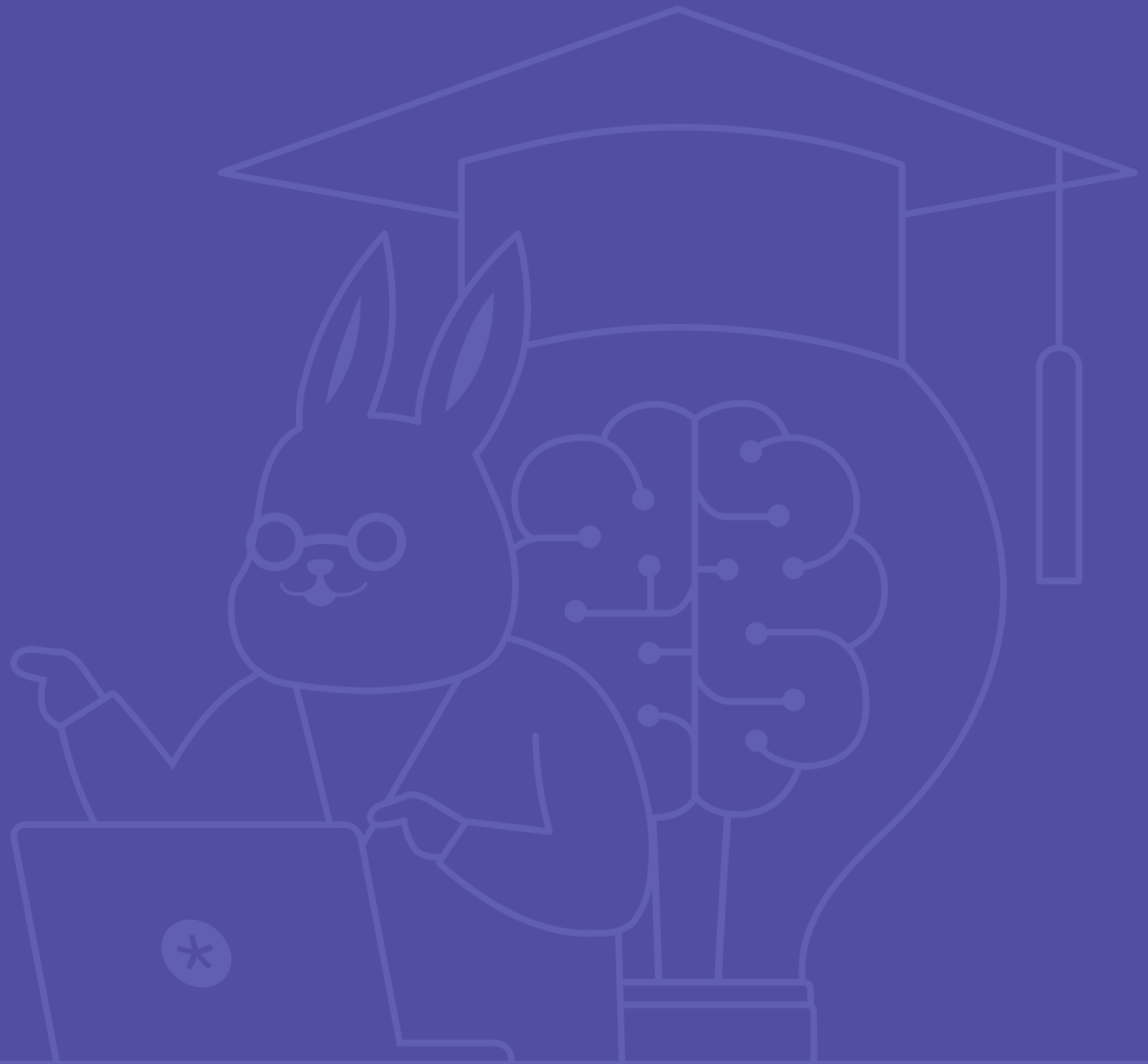
Classification

2장 분류

고혜선 선생님



Contents

Curriculum

What is Classification?

분류 (Classification)의 의미와 특성을 이해하고,
회귀 (Regression)과의 차이점을 알아본다.

Logistic / Softmax Regression

Logistic Regression과 Softmax Regression의 의미와
특성을 이해한다.

Evaluation Metrics

혼동행렬 (Confusion Matrix)를 이해하고, 분류에
사용되는 여러 지표에 대해 알아본다.

Curriculum

K-Nearest Neighbor (KNN)

kNN의 원리를 이해하고, kNN에 사용되는 distance function에 대해 알아본다.

Naïve Bayes Classifier

NBC의 기본이 되는 Bayes Theorem과 conditional independence를 이해하고, 계산방식에 대해 알아본다.

Support Vector Machines

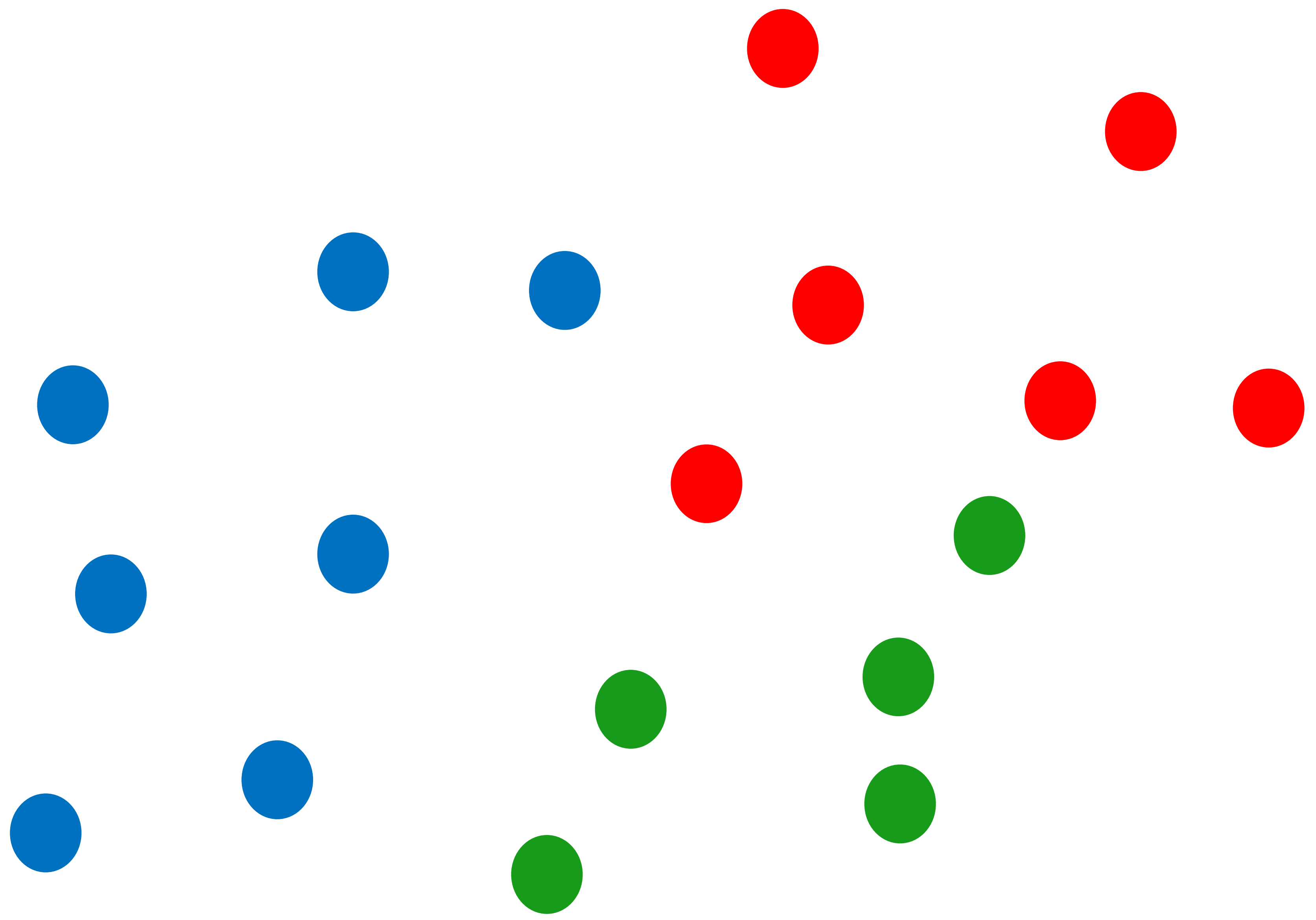
SVM의 원리를 직관적으로 이해하고, 이를 Lagrangian multiplier를 사용한 수식을 통해 계산할 수 있다.

04

K-Nearest Neighbor (KNN)

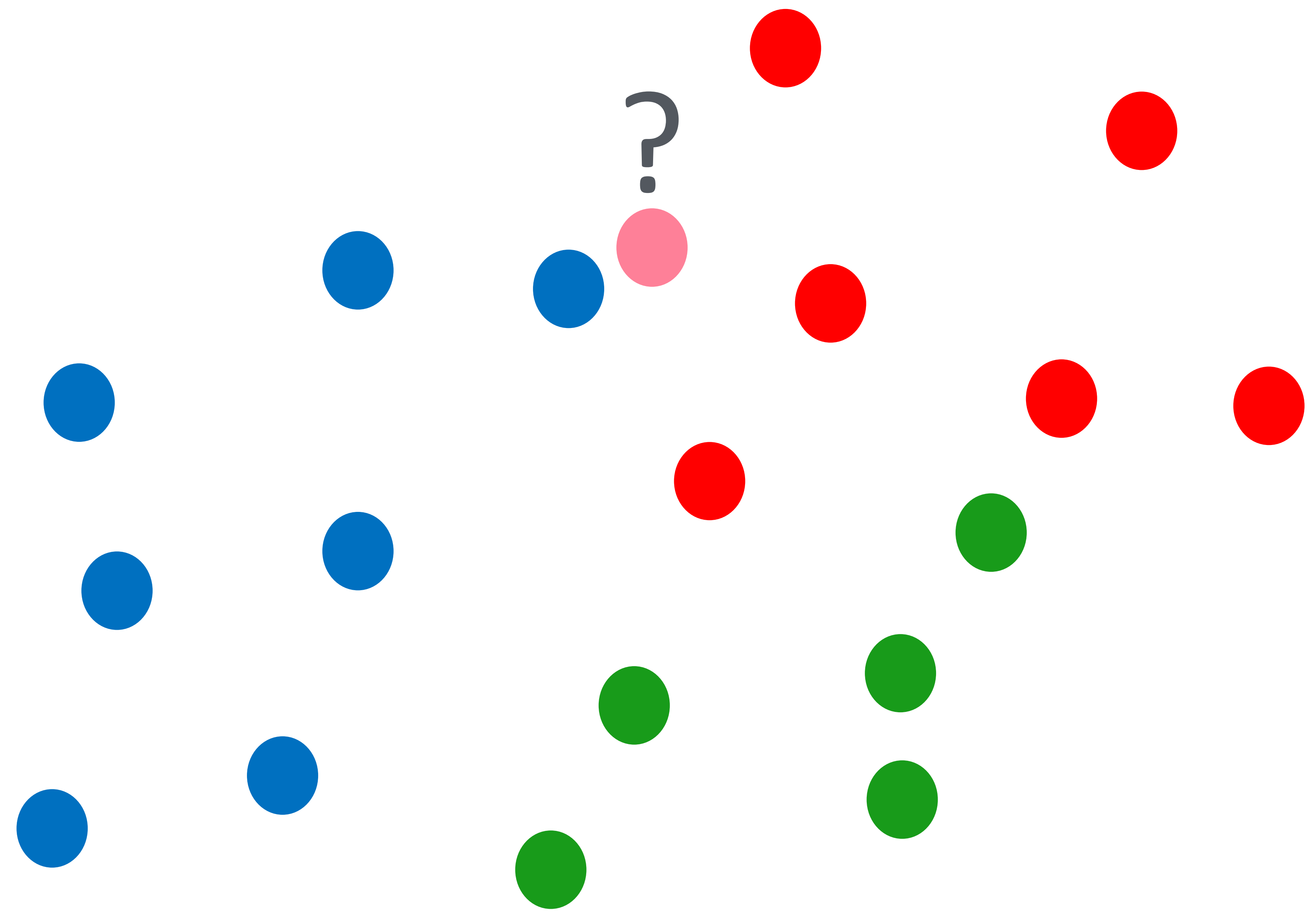
K-NN (최근접 이웃)이란?

클래스 = {RED, BLEU, GREEN}



K-NN (최근접 이웃)이란?

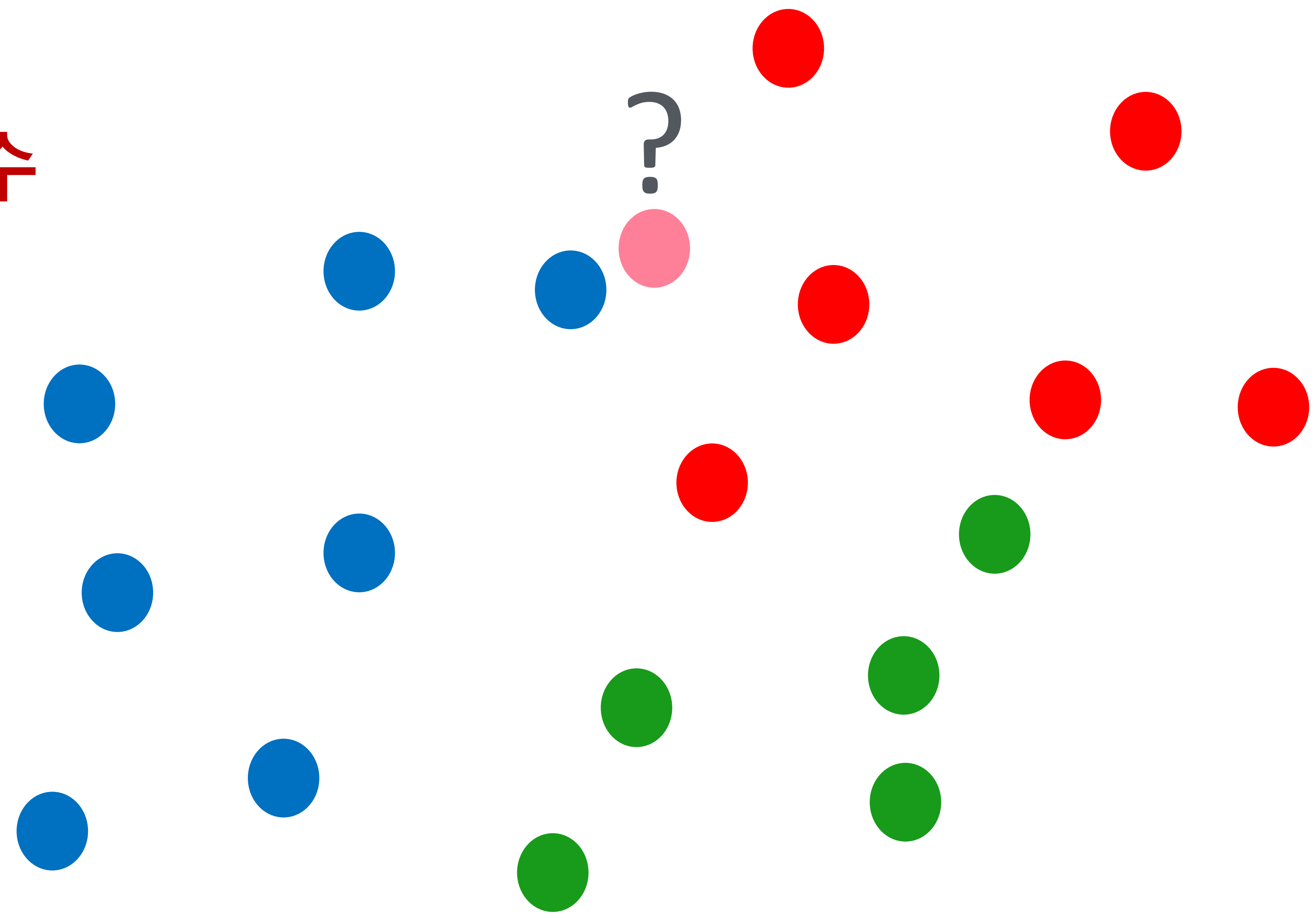
새로운 데이터 ( )는 어떤 class일까?



