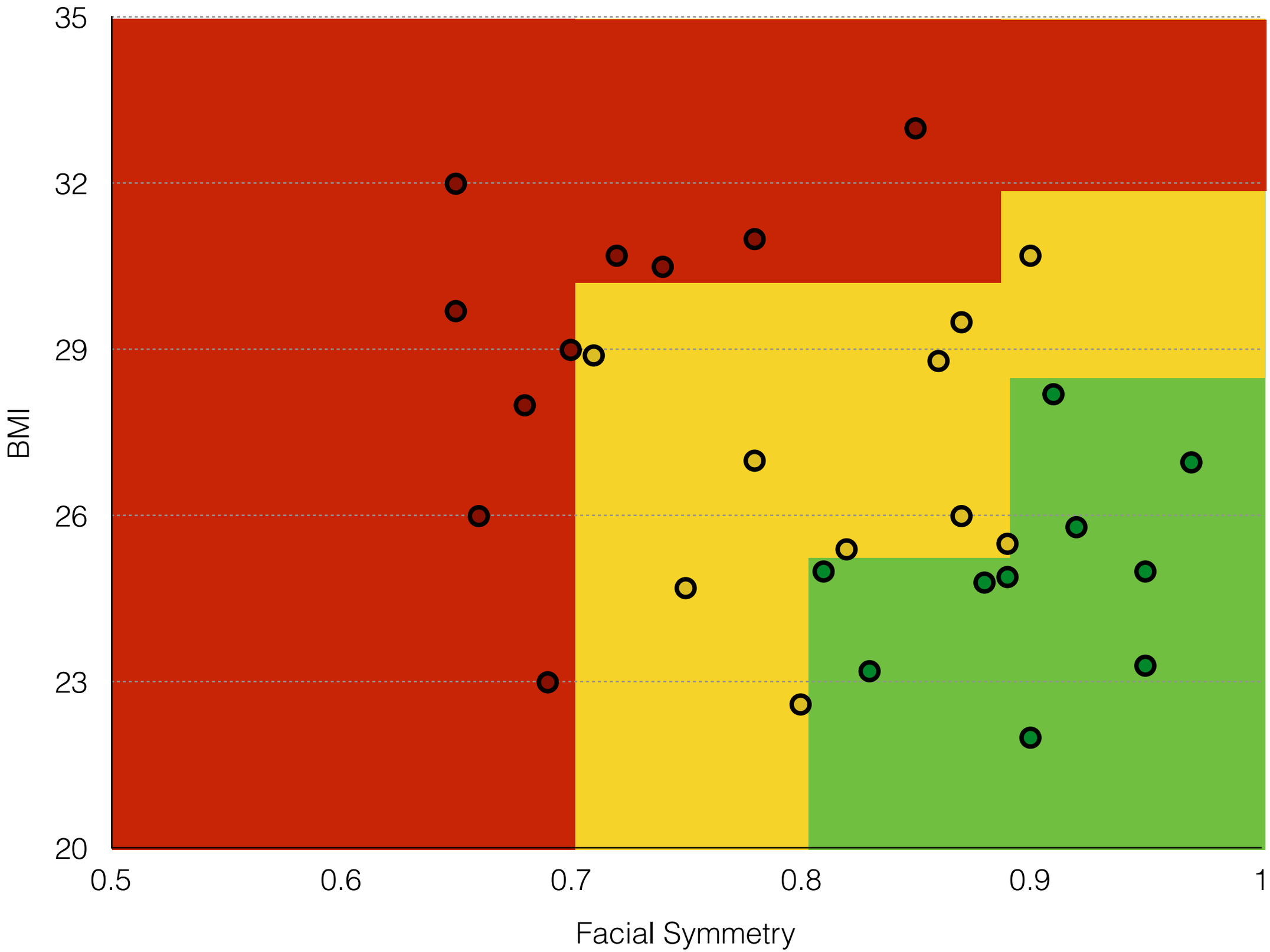

Symmetry >.8
samples = 30
count = [10, 10, 10]
true
false
BMI < 25
samples = 17
count = [10, 5, 2]
BMI < 24
samples = 13
count = [0, 5, 8]
Well-Groomed=1
samples = 12
count = [9, 3, 0]
Well-Groomed=1
samples =5
count = [1, 2, 2]
samples = 6
count = [6, 0, 0]
class = **attractive**
Hip-to-Waist >.84
samples = 6
count = [3, 3, 0]
Hip-to-Waist >.79
samples = 3
count = [1, 2, 0]
samples = 2
count = [0, 0, 2]
class = **unattractive**
samples = 3
count = [3, 0, 0]
class = **attractive**
samples = 3
count = [0, 3, 0]
class = **average**
samples = 1
count = [1, 0, 0]
class = **attractive**
samples = 2
count = [0, 2, 0]
class = **average**

[att, ave, un]

