k-NN, LDA, Naïve Bayes

강필성
고려대학교 산업경영공학부
pilsung_kang@korea.ac.kr

AGENDA

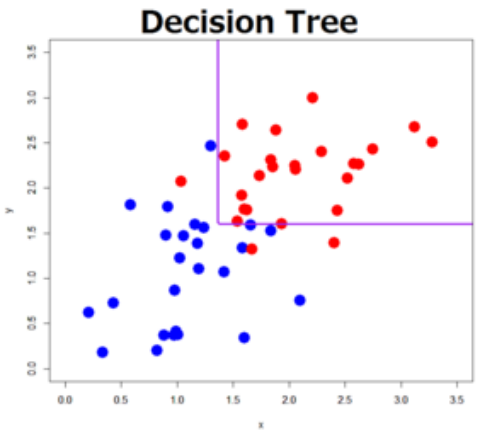
01 k-Nearest Neighbor

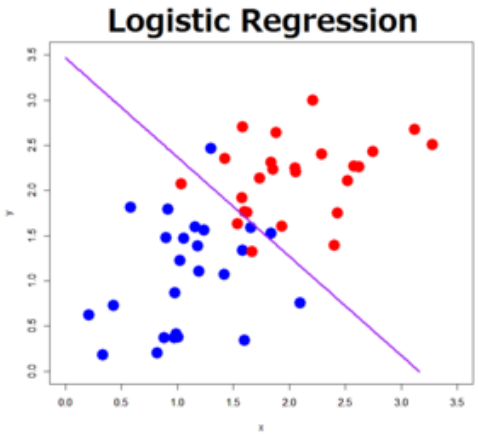
02 Linear Discriminant Analysis

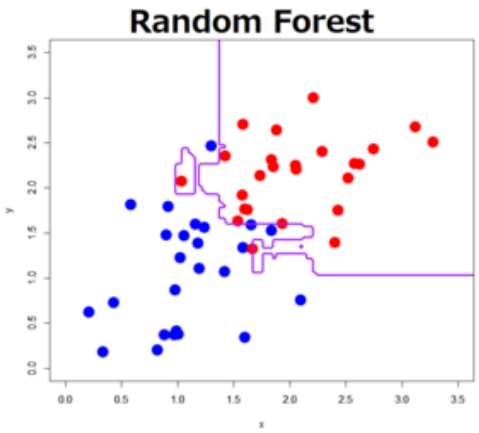
03 Naïve Bayesian Classifier

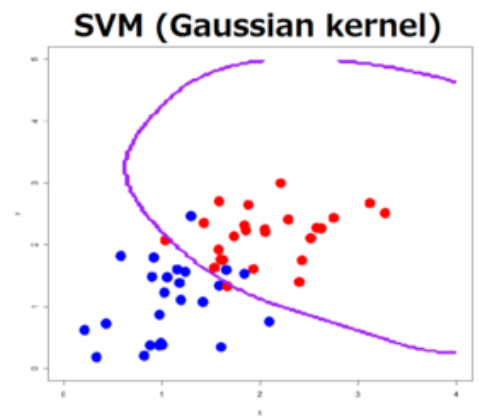
Backgrounds

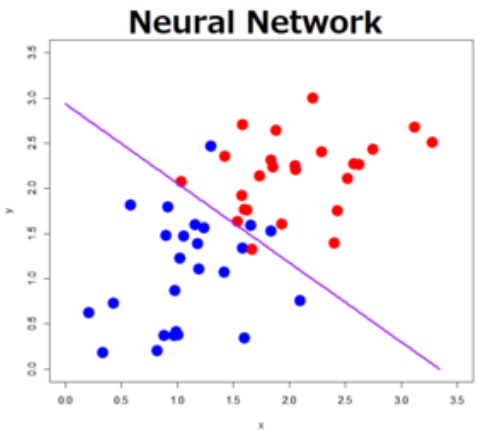
- 왜 여러가지의 Machine Learning 알고리즘을 알아야 하는가?
  - ✓특정한 알고리즘이 모든 상황에서 다른 알고리즘보다 우월하다는 결론을 내릴 수 없음