K-NN (최근접 이웃)이란?

새로운 데이터 ( )는 어떤 class일까?

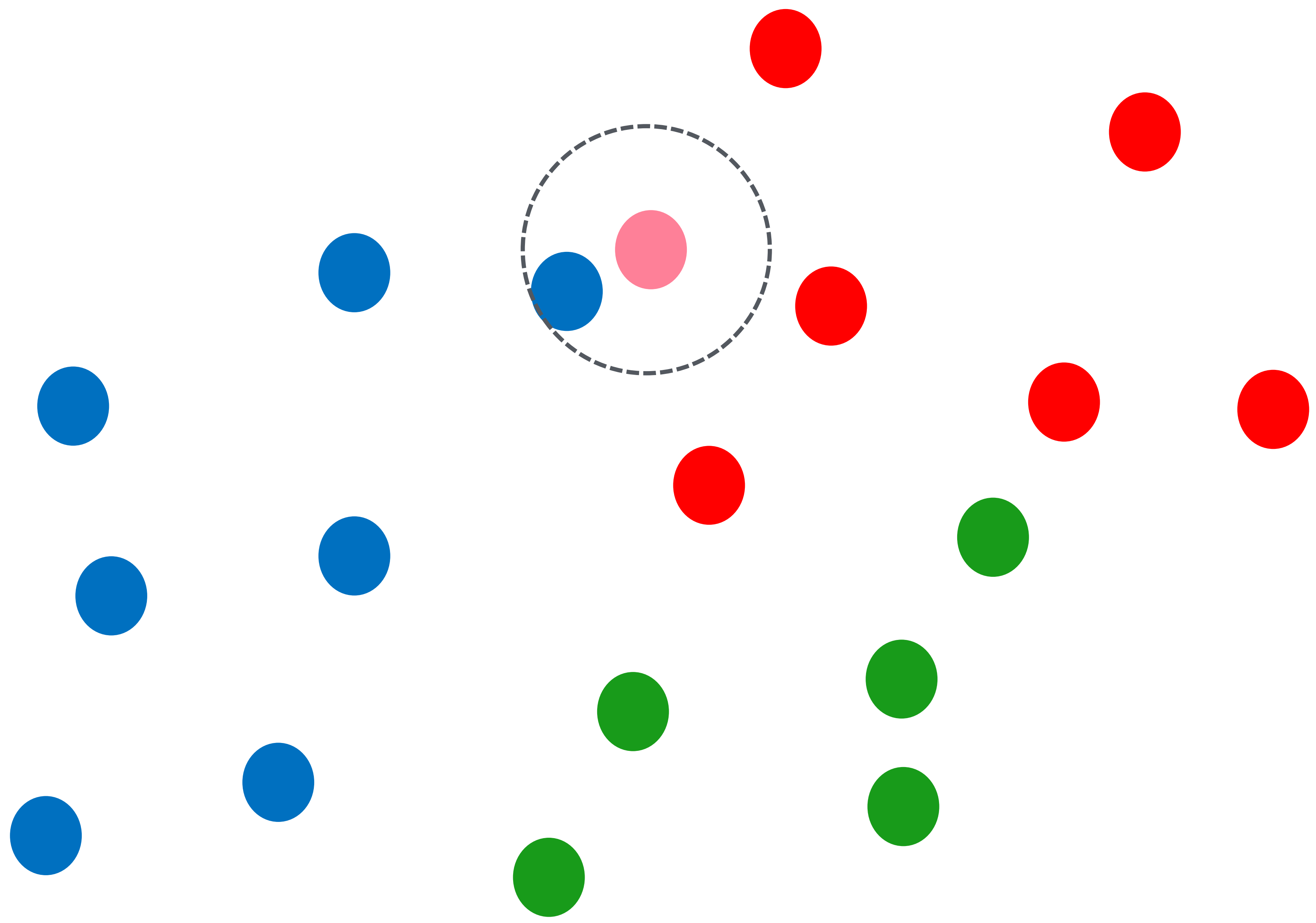
K는 주변의 개수



K-NN (최근접 이웃)이란?

새로운 데이터 ( )는 어떤 class일까? **BLEU**

K=1

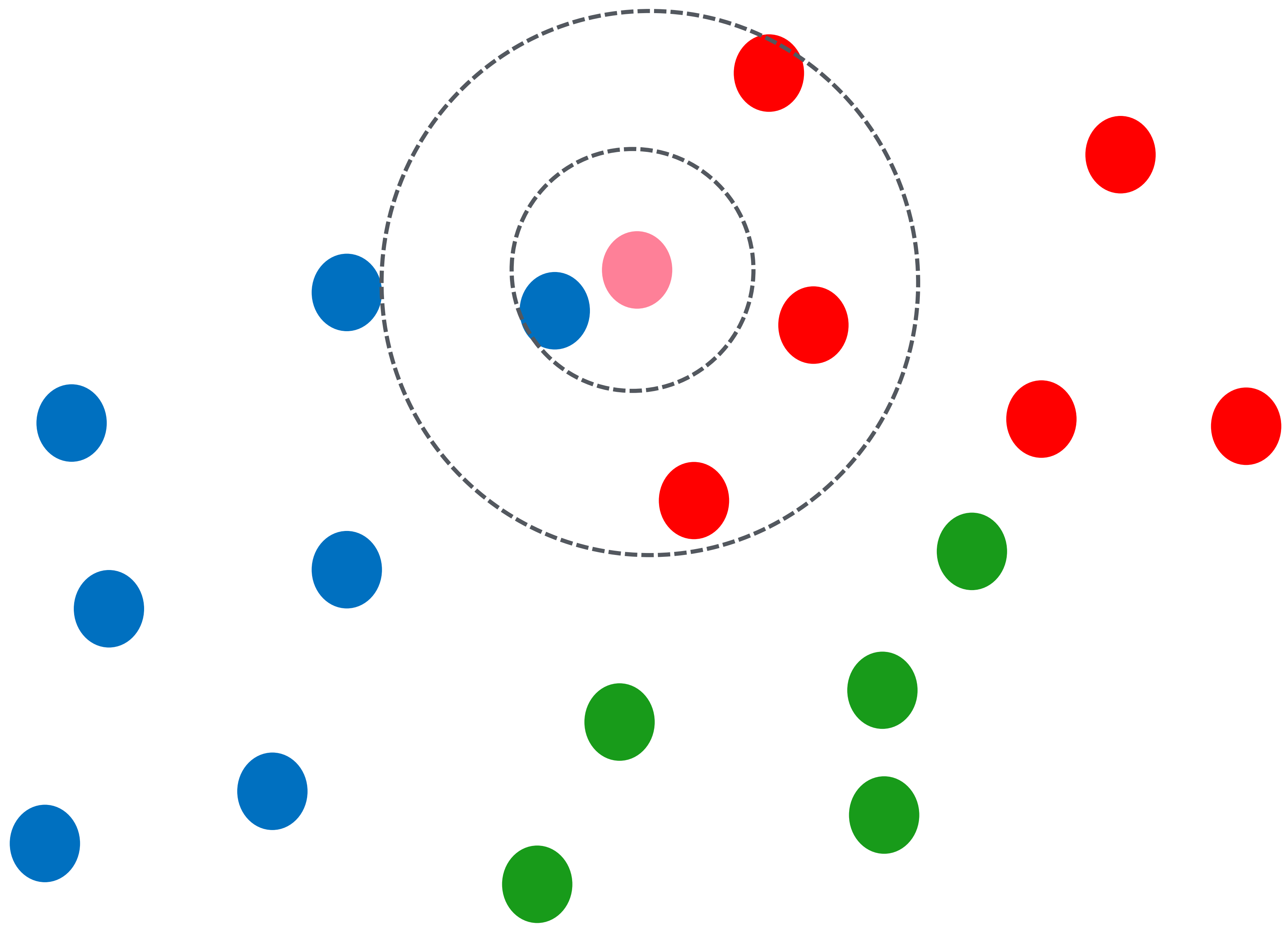


기준 = Euclidean distance

K-NN (최근접 이웃)이란?

새로운 데이터 ( )는 어떤 class일까? **RED**

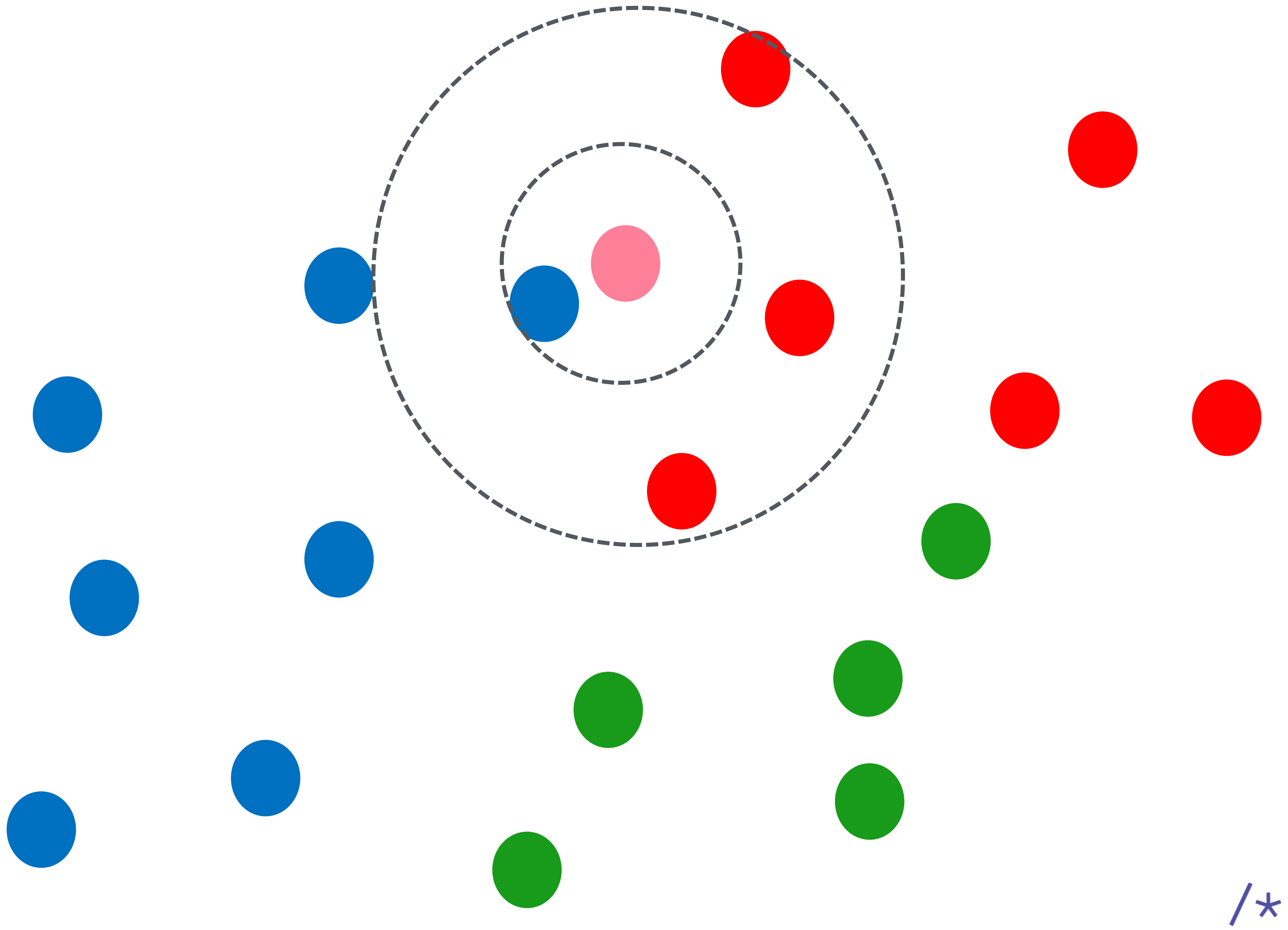
K=4



기준 = Euclidean distance

K-NN (최근접 이웃)이란?