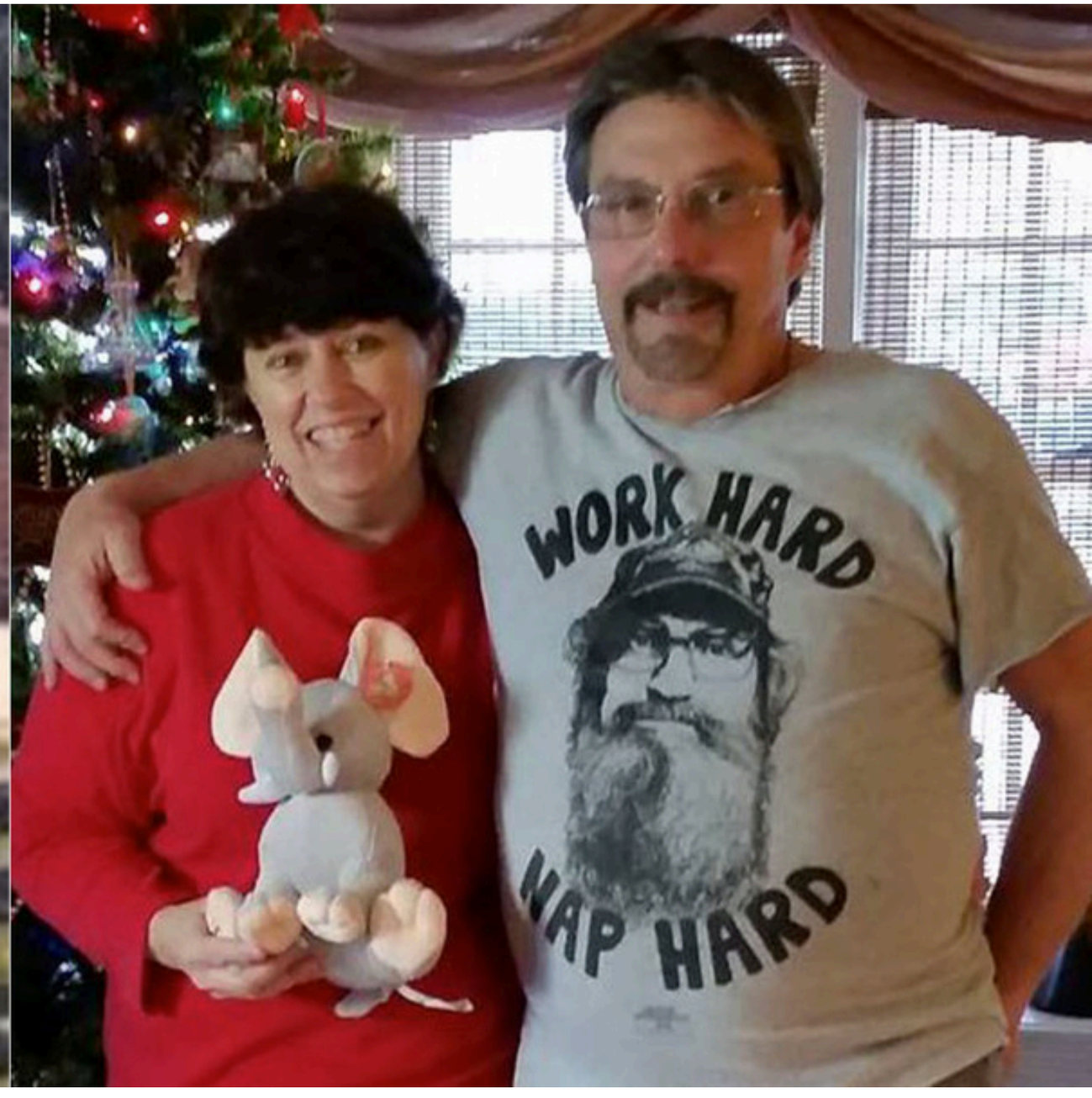

35
32
29
26
23
20
BMI
0.5
0.6
0.7
0.8
0.9
1
Facial Symmetry


35
32
29
26
23
20
BMI
0.5
0.6
0.7
0.8
0.9
1
Facial Symmetry


35
32
29
26
23
20
BMI
0.5
0.6
0.7
0.8
0.9
1
Facial Symmetry


35
32
29
26
23
20
BMI
0.5
0.6
0.7
0.8
0.9
1
Facial Symmetry

K-nearest Neighbor


WORK HARD
NAP HARD

	Number of Relationships	Grade of First Relationship
sample a	3	7
sample b	6	11
sample c	3	9
sample d	5	10

Euclidean Distance

point a = [a1, a2]

point b = [b1, b2]

Euclidean Distance

point a = [a1, a2]

point b = [b1, b2]

Two dimensions (features)

$$\sqrt{(a_1-b_1)^2+(a_2-b_2)^2}$$

Euclidean Distance

Feature Vector

a = [3, 7]

b = [6, 11]

Euclidean Distance

$$\sqrt{(3-6)^2+(7-11)^2}$$

Feature Vector

a = [3, 7]

b = [6, 11]

Euclidean Distance

$$\sqrt{(3-6)^2+(7-11)^2}$$

$$\sqrt{(-3)^2+(-4)^2}$$

Feature Vector

a = [3, 7]

b = [6, 11]

Euclidean Distance

Feature Vector

a = [3, 7]

b = [6, 11]

$$\sqrt{(3-6)^2+(7-11)^2}$$

$$\sqrt{(-3)^2+(-4)^2}$$

$$\sqrt{9+16}$$

Euclidean Distance

Feature Vector

a = [3, 7]

b = [6, 11]