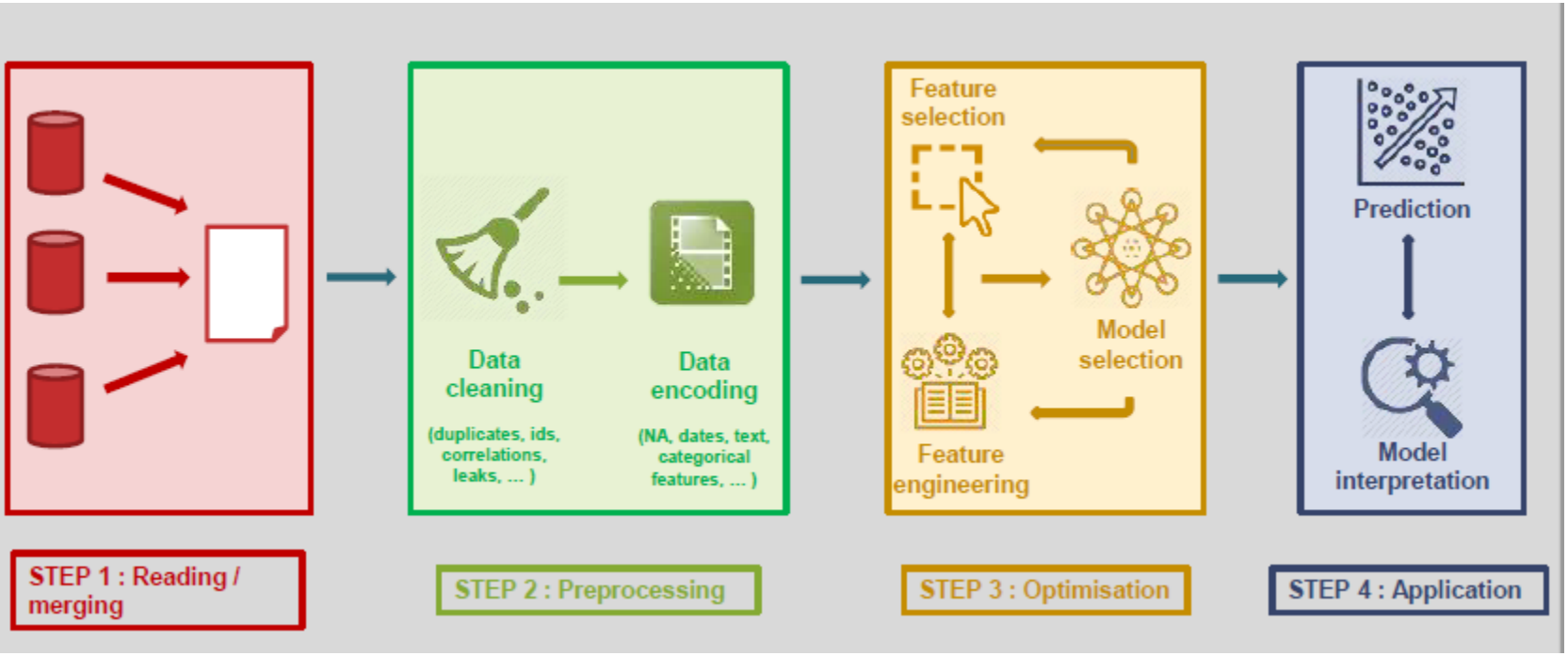






Backgrounds

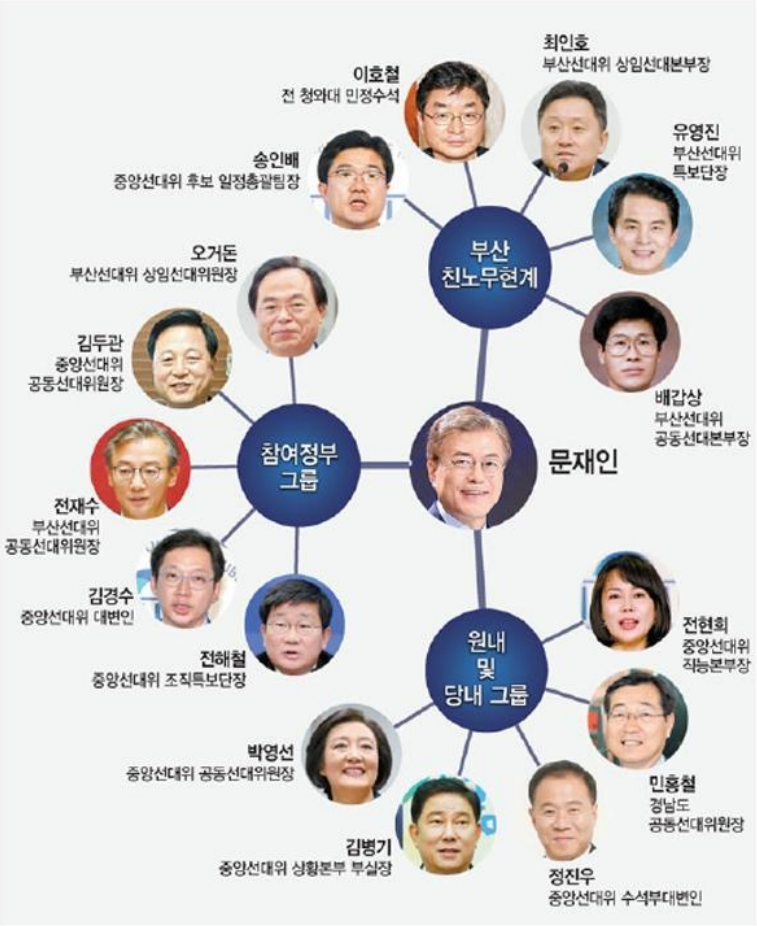
- AutoML
  - ✓ AS-IS: 아래 Machine Learning Pipeline을 숙달된 Data Scientist/Engineer가 수행
    - 얼마나 빠르게 우수한 모델을 완성시키는가? = 얼마나 높은 몸값이 요구되는가?
  - ✓ TO-BE: 데이터Data와 문제Task만 주세요, 나머지는 저희가 다 알아서 해드립니다…