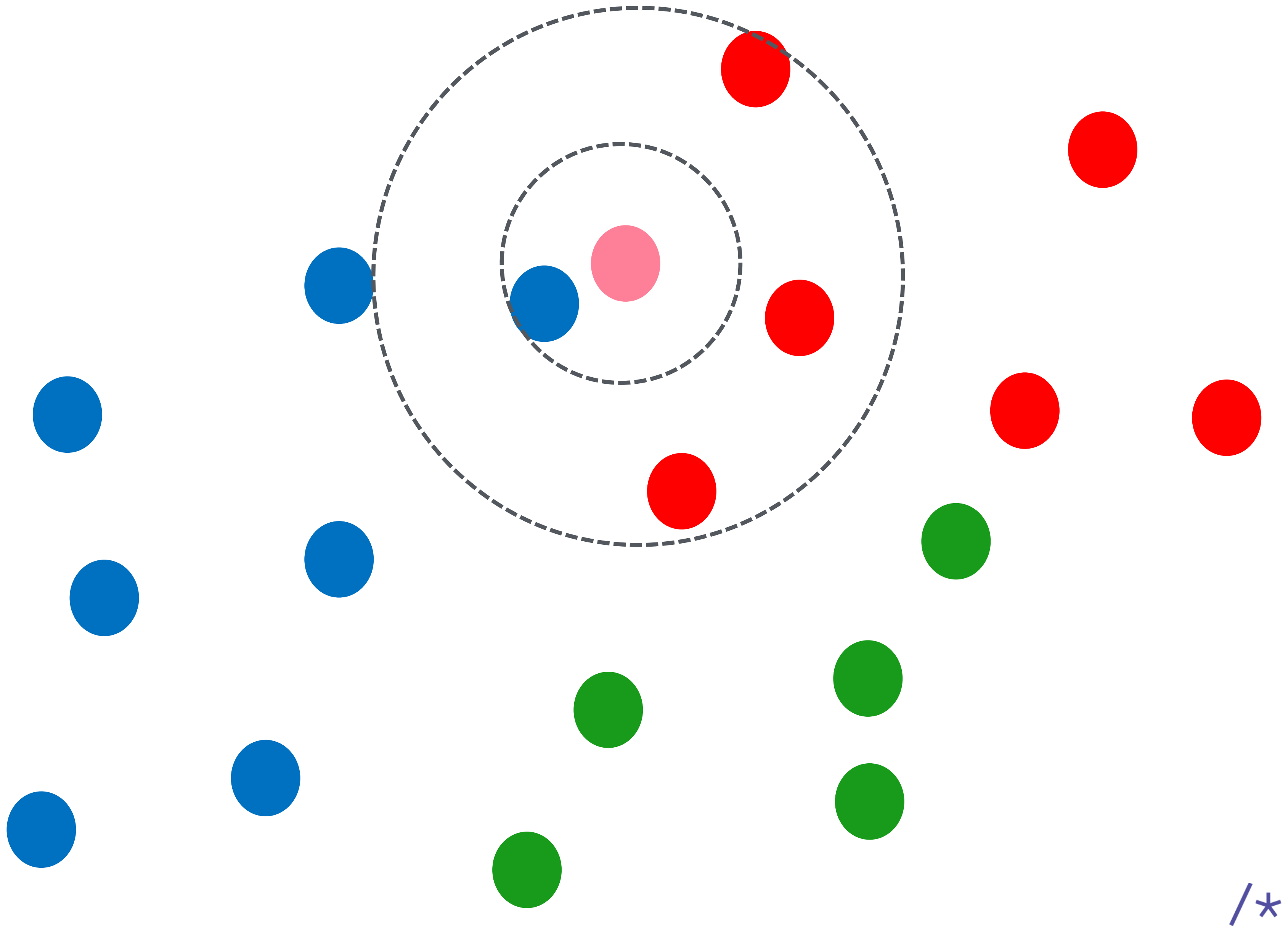
- 기준이 되는 distance function을 정한 후, 근접한 k개의 데이터를 확인한다.
- 다수결로 새로운 데이터에 대한 클래스를 결정한다.



K-NN (최근접 이웃)이란?

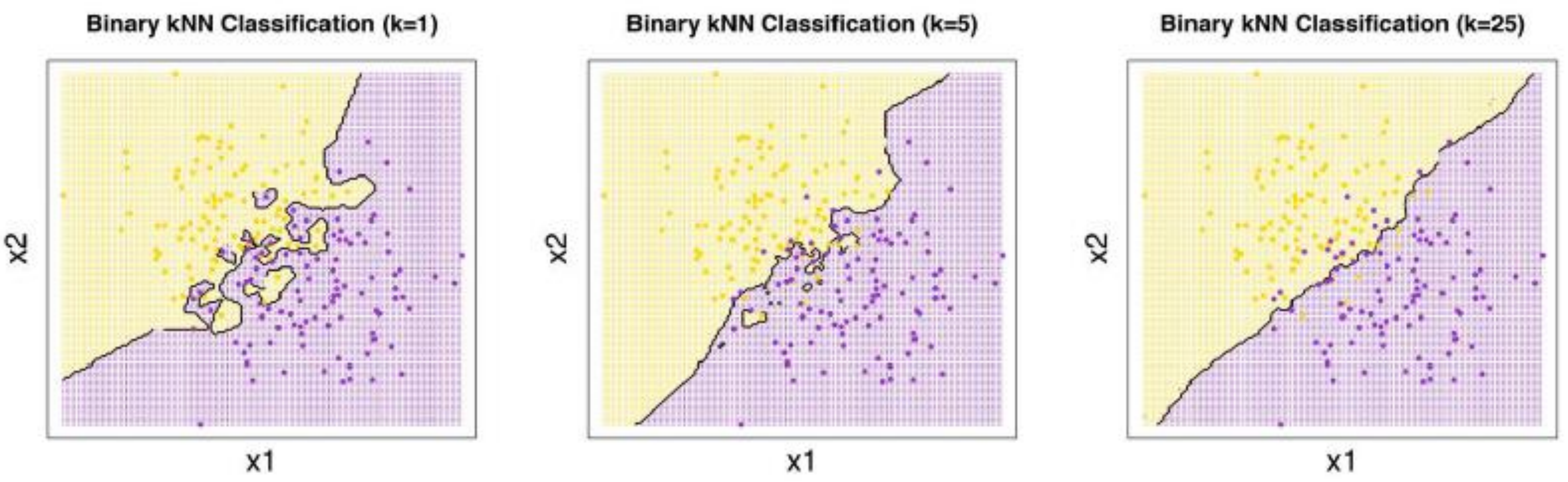
1 2

- 기준이 되는 distance function을 정한 후, 근접한 k개의 데이터를 확인한다.
- 다수결로 새로운 데이터에 대한 클래스를 결정한다.



K가 점점 커진다면..

- K=5일 때가, k=1일때보다 더 smoother boundary 를 그리고, label noise를 줄여준다.
- 하지만, k가 너무 클 때 (예: k=데이터개수), 항상 **majority class**로 분류하게 된다.



*/

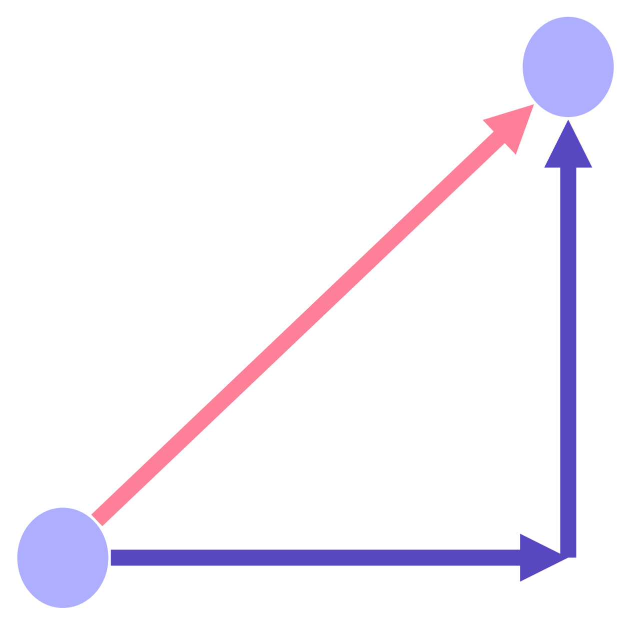
Distance function

- Manhatton distance (L1-norm)

$$dist(x, y) = \sum_{i=1}^{N} |x_i - y_i|$$

- Euclidean distance (L2-norm)

$$dist(x, y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$$



$$dist(x, y) = (\sum_{i=1}^{N} |x_i - y_i|^p)^{1/p}$$

Generalized version (p-norm)

이외에도 distance function은 많습니다. (e.g., Mahalanobis distance, cosine similarity, Pearson correlation, Jaccard similarity)

How: Not just counting

$$S(x', \boldsymbol{RED}) = \sum_{x \in N(x', RED)}^{N} w(x)$$

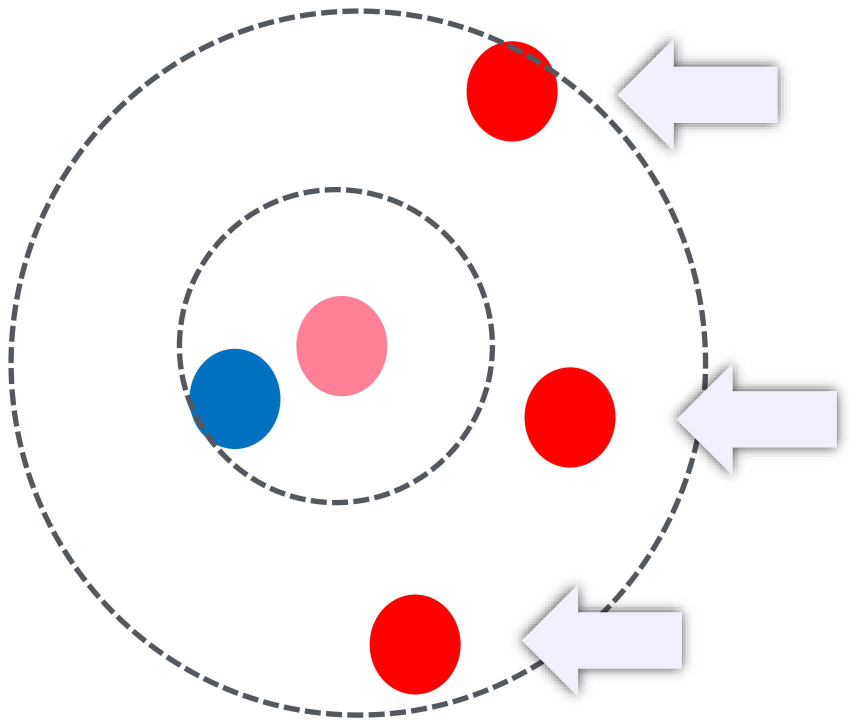
$$S(x', \boldsymbol{BLUE}) = \sum_{x \in N(x', BLUE)}^{N} w(x)$$

How: Not just counting

$$S(x', \textbf{\textit{RED}}) = \sum_{x \in N(x', RED)}^{N} w(x)$$

$$S(x', \textbf{\textit{BLUE}}) = \sum_{x \in N(x', BLUE)}^{N} w(x)$$

The set of **RED** data among the nearest neighbors of x’



How: Not just counting

$$S(x', \mathbf{RED}) = \sum_{x \in N(x', RED)}^{N} w(x) \quad > \quad S(x', \mathbf{BLUE}) = \sum_{x \in N(x', BLUE)}^{N} w(x)$$

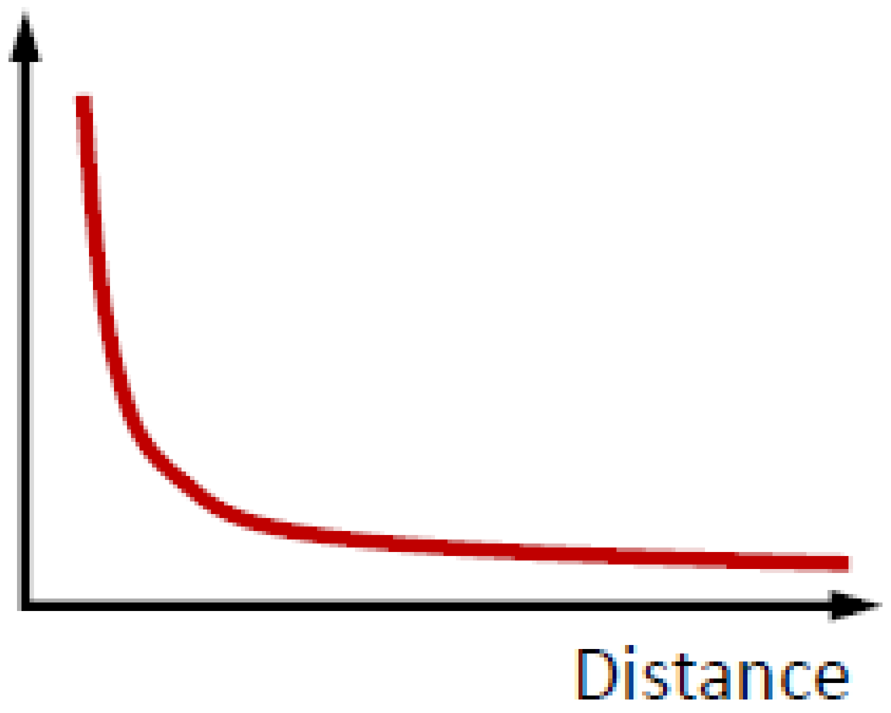
Predicted Class = "RED"