$$\sqrt{(3-6)^2+(7-11)^2}$$

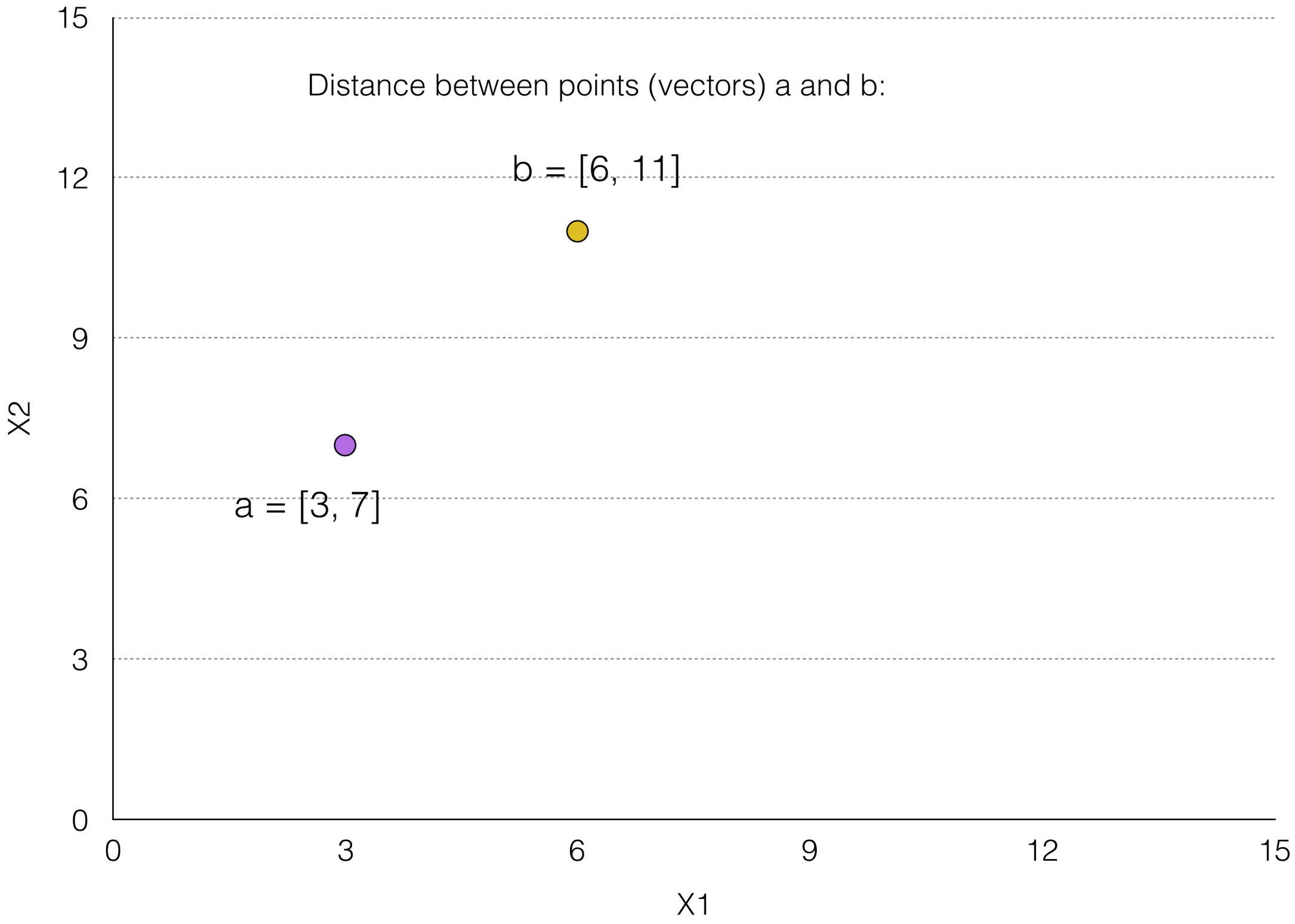
$$\sqrt{(-3)^2+(-4)^2}$$

$$\sqrt{9+16}$$

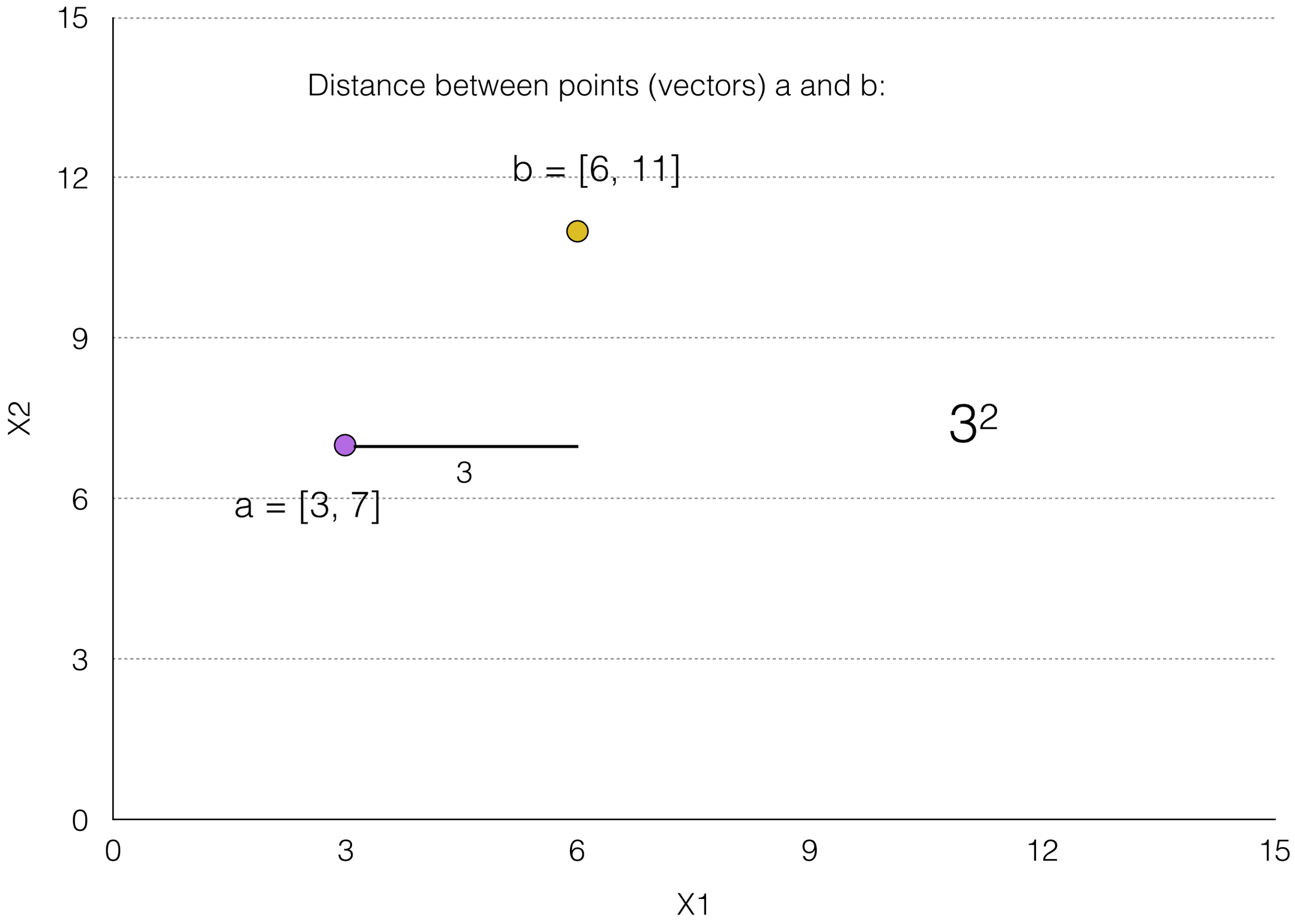
Distance between points (vectors) a and b:

$$\sqrt{25}=5$$

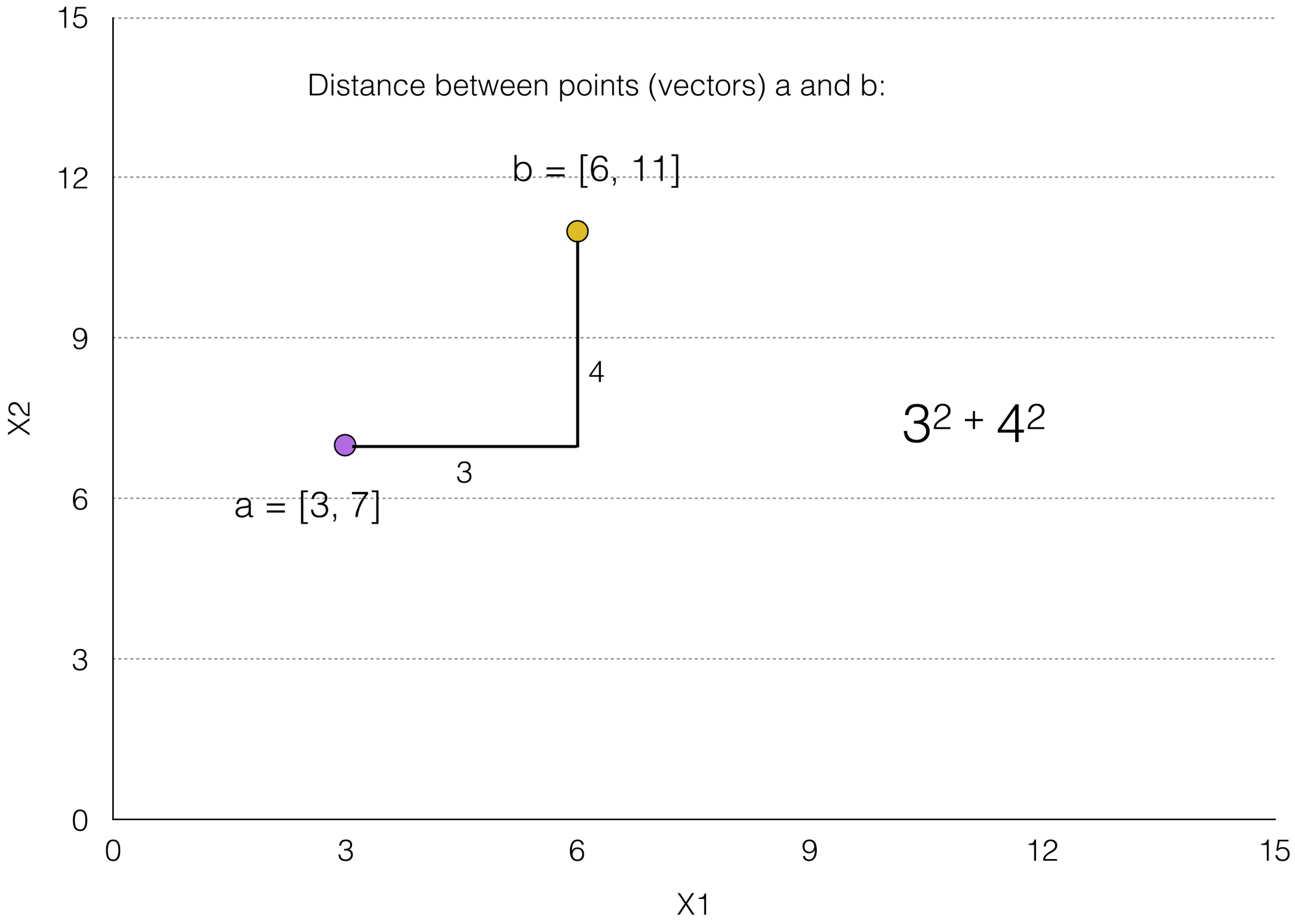
Euclidean Distance



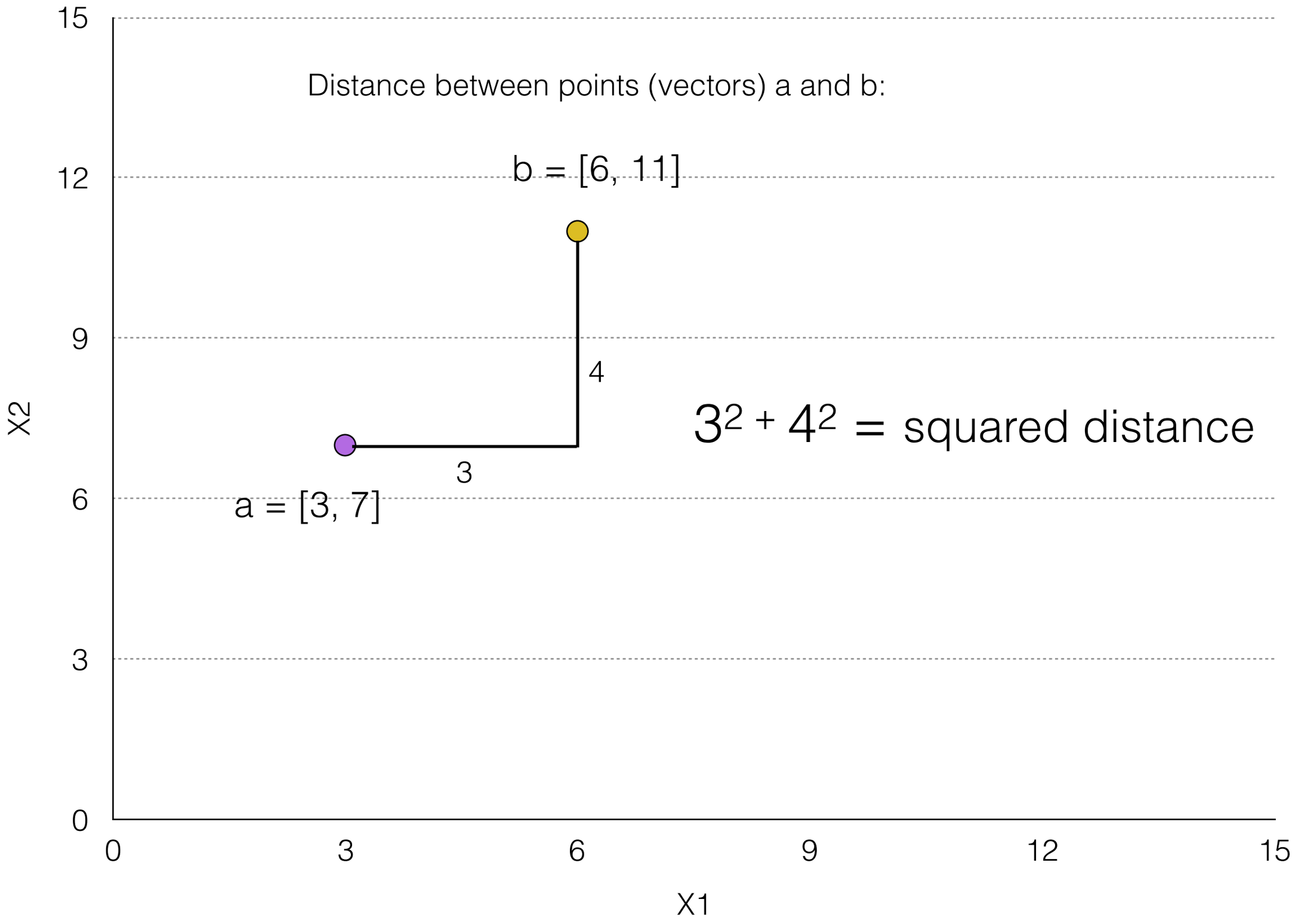
Euclidean Distance



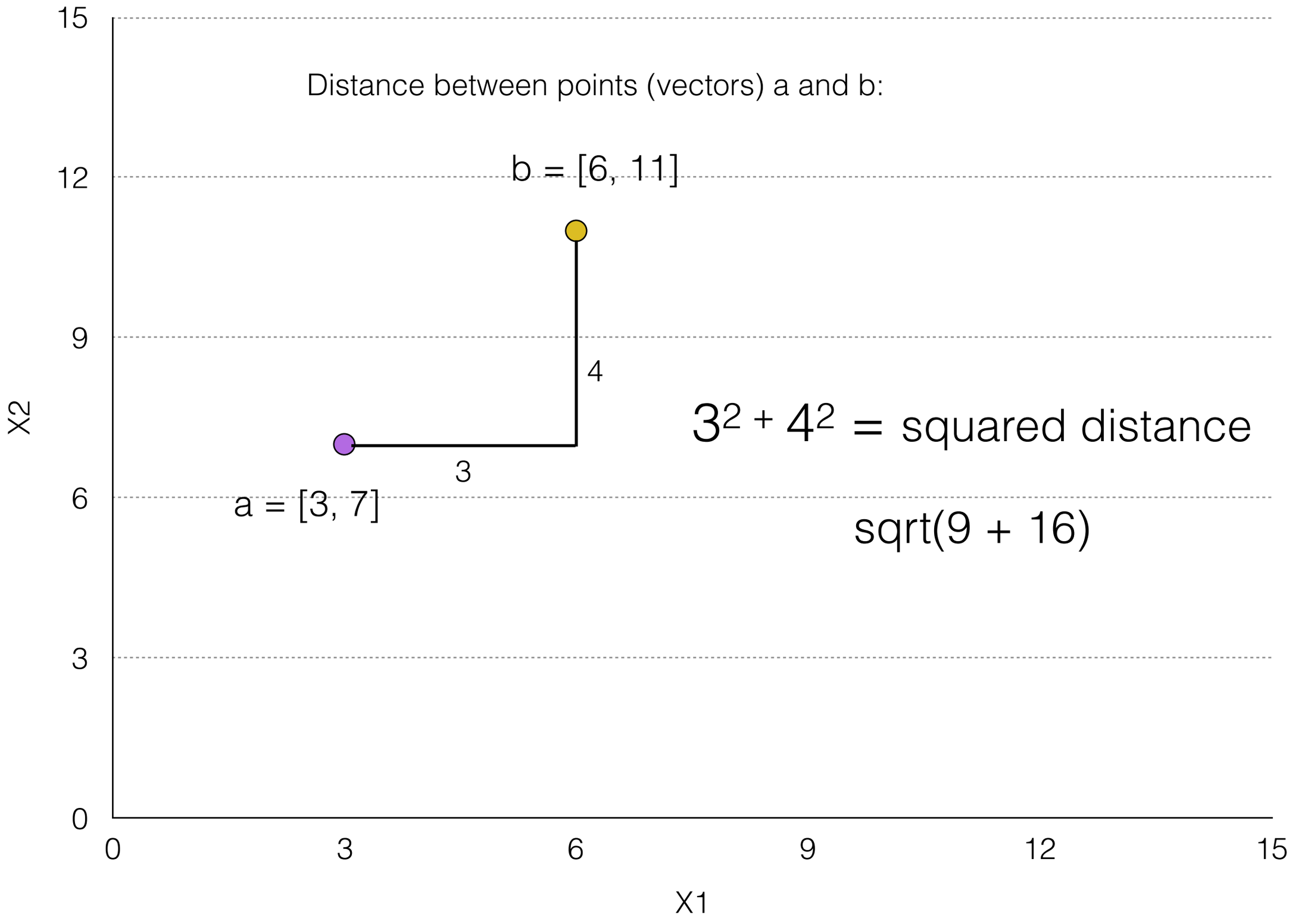
Euclidean Distance