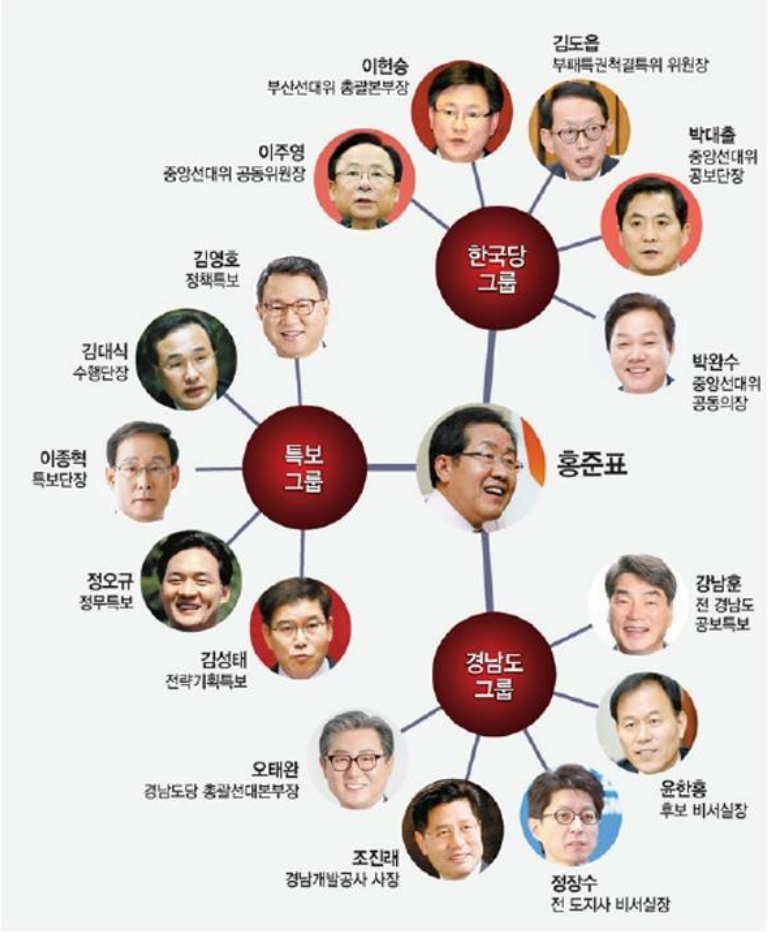


https://heartbeat.fritz.ai/automl-the-next-wave-of-machine-learning-5494baac615f

k-Nearest Neighbor