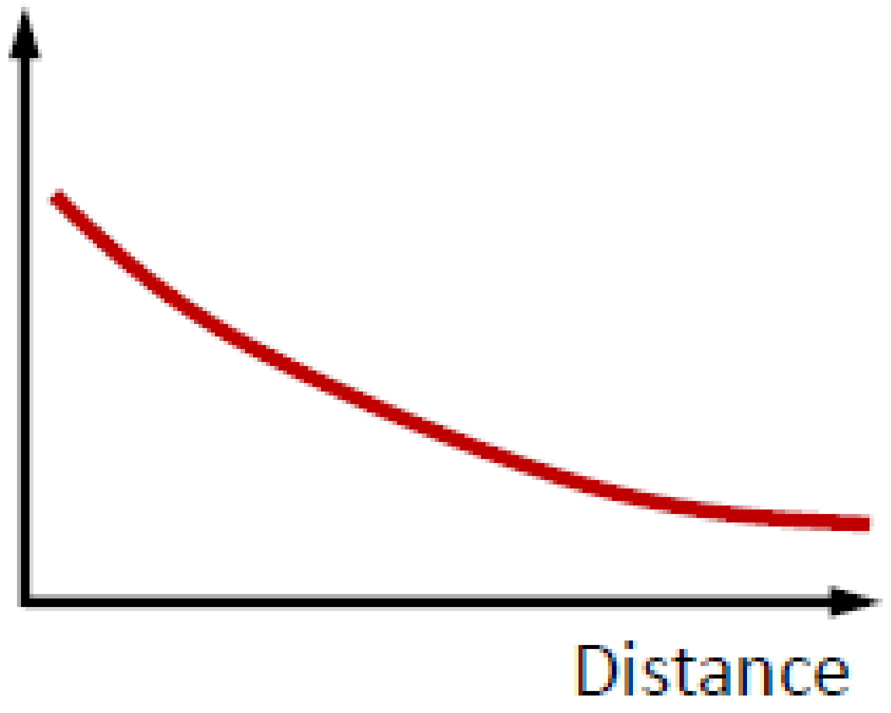
How: Not just counting

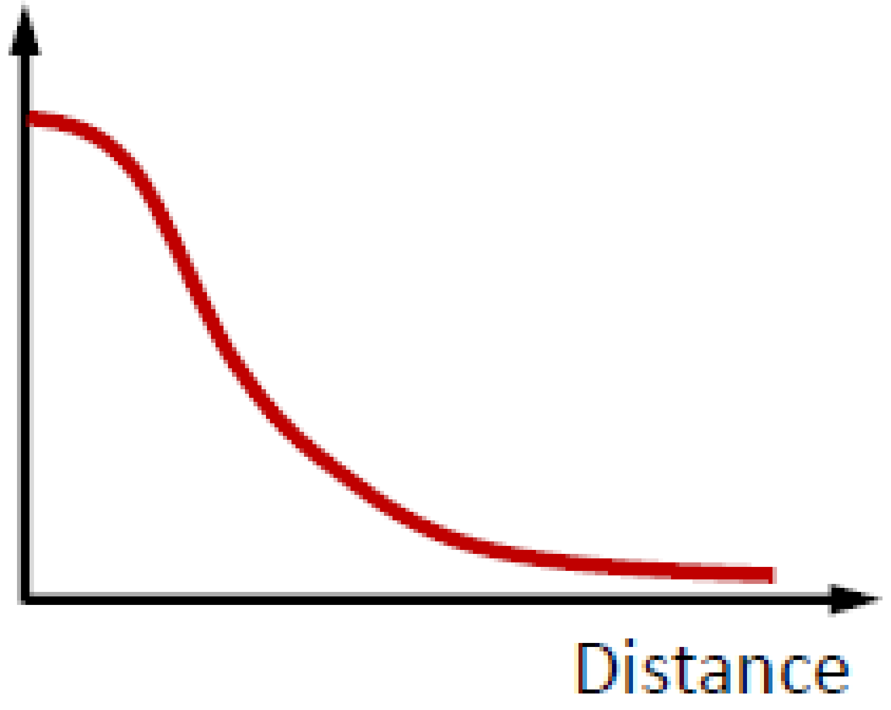
$$S(x', \mathbf{\mathit{RED}}) = \sum_{x \in N(x', RED)}^{N} w(x) < S(x', \mathbf{\mathit{BLUE}}) = \sum_{x \in N(x', BLUE)}^{N} w(x)$$

Predicted Class = “BLUE”

How: Determine weights







$$w(\mathbf{x}) = \frac{1}{dist(\mathrm{x}, \mathrm{x}')}$$

$$w(\mathbf{x}) = \exp(-dist(\mathrm{x}, \mathrm{x}'))$$

$$w(\mathbf{x}) = \exp\left(-\frac{dist(\mathrm{x}, \mathrm{x}')^2}{2\lambda^2}\right)$$

Summary

- (+) No training step (only inference step)
- (-) Need to calculate the distance from all training data → Time
- (-) Sensitive to noise
- (-) If training data is imbalanced, major class may dominate

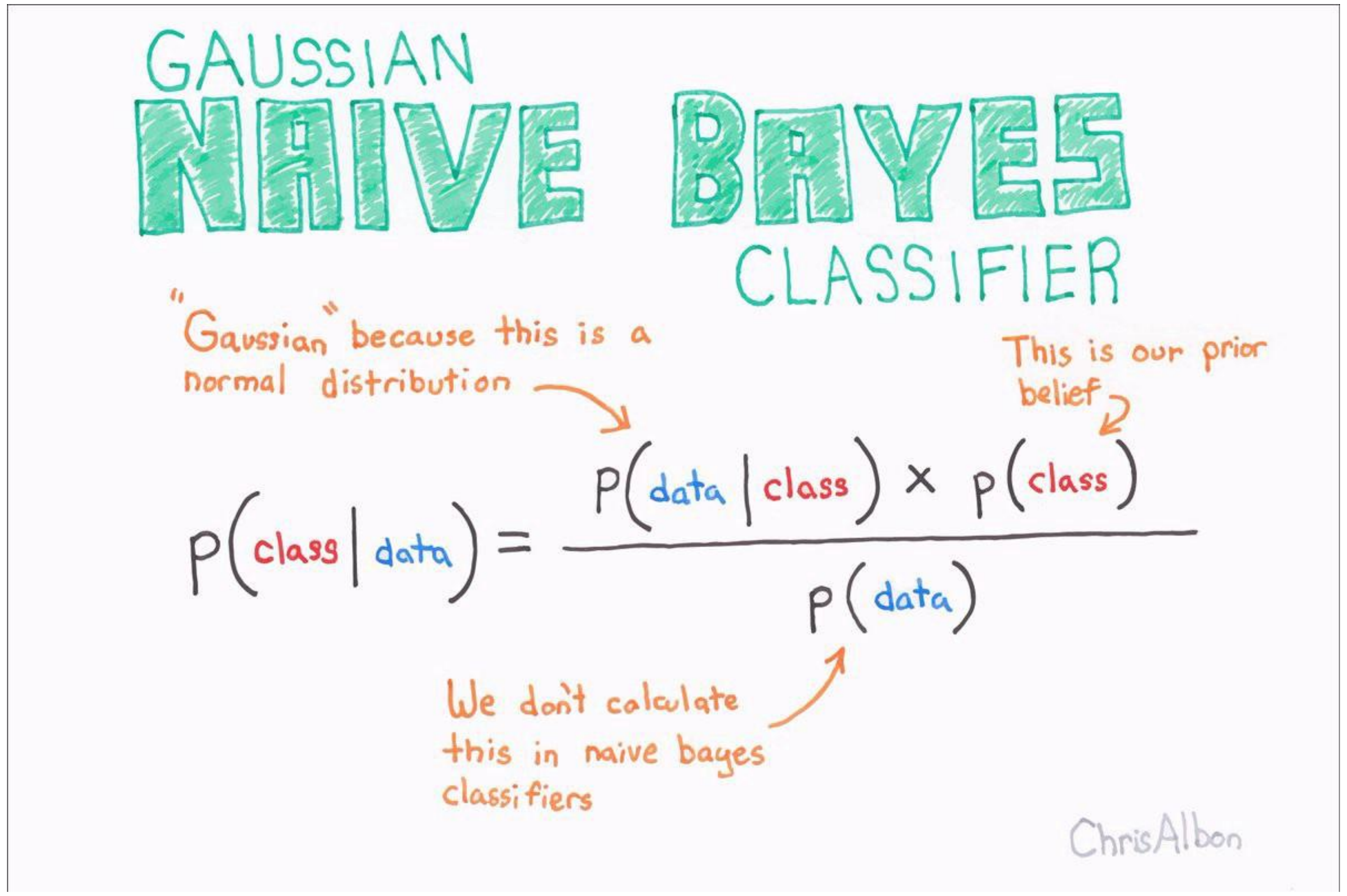
- **Which k is better?**
- > Small k: Higher variance (overfitting)
- > Large k: Higher bias (underfitting)

05

Naïve Bayes Classifier

Naïve Bayes Classifier란

- Bayes 이론에 기반하여, 어떤 데이터가 주어졌을 때 Conditional Independence를 가정하여 “특정“ 클래스에 속할 확률을 계산하여 분류하는 모델



Review: Bayes' Theorem

$$P(\text{class} \mid X) = \frac{P(X \mid \text{class})\, P(c)}{P(X)}$$

Review: Bayes' Theorem

Likelihood

Prior

$$\boxed{P(\text{class} \mid X)} = \frac{\boxed{P(X \mid \text{class})}\,\boxed{P(c)}}{P(X)}$$

Posterior

Bayesian vs. Frequentist

Probability vs. Likelihood

- Probability
  - 모델(모수; θ)을 기반으로 데이터가 나타날 확률
  - 예: P(X>=30 | mean=20, std=1)

$$Pr(\text{data} \mid \text{fixed distribution}; \theta)$$

- Likelihood
  - 데이터(관측치)를 기반으로 분포의 모수를 추정하는 것
  - 예: P(mean=20, std=1| X=30)

$$Pr(\theta \mid \text{data})$$

Probability vs. Likelihood

- Probability
  - 모델(모수; θ)을 기반으로 데이터가 나타날 확률
  - 예: P(X>=30 | mean=20, std=1)

- Likelihood
  - 데이터(관측치)를 기반으로 분포의 모수를 추정하는 것
  - 예: P(mean=20, std=1| X=30)

Conditional Independence

- Two events are **conditionally independent** given an event C with $P(C)>0$ if

$$P(A, B \mid C) = P(A \mid C)\, P(B \mid C)$$

Conditional Independence

- Two events are **conditionally independent** given an event C with P(C)>0 if

$$P(A, B \mid C) = P(A \mid C)\, P(B \mid C)$$

$$\mathrm{P}(\mathrm{class} \mid x_1, x_2, \dots, x_d) = \frac{\mathrm{P}(x_1, x_2, \dots, x_d \mid \mathrm{class})\, \mathrm{P}(\mathrm{c})}{\mathrm{P}(\mathrm{X})}$$

X를 여러 개의 feature로 표현

Conditional Independence

- Two events are **conditionally independent** given an event C with P(C)>0 if

$$P(A, B \mid C) = P(A \mid C)\, P(B \mid C)$$

각 특성 별 확률 곱으로 계산 될 수 있음

$$P(x_1 \mid \text{class}) P(x_2 \mid \text{class}) \ldots P(x_d \mid \text{class})$$

$$P(\text{class} \mid x_1, x_2, \ldots, x_d) = \frac{P(x_1, x_2, \ldots, x_d \mid \text{class})\; P(c)}{P(X)}$$

x를 여러 개의 feature로 표현

Classification example: Good vs. Bad

	sex	*mask*	*cape*	*tie*	*ears*	*smokes*	*Label*
batman	male	yes	yes	no	yes	no	Good
robin	male	yes	yes	no	no	no	Good
alfred	male	no	no	yes	no	no	Good
penguin	male	no	no	yes	no	yes	Bad
catwoman	female	yes	no	no	yes	no	Bad
joker	male	no	no	no	no	no	Bad

Classification example: Good vs. Bad

??

	sex	*mask*	*cape*	*tie*	*ears*	*smokes*	*Label*
batman	male	yes	yes	no	yes	no	Good
robin	male	yes	yes	no	no	no	Good
alfred	male	no	no	yes	no	no	Good
penguin	male	no	no	yes	no	yes	Bad
catwoman	female	yes	no	no	yes	no	Bad
joker	male	no	no	no	no	no	Bad
superman	male	yes	yes	no	no	no	??

Classification example: Good vs. Bad

superman	male	yes	yes	no	no	no

- P(class=Good | X)와 P(class=Bad | X)를 비교해서 더 큰 값을 가지는 클래스로 분류!

Classification example: Good vs. Bad

superman	male	yes	yes	no	no	no

Prior probability

- $P(Good) = 0.5, \ P(Bad) = 0.5$

Conditional probability

- $P(sex = male \mid Good) = 1.0, \quad P(sex = male \mid Bad) = 0.67$
- $P(mask = yes \mid Good) = 0.67, \quad P(mask = yes \mid Bad) = 0.33$
- $P(cape = yes \mid Good) = 0.67, \quad P(cape = yes \mid Bad) = 1.0$
- $P(tie = no \mid Good) = 0.67, \quad P(tie = no \mid Bad) = 0.67$
- $P(ears = no \mid Good) = 0.67, \quad P(ears = no \mid Bad) = 0.67$
- $P(smokes = no \mid Good) = 0.67, \quad P(smokes = no \mid Bad) = 0.67$

Classification example: Spam vs. Not-Spam

IDX	type	text
1	ham	Hope you are having a good week. Just checking in
2	ham	K..give back my thanks.
3	ham	Am also doing in cbe only. But have to pay.
4	spam	complimentary 4 STAR Ibiza Holiday or 10,000 cash needs your URGENT
5	spam	okmail: Dear Dave this is your final notice to collect your 4* Tenerife Holiday or ······
~		
5559	ham	Shall call now dear having food

idx	check	good	thanks	pay	~
1	1	1	0	0	~
2	0	0	1	1	
3	0	0	0	0	
~					

$$P(스팸|단어1, 단어2, 단어3 \ldots) > P(정상|단어1, 단어2, 단어3 \ldots) \text{ 이면 스팸}$$

Discussion: Zero Probability

$$P(x_1 | \text{class}) P(x_2 | \text{class}) \ldots P(x_d | \text{class})$$
$$= 0$$

Example: Laplacian Correction (Laplacian estimator)

- 예: 100개의 데이터 중, score = A(10), score = B(90), score = C(0) 라고 가정하자.
- 문제점: P(score=A | new X)를 계산하는 과정에서 score C의 경우, zero probability가 발생함.