Euclidean Distance



Euclidean Distance



Euclidean Distance

Distance between points (vectors) a and b:

b = [6, 11]

distance = **5**

4

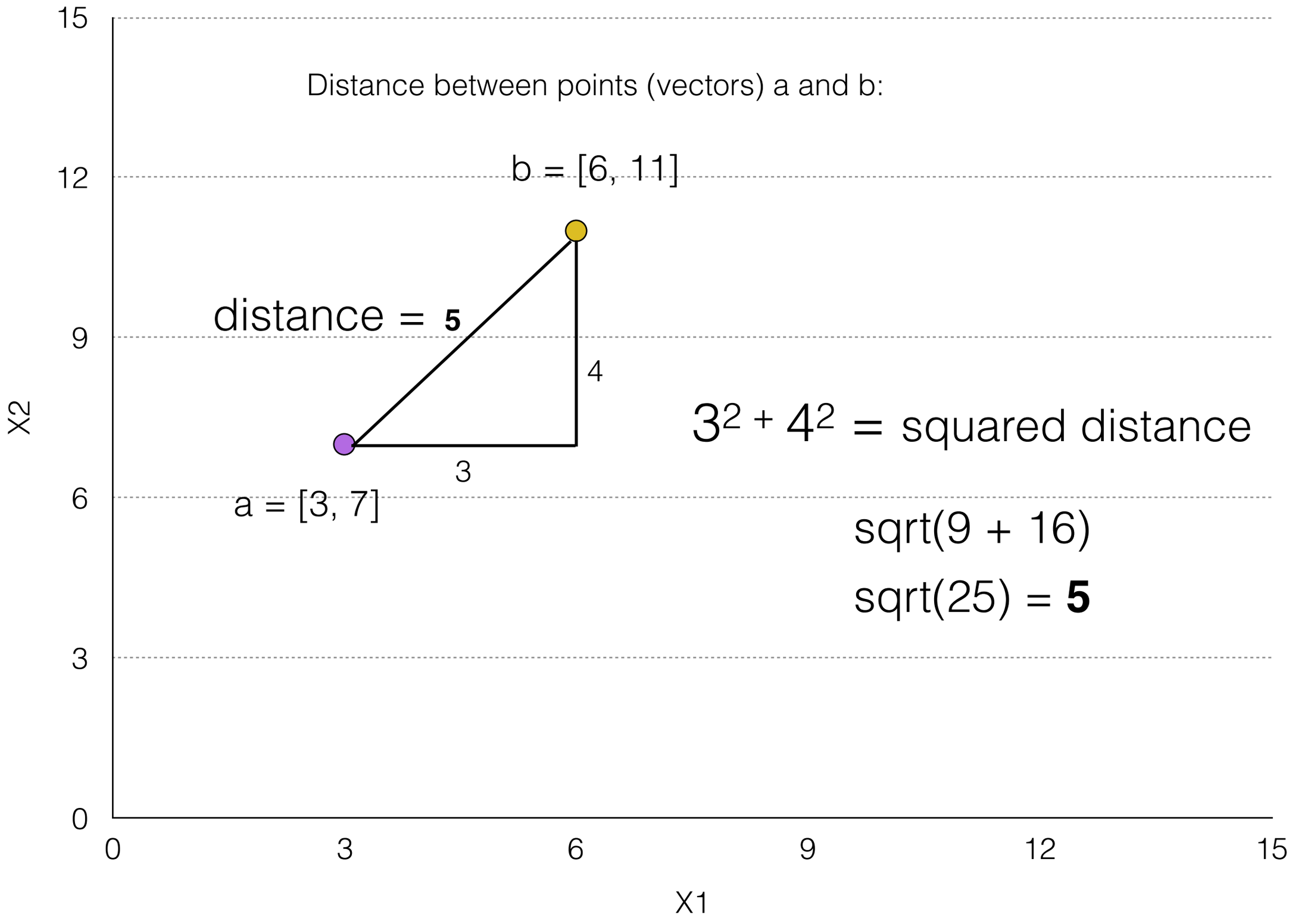
3

a = [3, 7]