- 사람을 알려거든 그 사람의 친구가 누구인지를 보면 된다





k-Nearest Neighbor

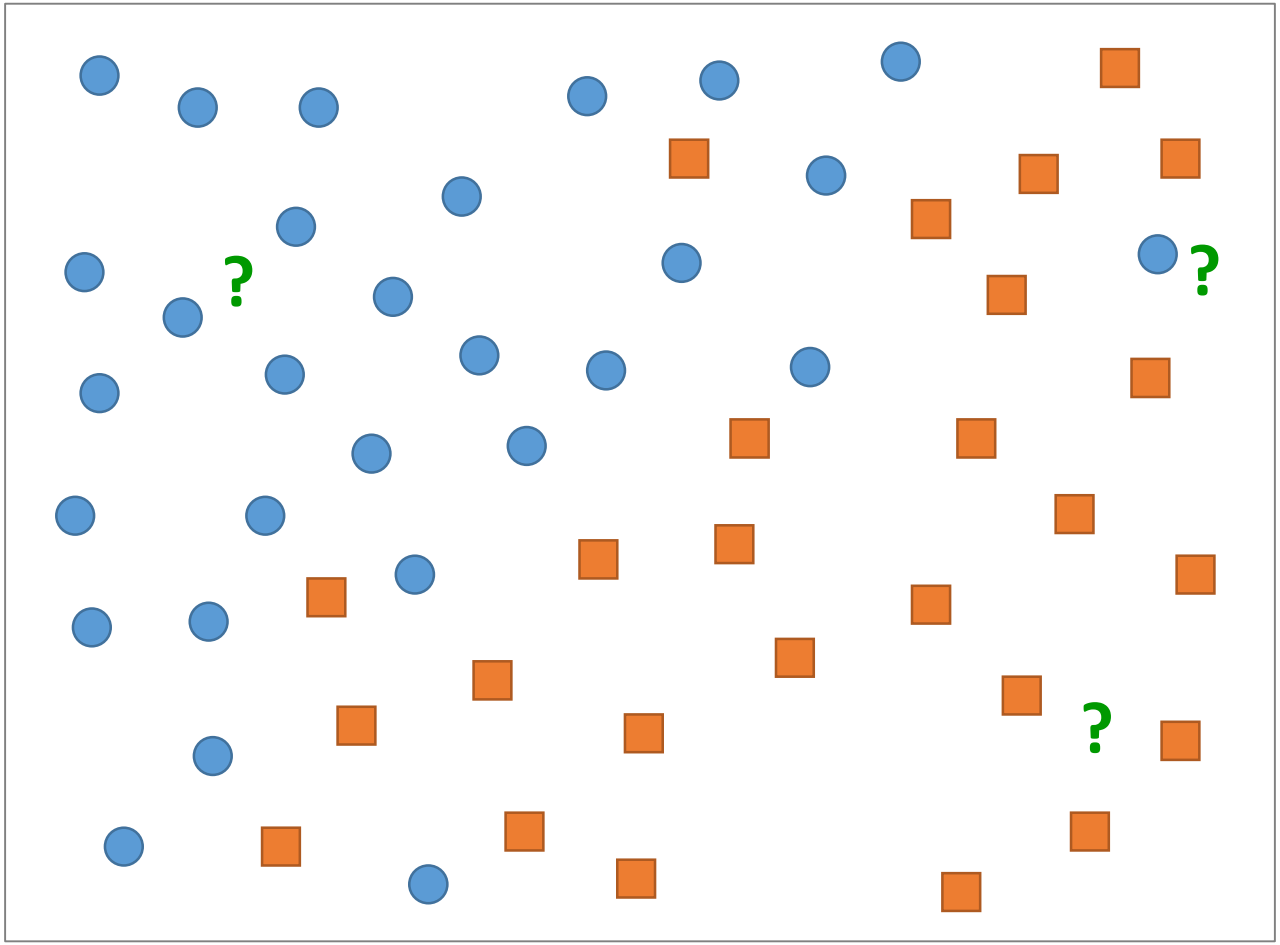
- 배경

類類相從　近墨者黑

"Birds of a feather flock together"