- **해결책**: Laplacian Correction (Laplacian Estimator)

  (1) 각 데이터에 대해 1씩 더한다.
- **계산해보면?**

  P(score=A) = (10+1)/(100+3) = 1/103

  P(score=B) = (90+1)/103 = 91/103

  P(score=C) = (0+1)/103 = 1/103

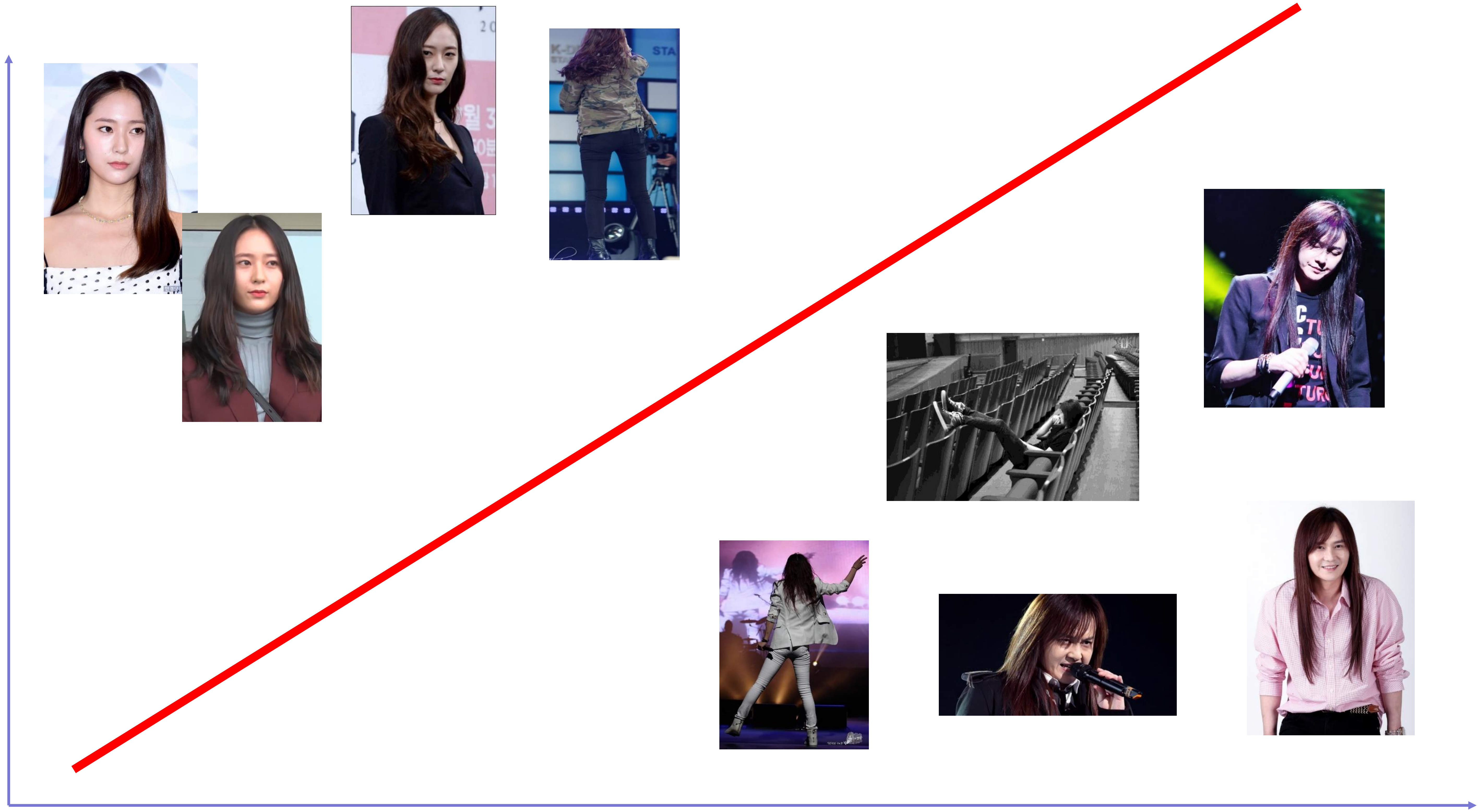
Summary

- (+) Easy to implement
- (+) Good results can be obtained when cond.independence satisfies
- (-) Practically, dependencies exist among variables. ➔ loss of Accuracy
- (-) cannot model the dependency between attributes

06

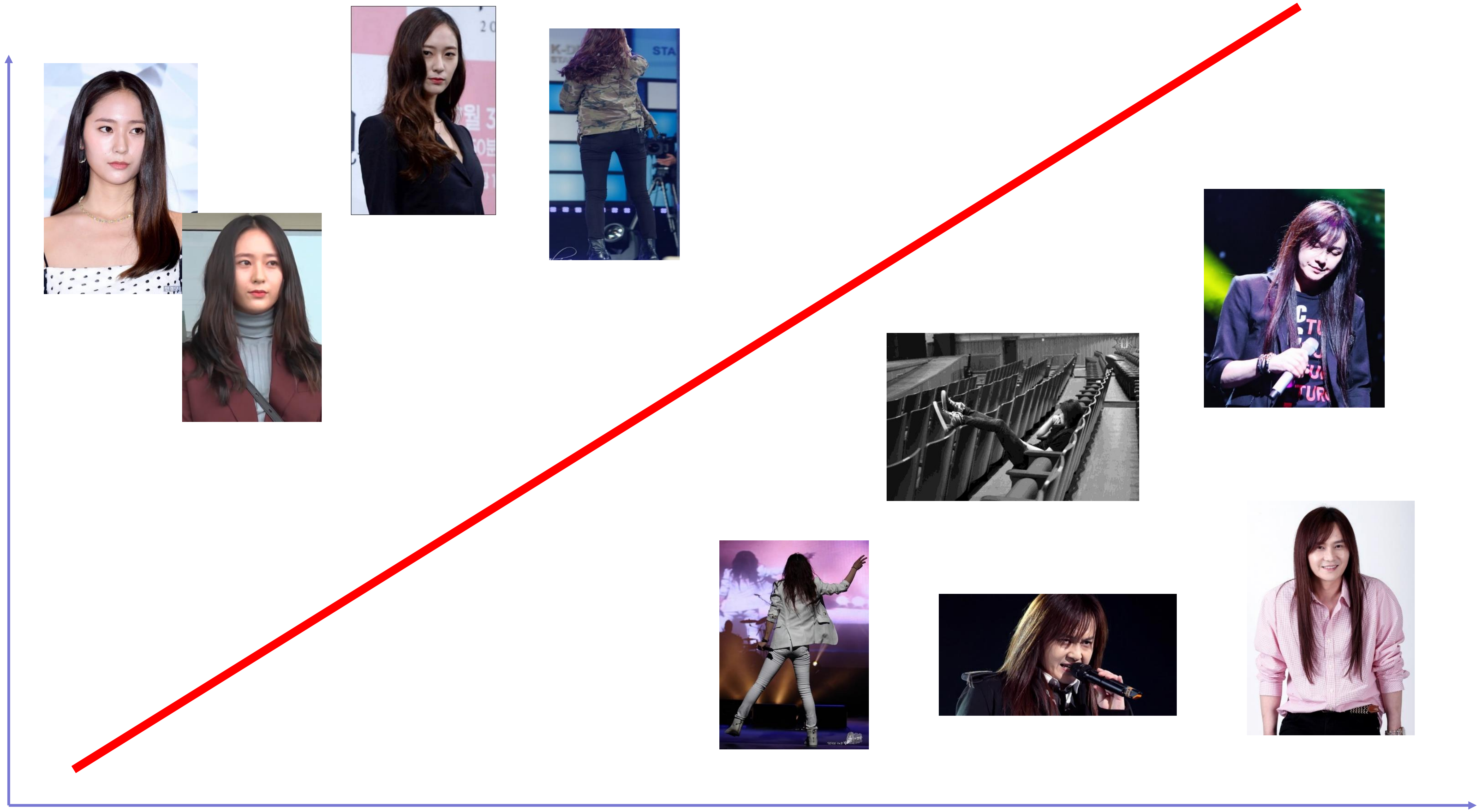
Support Vector Machines (SVM)