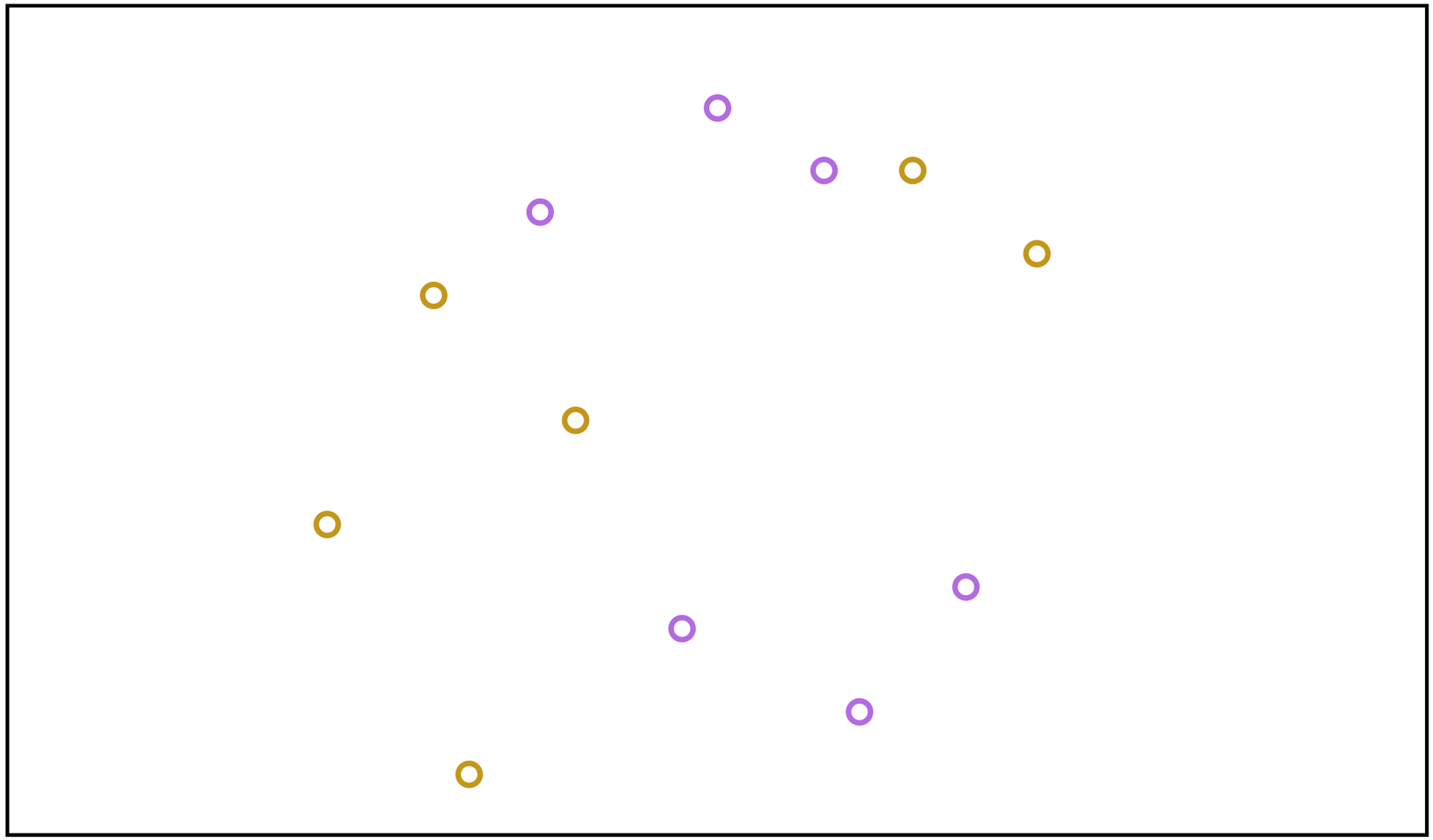
$3^2 + 4^2$ = squared distance

sqrt(9 + 16)