Classification: 김경호 vs. 크리스탈



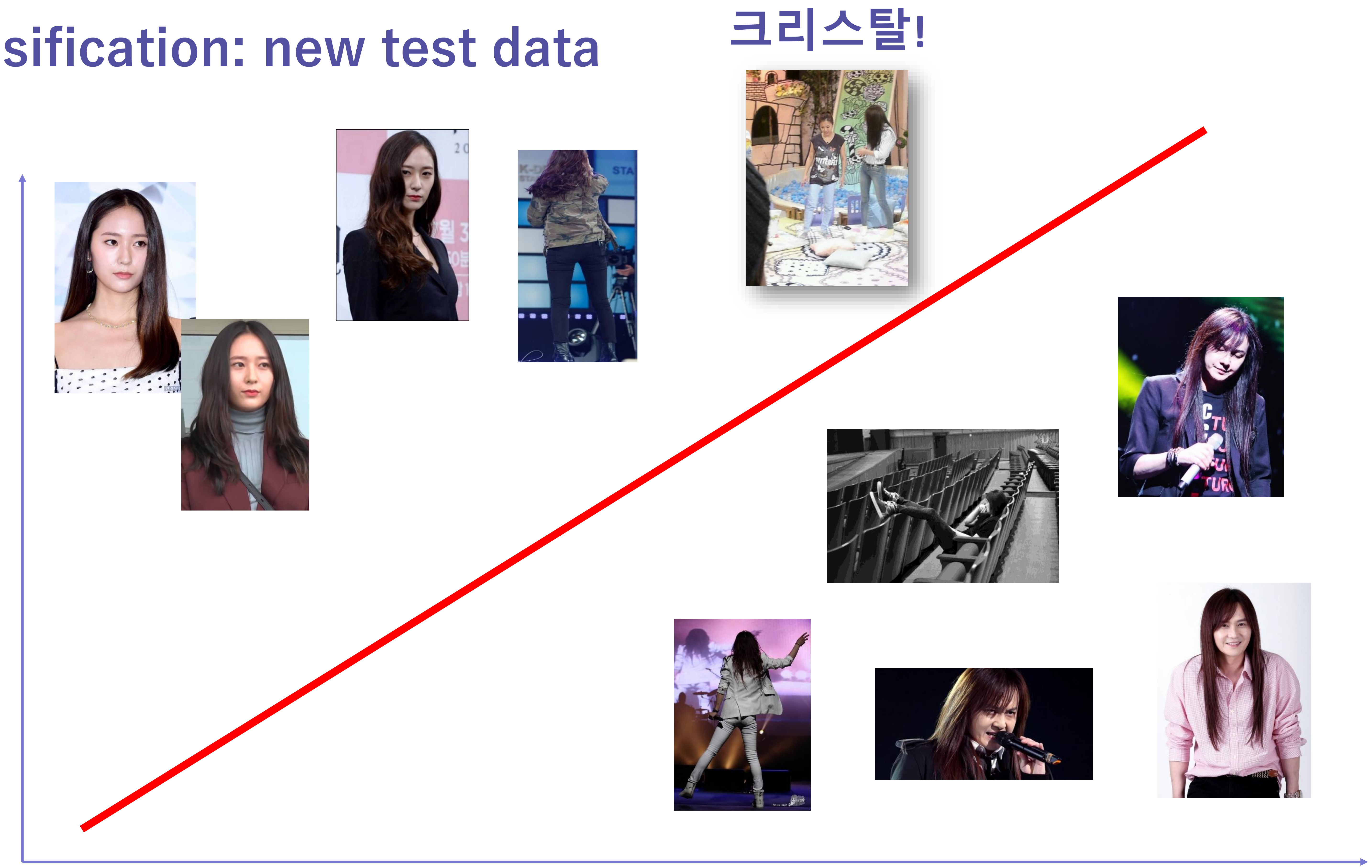
Classification: 김경호 vs. 크리스탈

Classification: new test data



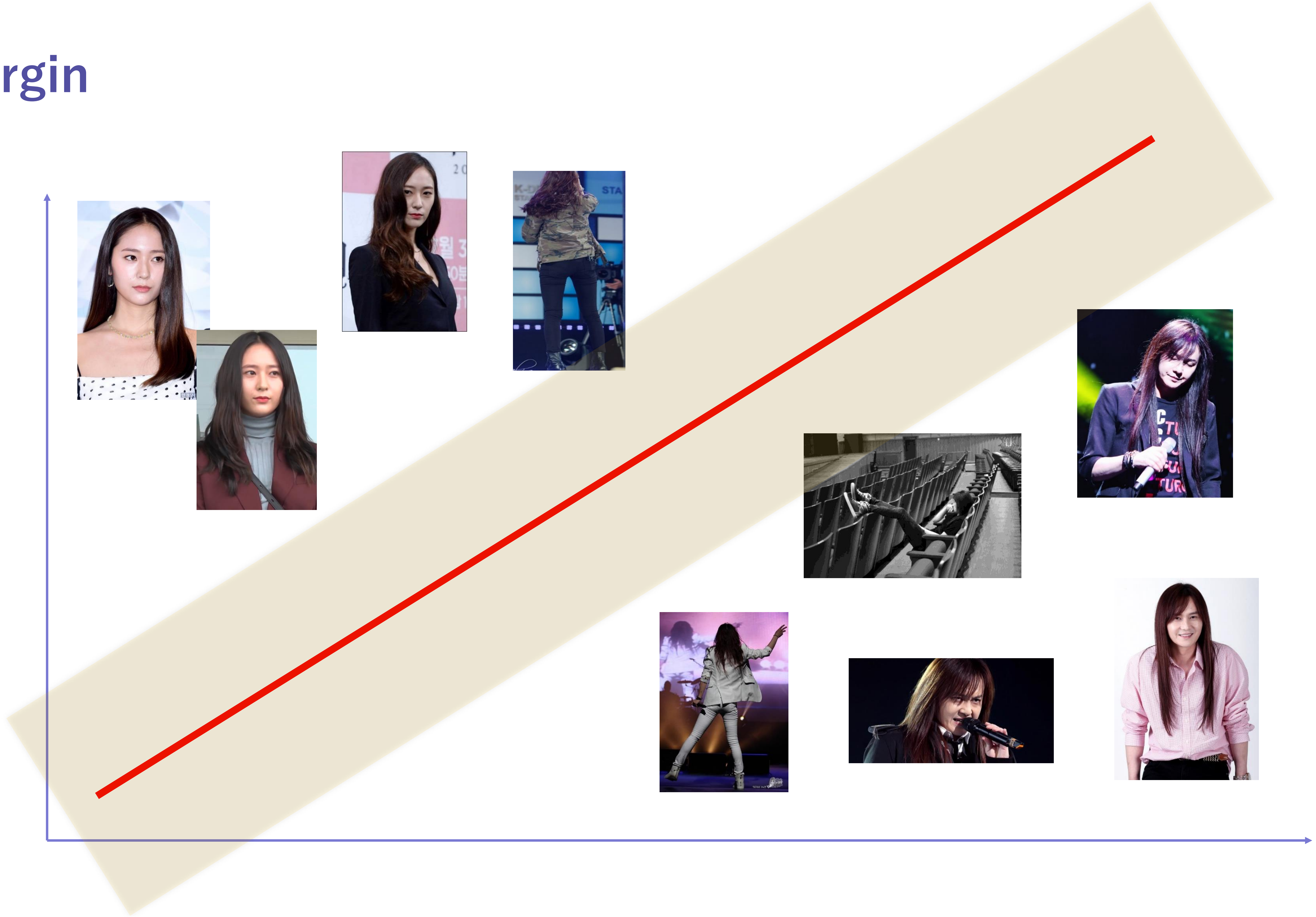
Classification: new test data

Classification: new test data

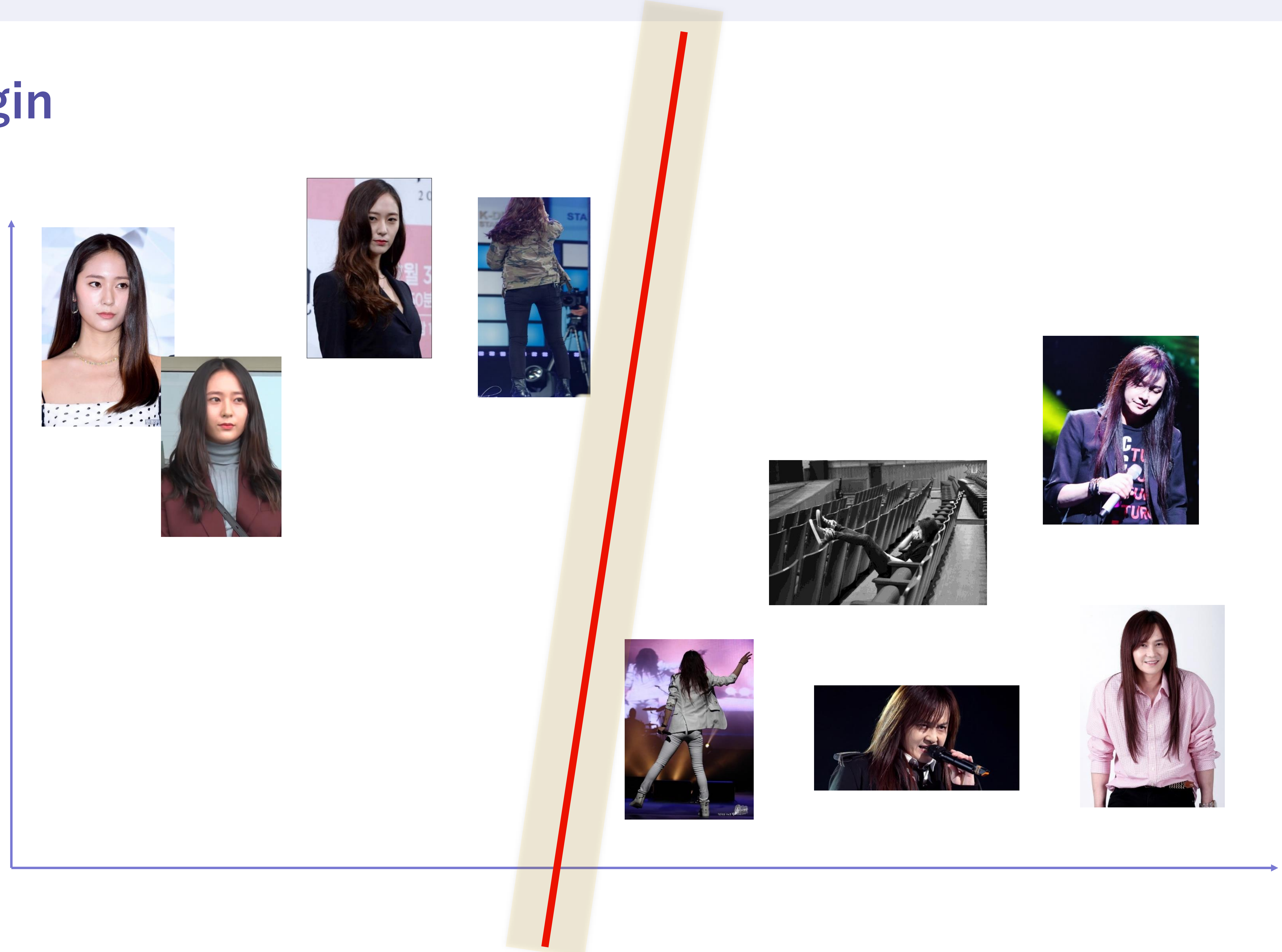


Classification: new test data

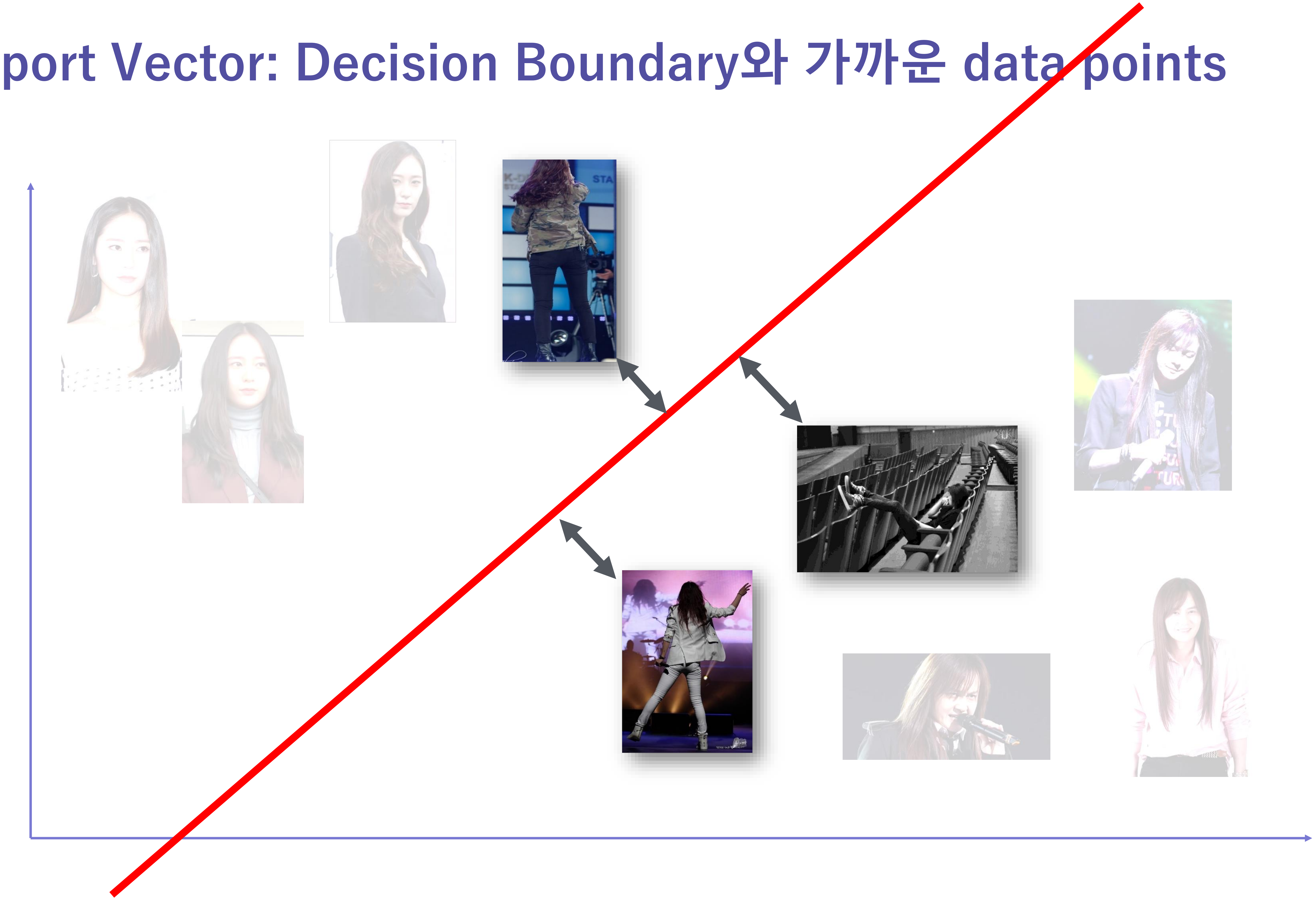
Margin



Margin



- Support Vector: Decision Boundary와 가까운 data points

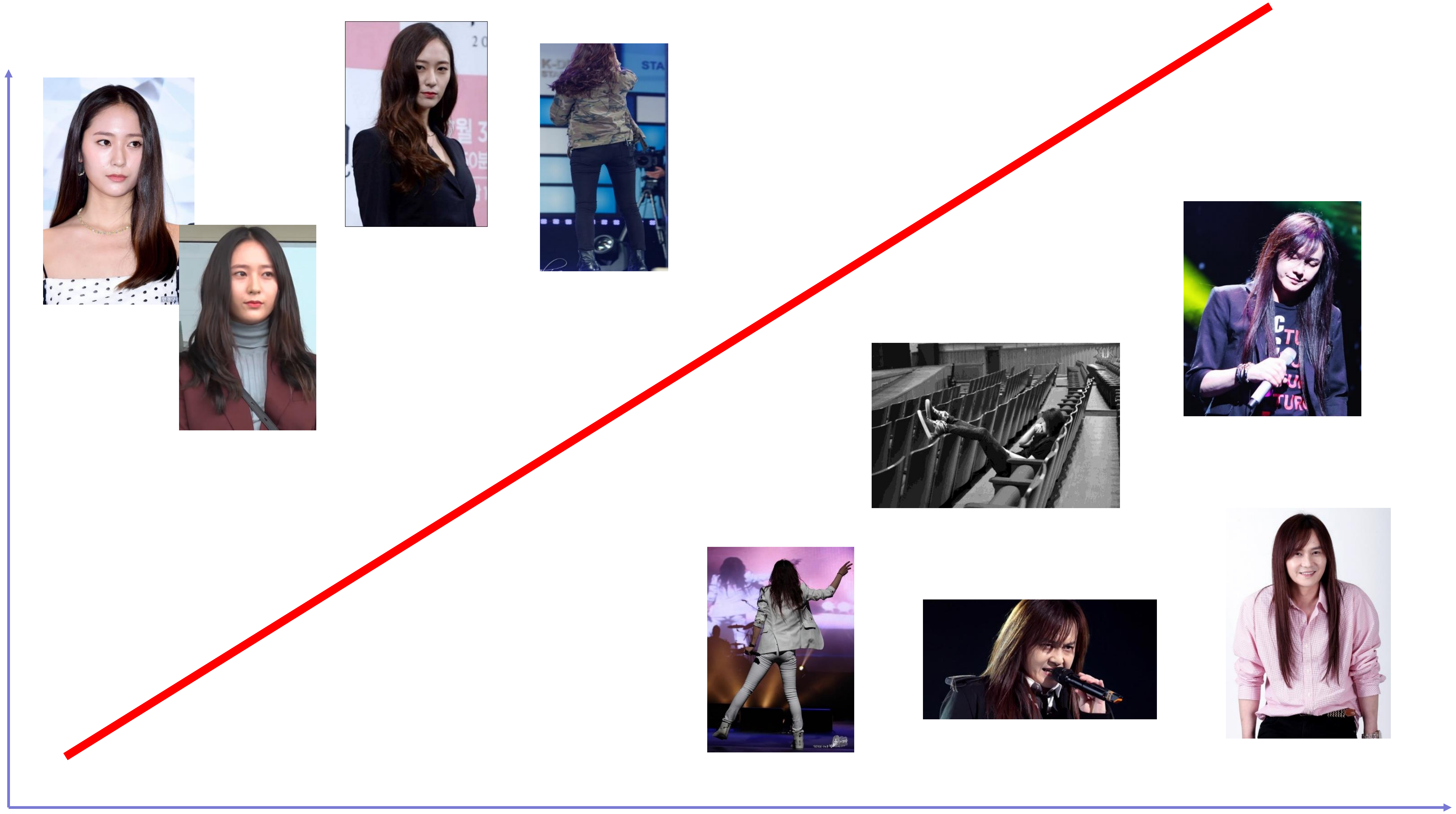


Support Vector: Decision Boundary와 가까운 data points

- Only support vectors are important; other examples are ignorable

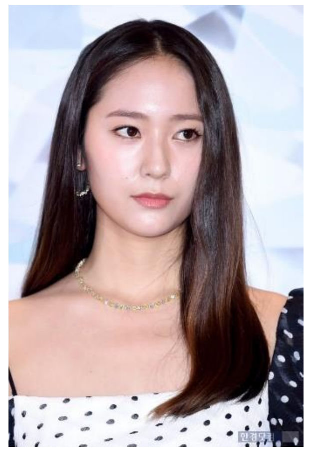
  → **Less computation**!

Linear Support Vector Machines (LSVM)



Non-Linear SVM

김



크

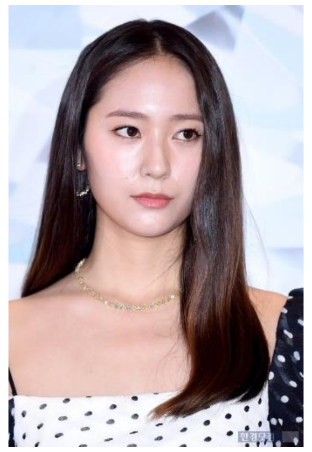
크

크

김

Non-Linear SVM

김



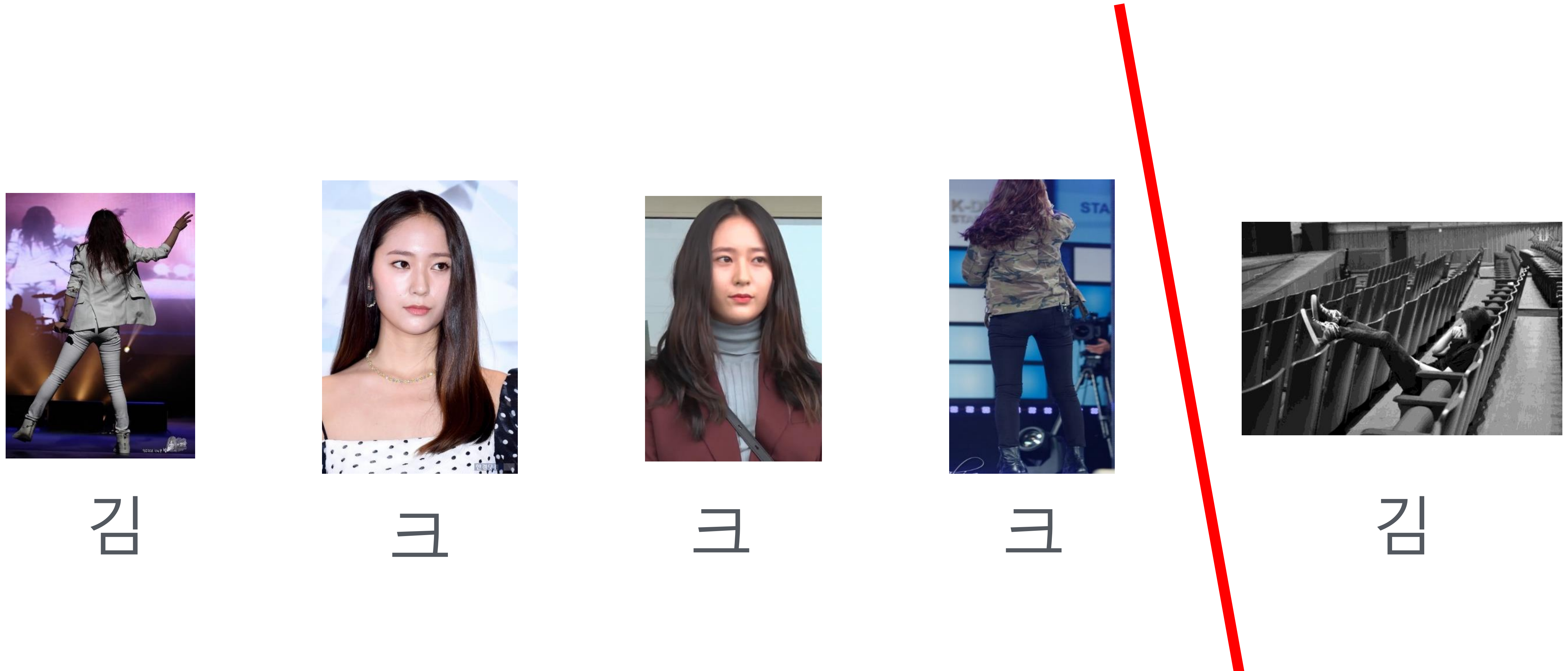
크

크

크

김

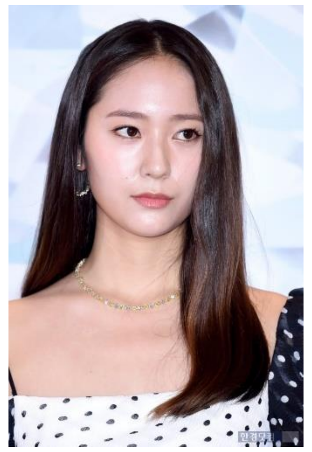
Non-Linear SVM



Non-Linear SVM: Kernel Trick $(y = x^2)$

김

김

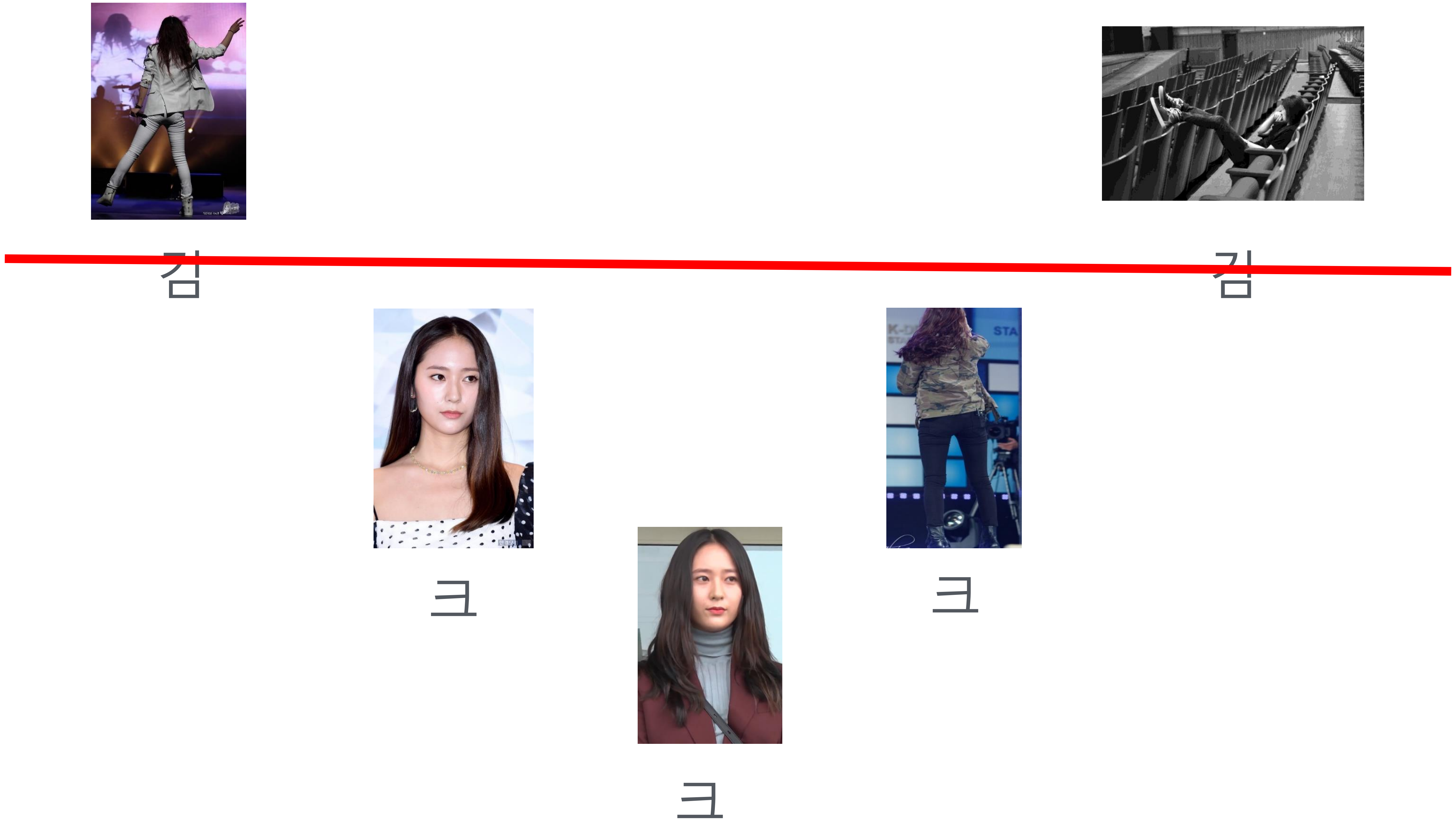


크

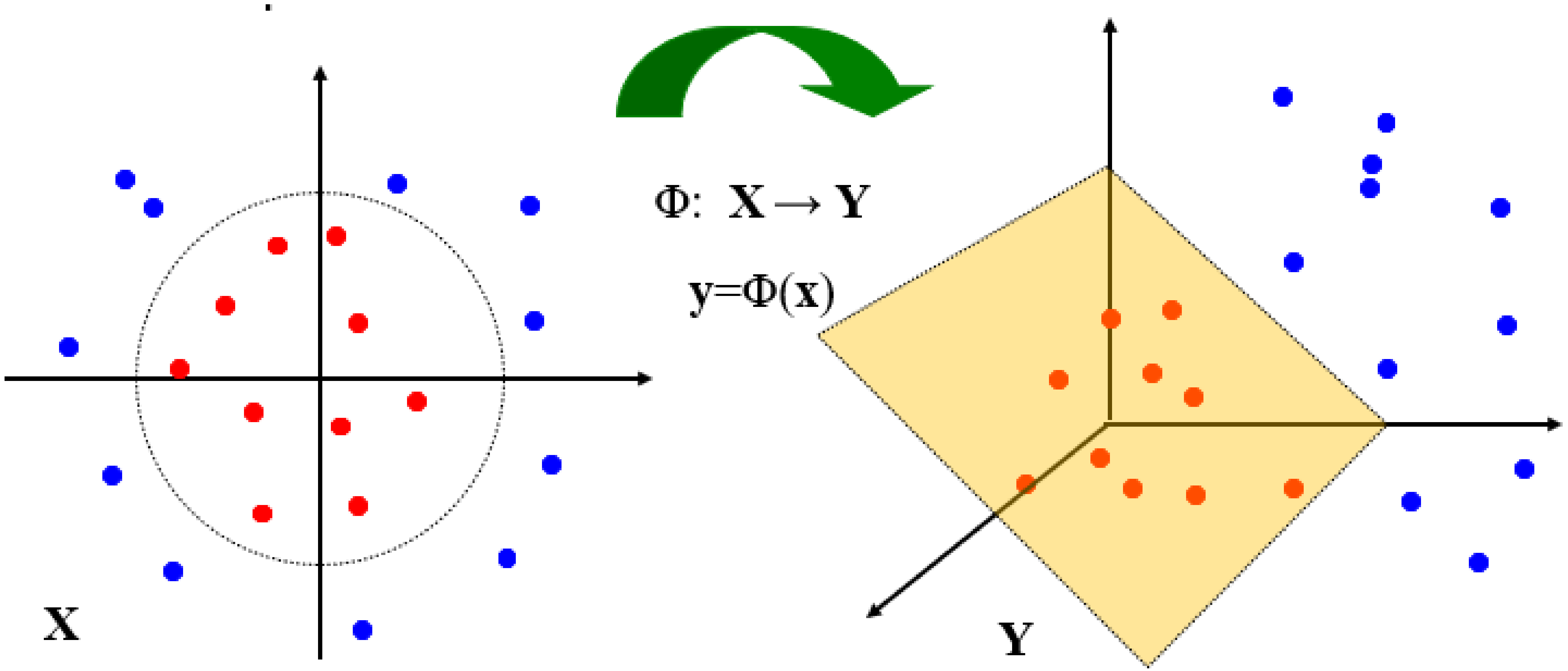
크

크

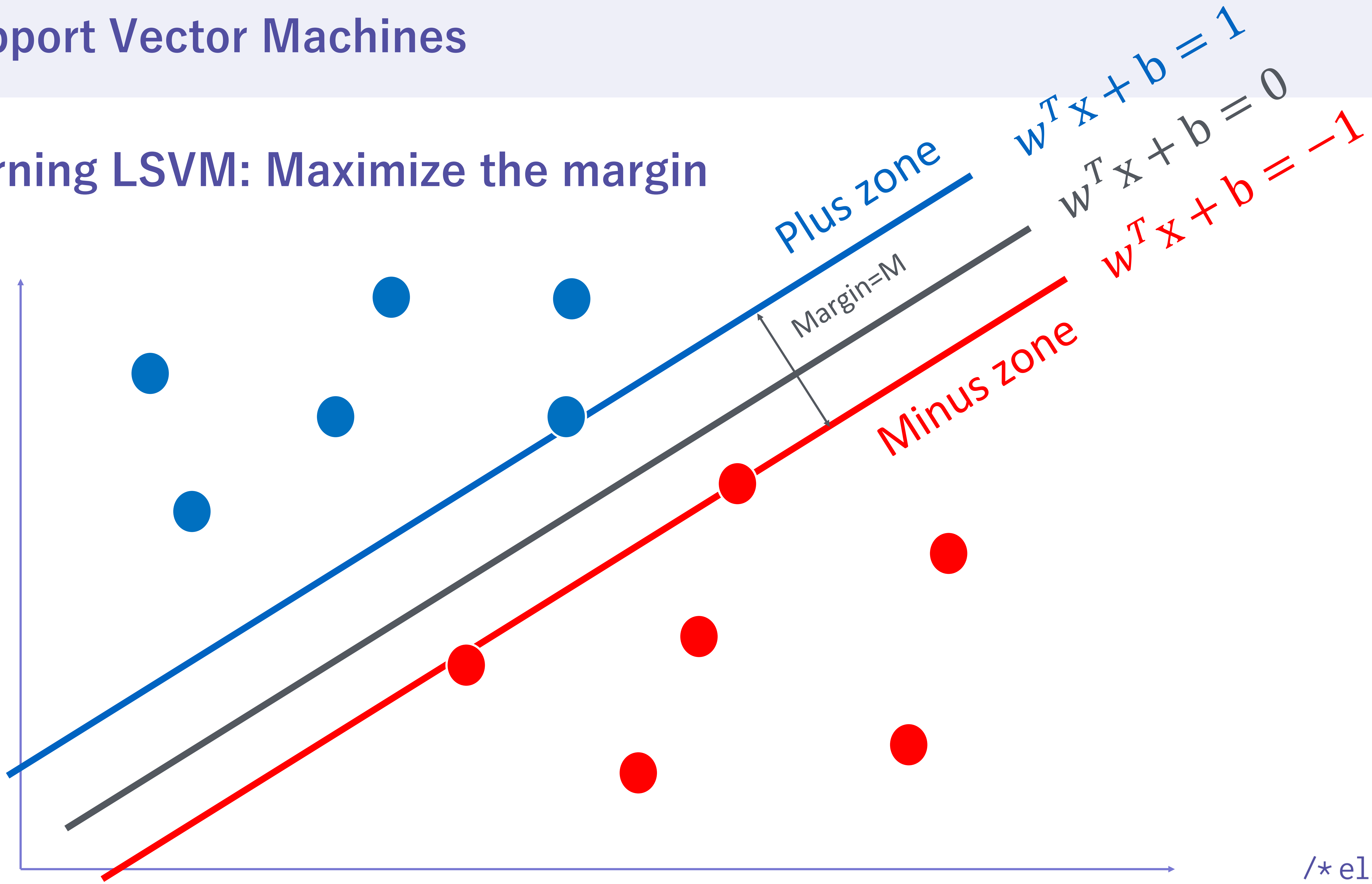
Non-Linear SVM: Kernel Trick ($y = x^2$)