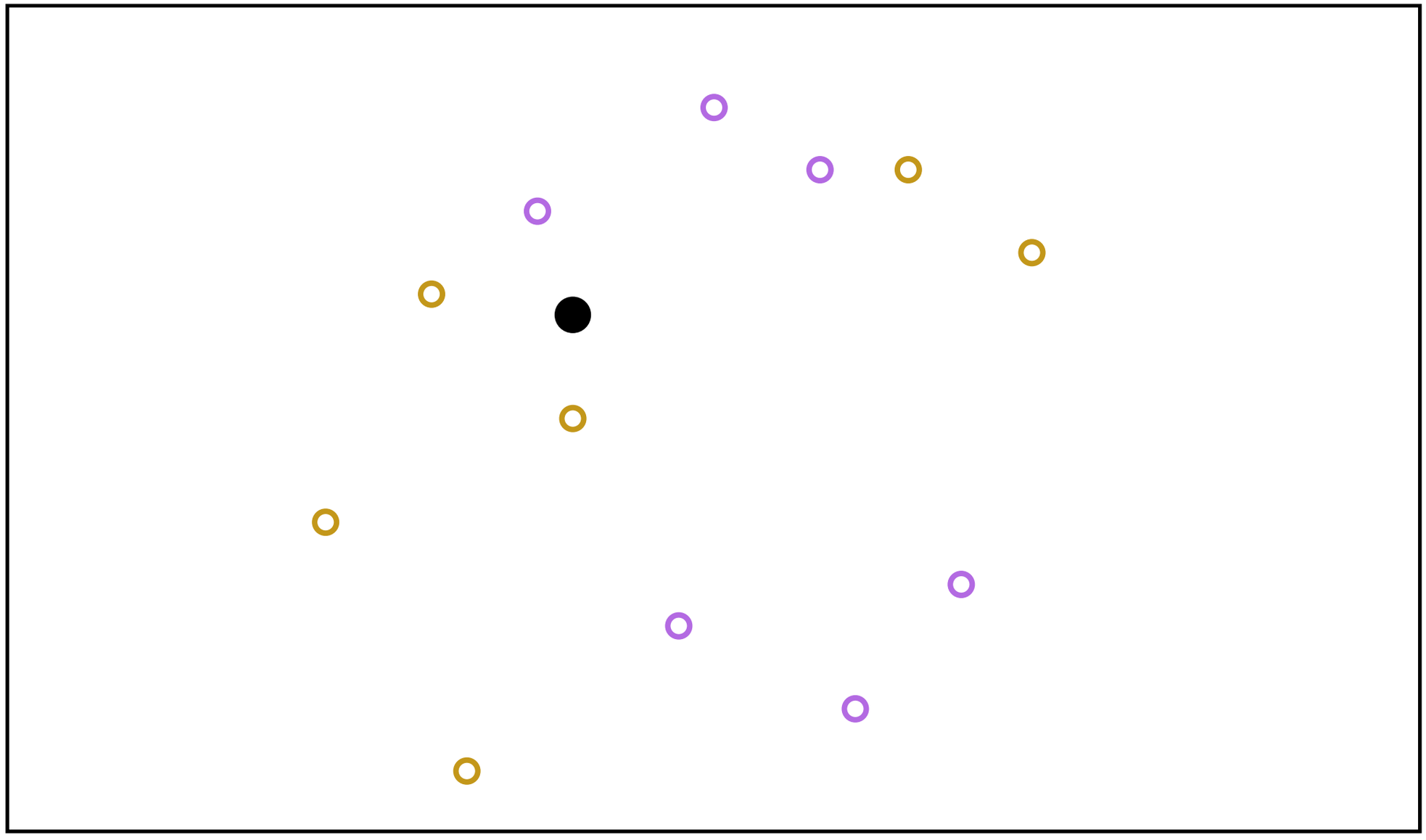
sqrt(25) = **5**



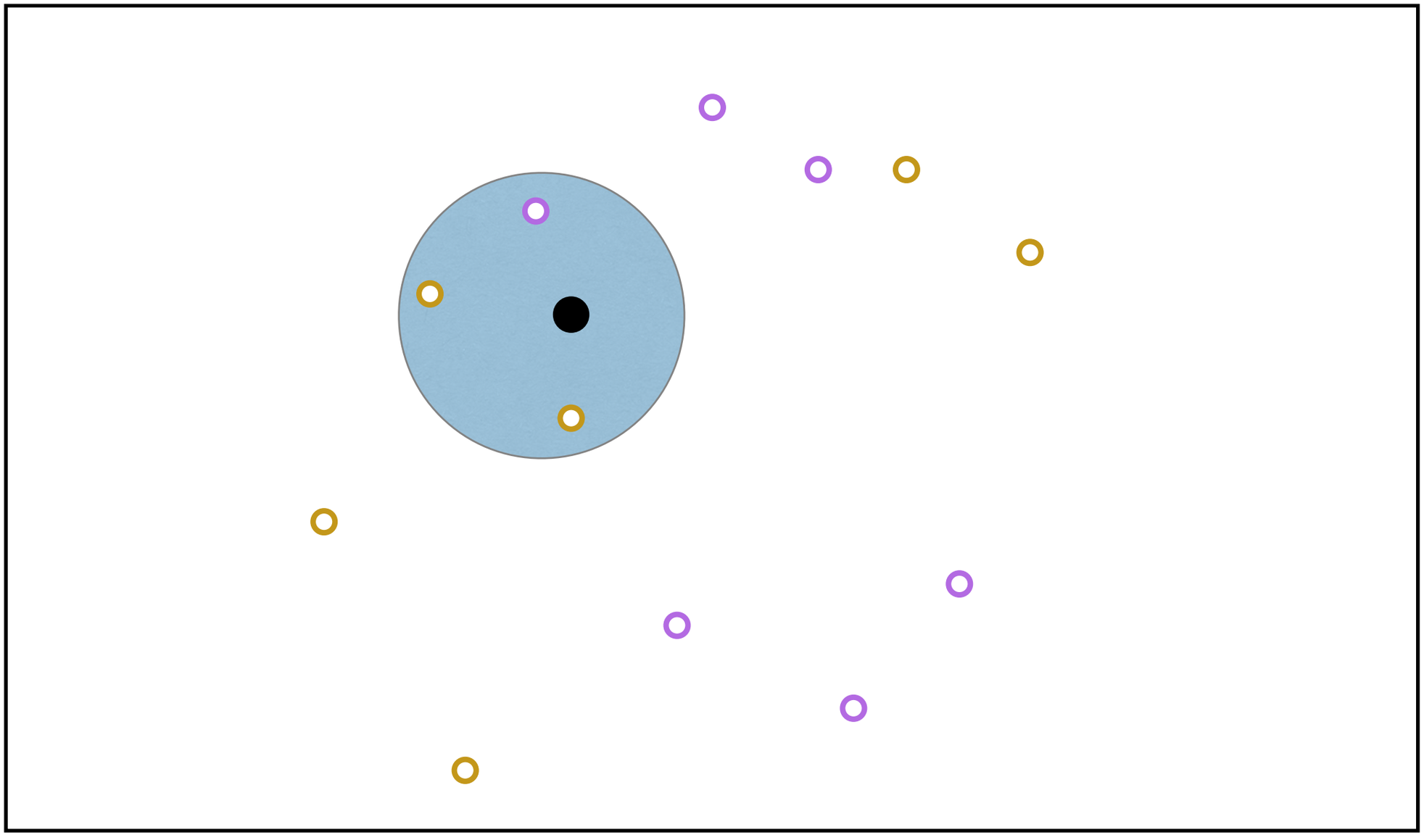
K-nearest Neighbor



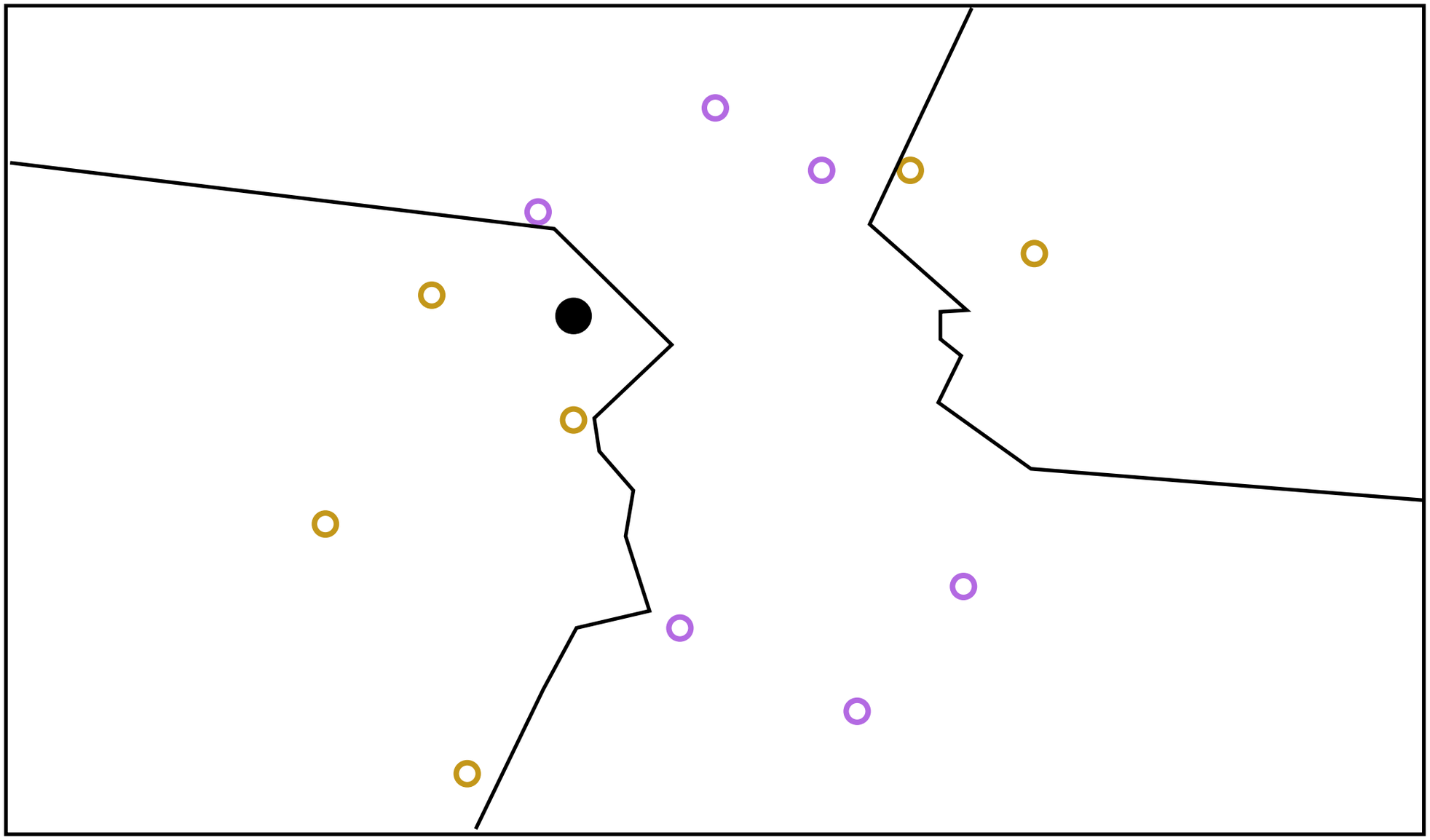
K-nearest Neighbor



K = 3



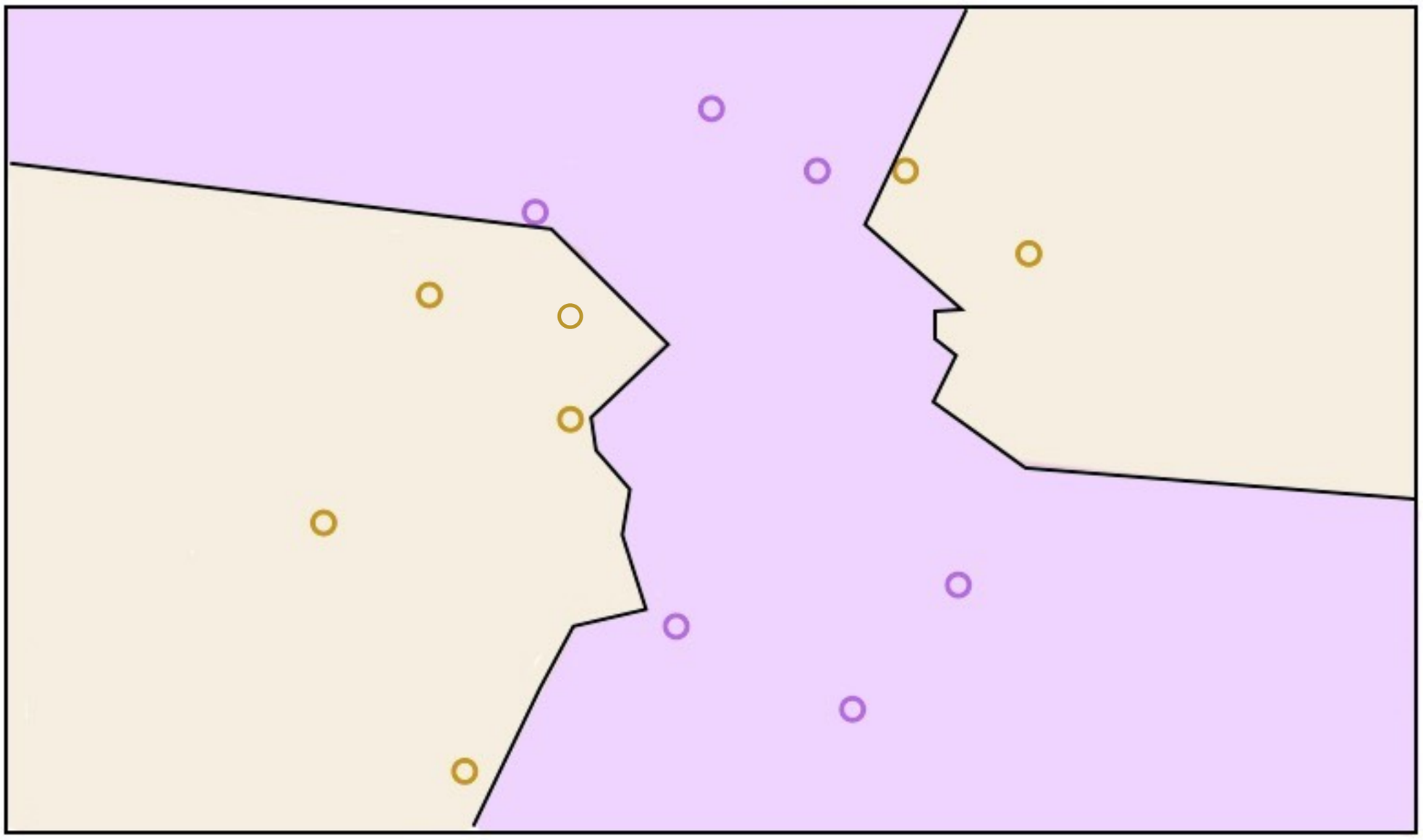
K = 3



Stays together

Breaks up

K = 3



K-Means

Scikit-learn Intro

```
model.fit(train_X, train_y)
```

```
model.fit(train_X, train_y)

model.predict(test_X)
```

```
model.fit(train_X, train_y)

model.predict(test_X)

model.score(test_X, test_y)
```

Iris Data