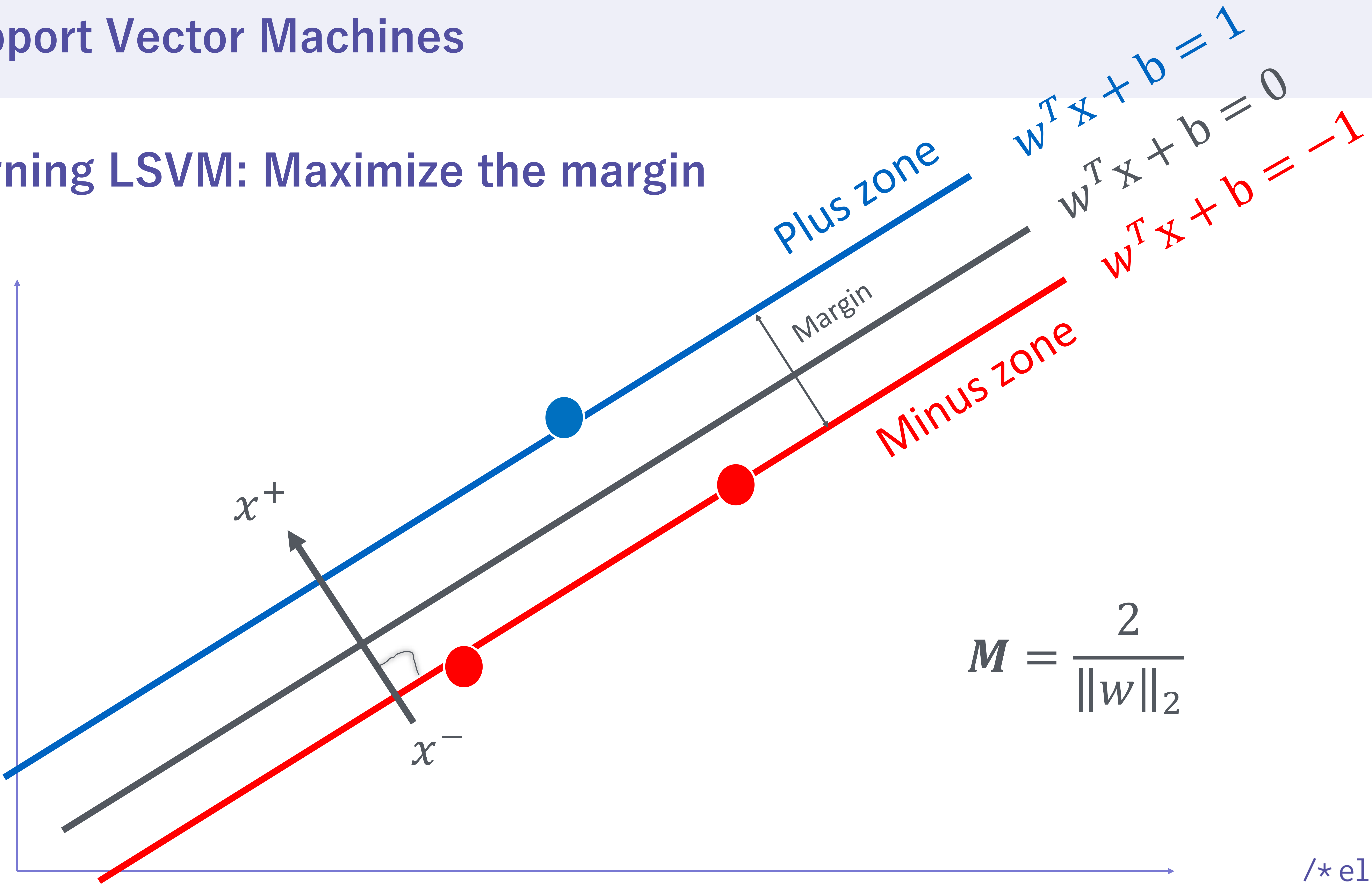
Non-Linear SVM: Kernel Trick



Learning LSVM: Maximize the margin



Learning LSVM: Maximize the margin



$$M = \frac{2}{\|w\|_2}$$

Learning LSVM: Maximize the margin

$$M = \|x^+ - x^-\|_2$$

Learning LSVM: Maximize the margin

$$M = \|x^+ - x^-\|_2$$

$$x^+ - x^- = \lambda w$$

Learning LSVM: Maximize the margin

$$M = \|x^{+} - x^{-}\|_2 = \|\lambda w\|_2 = \lambda\sqrt{w^T w}$$

$$x^{+} - x^{-} = \lambda w$$

Learning LSVM: Maximize the margin

$$M = \|x^+ - x^-\|_2 = \|\lambda w\|_2 = \lambda\sqrt{w^T w}$$

$$x^+ - x^- = \lambda w$$

$$w^T(x^- + \lambda w) + b = 1$$

Learning LSVM: Maximize the margin

$$M = \|x^+ - x^-\|_2 = \|\lambda w\|_2 = \lambda\sqrt{w^T w}$$

$$x^+ - x^- = \lambda w$$

$$w^T(x^- + \lambda w) + b = 1$$

$$\underbrace{w^T x^- + b}_{-1} + \lambda w^T w = 1 \qquad \therefore\ \lambda = \frac{2}{w^T w}$$

Learning LSVM: Maximize the margin

$$M = \|x^+ - x^-\|_2 = \|\lambda w\|_2 = \lambda\sqrt{w^T w} \Rightarrow M = \frac{2}{\sqrt{w^T w}} = \frac{2}{\|w\|_2}$$

$$x^+ - x^- = \lambda w$$

$$w^T(x^- + \lambda w) + b = 1$$

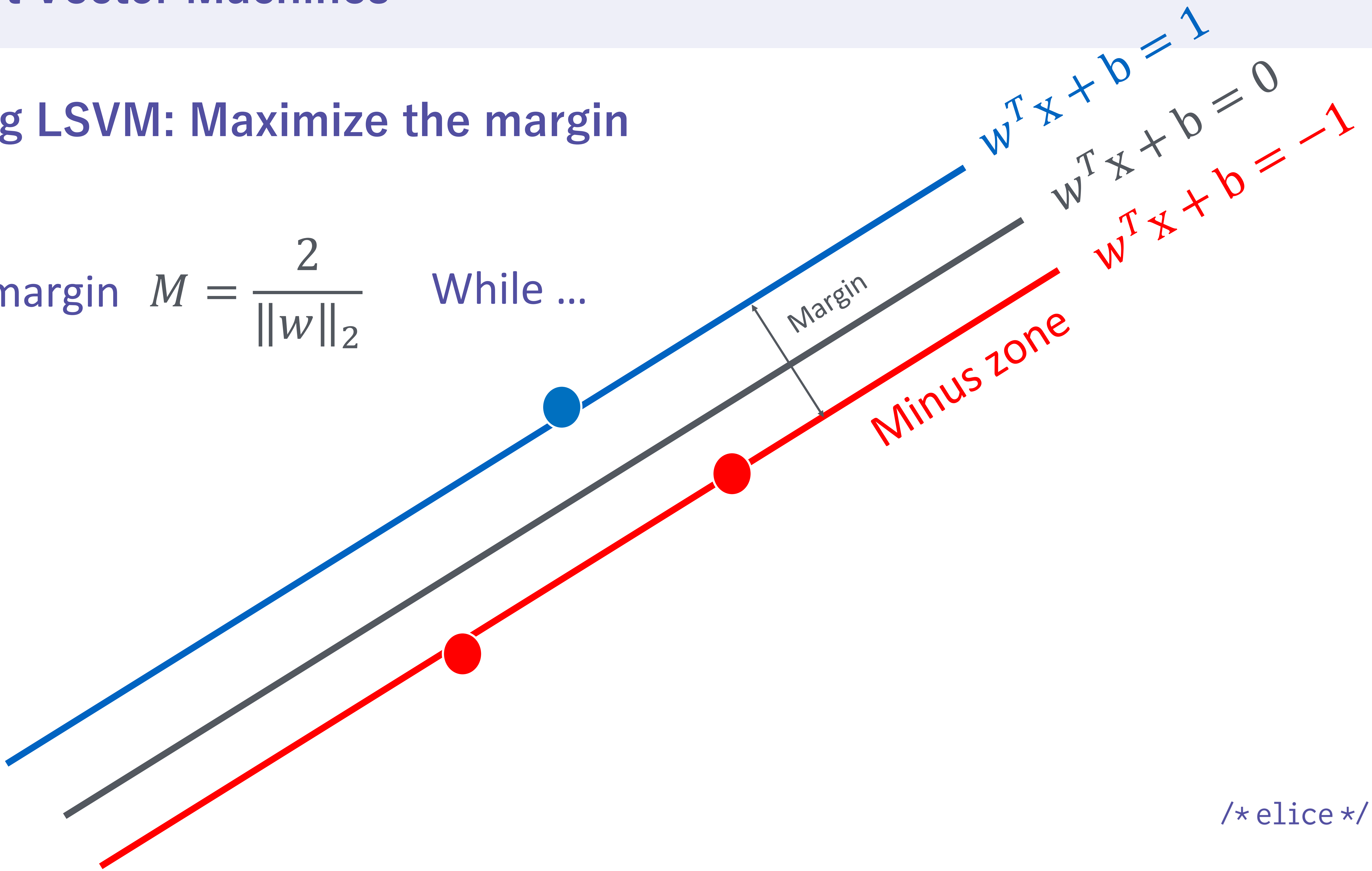
$$\underbrace{w^T x^- + b}_{-1} + \lambda w^T w = 1 \qquad \therefore\ \lambda = \frac{2}{w^T w}$$

Learning LSVM: Maximize the margin

Max margin $M = \dfrac{2}{\|w\|_2}$ While …

Learning LSVM: Maximize the margin

Max margin $M = \dfrac{2}{\|w\|_2}$ While …



Learning LSVM: Maximize the margin

Max margin $M = \frac{2}{\|w\|_2}$ While ... $w^T\mathrm{x} + \mathrm{b} \geq 1 \Rightarrow \mathrm{y} = 1$

$w^T\mathrm{x} + \mathrm{b} \leq -1 \Rightarrow \mathrm{y} = -1$

Learning LSVM: Maximize the margin

Max margin $M = \frac{2}{\|w\|_2}$ While … $y_i(w^T x_i + \mathrm{b}) \geq 1$

Learning LSVM: Maximize the margin

Max margin $M = \dfrac{2}{\|w\|_2}$ While … $y_i(w^T x_i + \mathrm{b}) \geq 1$

$$\Rightarrow \min \frac{1}{2}\|w\|_2 \Rightarrow \min \frac{1}{2}\|w\|_2^2$$

$\|w\|_2$이 제곱근을 포함하고 있기 때문에 계산이 어려워서

계산상의 편의를 위해 다음과 같은 형태로 변경

SVM 목적함수

Original Problem

$$\text{minimize } \frac{1}{2}\|w\|_2^2$$

$$\text{subject to } y_i(w^T x_i + b) \geq 1, i = 1,2, \dots \dots, n$$

SVM 목적함수: Original Form

Original Problem

$$\text{minimize } \frac{1}{2}\|w\|_2^2$$

$$\text{subject to } y_i(w^T x_i + b) \geq 1, i = 1,2, \dots \dots, n$$

Lagrangian Primal

$$\max_{\alpha} \min_{w,b} \mathcal{L}(w, b, \alpha) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} \alpha_i(y_i(w^T x_i + b) - 1)$$

$$\text{subject to } \alpha_i \geq 0, i = 1,2, \dots \dots, n$$

Lagrangian multiplier를 사용하여
목적식과 제약식을 하나의 식으로 표현가능

SVM 목적함수

Lagrangian Primal

$$\max_{\alpha} \min_{w,b} \mathcal{L}(w,b,\alpha) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} \alpha_i (y_i (w^T x_i + b) - 1)$$

$$subject\ to\ \alpha_i \geq 0, i = 1,2, \dots \dots, n$$

Convex, continuous이기 때문에 미분 = 0에서 최소값을 가짐

① $$\frac{\partial \mathcal{L}(w,b,\alpha)}{\partial w} = 0 \longrightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

② $$\frac{\partial \mathcal{L}(w,b,\alpha)}{\partial b} = 0 \longrightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

α 값을 알아야함

SVM 목적함수

$$\underbrace{\frac{1}{2}\|w\|_2^2}_{①} - \underbrace{\sum_{i=1}^{n} \alpha_i\left(y_i\left(w^T x_i + b\right) - 1\right)}_{②}$$

① $\frac{1}{2}\|w\|_2^2 = \frac{1}{2} w^T w$

$$= \frac{1}{2} w^T \textstyle\sum_{j=1}^{n} \alpha_j y_j x_j$$

$$= \frac{1}{2} \textstyle\sum_{j=1}^{n} \alpha_j y_j (w^T x_j)$$

$$= \frac{1}{2} \textstyle\sum_{j=1}^{n} \alpha_j y_j \left(\sum_{i=1}^{n} \alpha_i y_i x_i^T x_j\right)$$

$$= \frac{1}{2} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

RECALL

$$\boxed{\frac{\partial \mathcal{L}(w,b,\alpha)}{\partial w} = 0 \longrightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i}$$

SVM 목적함수

$$\underbrace{\frac{1}{2}\|w\|_2^2}_{①} - \underbrace{\sum_{i=1}^{n} \alpha_i\left(y_i\left(w^T x_i + b\right) - 1\right)}_{②}$$

② $-\sum_{i=1}^{n} \alpha_i\left(y_i\left(w^T x_i + b\right) - 1\right)$

$$= -\sum_{i=1}^{n} \alpha_i y_i\left(w^T x_i + b\right) + \sum_{i=1}^{n} \alpha_i$$

$$= -\sum_{i=1}^{n} \alpha_i y_i w^T x_i - b\sum_{i=1}^{n} \alpha_i y_i + \sum_{i=1}^{n} \alpha_i$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{n} \alpha_i$$

RECALL

$$\frac{\partial \mathcal{L}(w,b,\alpha)}{\partial b} = 0 \implies \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial \mathcal{L}(w,b,\alpha)}{\partial w} = 0 \implies w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

SVM 목적함수: Dual Form

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \left(\mathbf{x}_i \cdot \mathbf{x}_j\right)$$

$$\text{subject to} \begin{cases} \alpha_i \geq 0 \quad \text{for } i = 1, \cdots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

$$b = y_k - \mathbf{w} \cdot \mathbf{x}_k \quad \text{for any } \mathbf{x}_k \text{ such that } \alpha_k > 0$$

SVM Example: W와 b는?

* 시간: 10분

$$D = \{(1, 1, -1), (2, 2, +1)\}$$

S

D

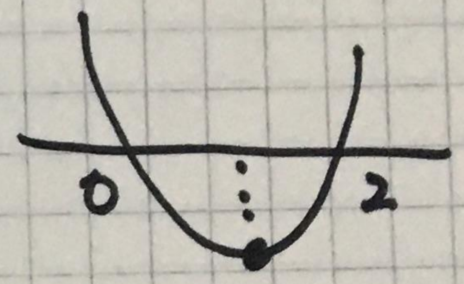
$$\max_{\alpha} (\alpha_1+\alpha_2) - \frac{1}{2}\left(y_1y_1\alpha_1\alpha_1 x_1\cdot x_1 + y_1y_2\alpha_1\alpha_2 x_1\cdot x_2 + y_2y_1\alpha_2\alpha_1 x_2\cdot x_1 + y_2y_2\alpha_2\alpha_2 x_2\cdot x_2\right)$$

subject to $\alpha_1 y_1 + \alpha_2 y_2 = 0$

대입!! (1,1,−1) (2,2,+1)

$$\Rightarrow \max_{\alpha} (\alpha_1+\alpha_2) - \frac{1}{2}\left(2\alpha_1^2 - 4\alpha_1\alpha_2 - 4\alpha_1\alpha_2 + 8\alpha_2^2\right)$$ subject to $-\alpha_1 + \alpha_2 = 0$

$$\Rightarrow \max_{\alpha} \left(2\alpha_1 - \alpha_1^2 + 4\alpha_1^2 - 4\alpha_1^2\right) \Rightarrow \min_{\alpha} \left(\alpha_1^2 - 2\alpha_1\right)$$



$\therefore \alpha_1 = \alpha_2 = 1$

$$\therefore w = \sum_{1}^{2} \alpha_i y_i x_i = (-1)\begin{bmatrix}1\\1\end{bmatrix} + \begin{bmatrix}2\\2\end{bmatrix} = \begin{bmatrix}1\\1\end{bmatrix}$$

$\therefore x_1 + x_2 - 3 = 0$ ■

$$\therefore b = 1 - \begin{bmatrix}1\\1\end{bmatrix}^T \begin{bmatrix}2\\2\end{bmatrix} = 1 - 4 = -3.$$

Credit

/* elice */

코스 매니저
이호정

콘텐츠 제작자
고혜선

강사
고혜선

감수자
이해솔

디자인
이해솔

Contact

TEL
070-4633-2015

WEB
https://elice.io

E-MAIL
contact@elice.io