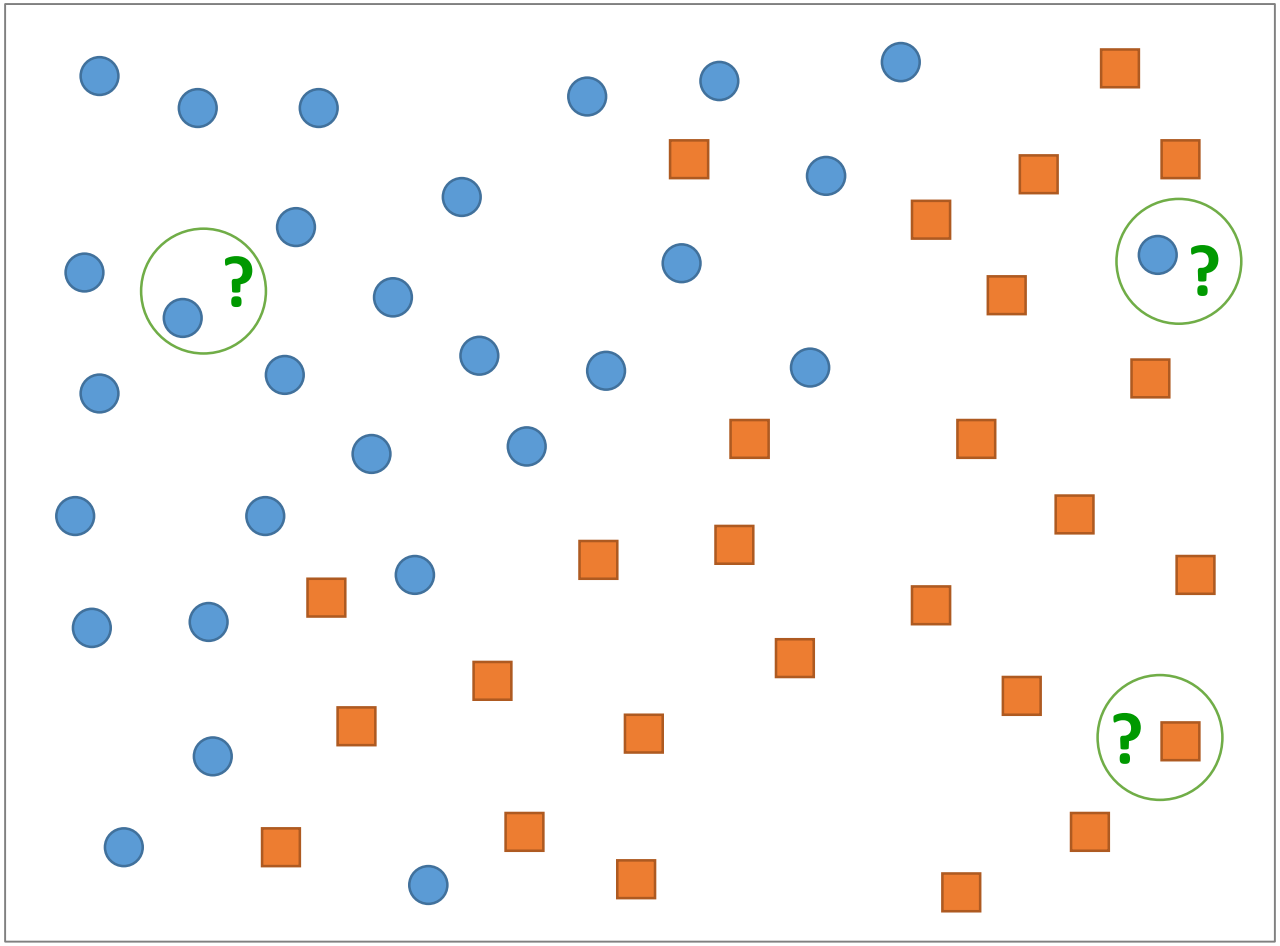
k-Nearest Neighbor

- 예시: 아래 물음표들은 동그라미인가? 아니면 네모인가?
  - ✓물음표들을 보는 순간 시야를 어디에 집중했는지 생각해 봅시다



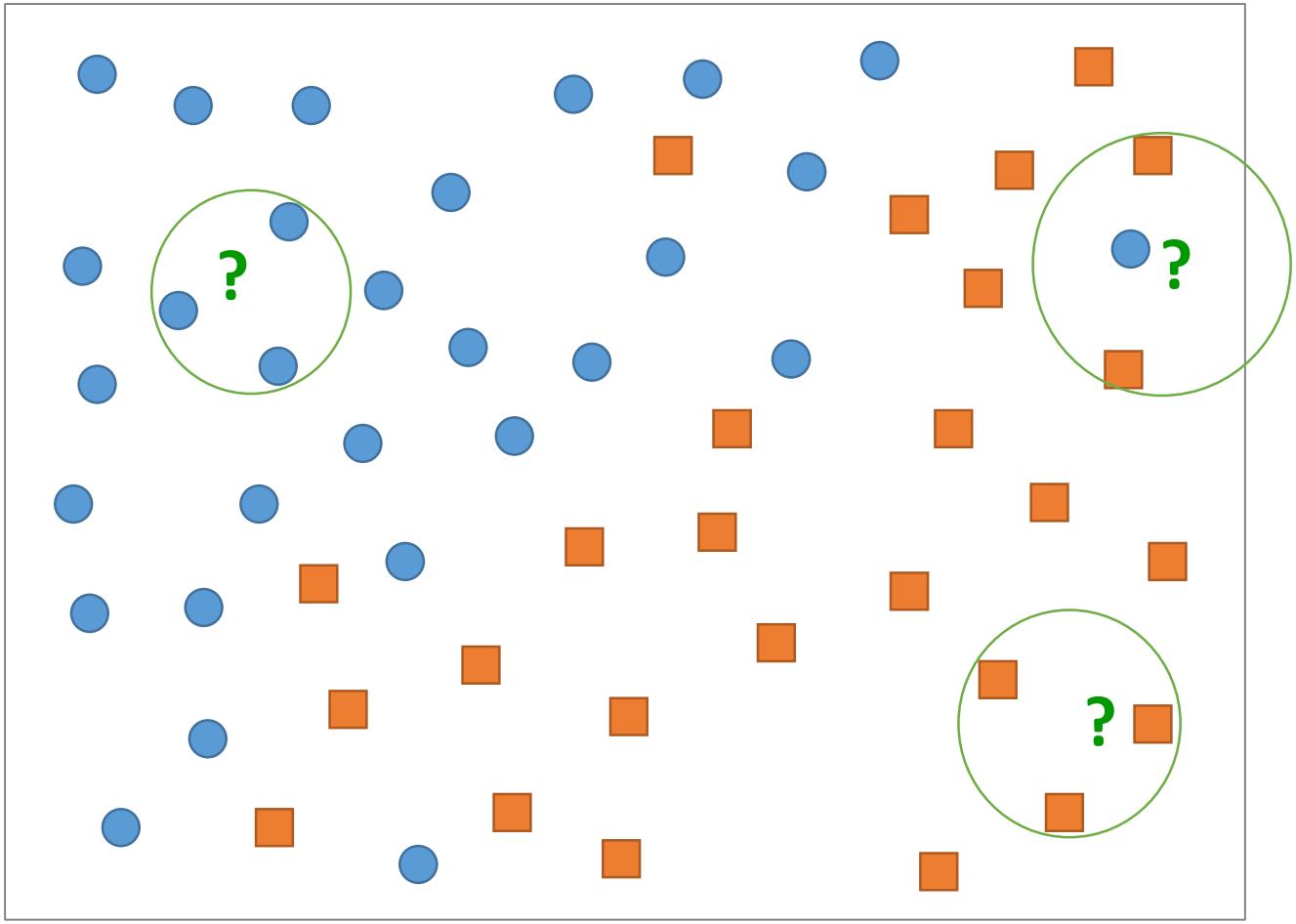
k-Nearest Neighbor

- 예시: 아래 물음표들은 동그라미인가? 아니면 네모인가?
  - ✓가장 가까운 객체 하나만 참고하면…



k-Nearest Neighbor

- 예시: 아래 물음표들은 동그라미인가? 아니면 네모인가?
  - ✓가까운 세 개의 객체들을 참고하면…



k-Nearest Neighbor: Classification

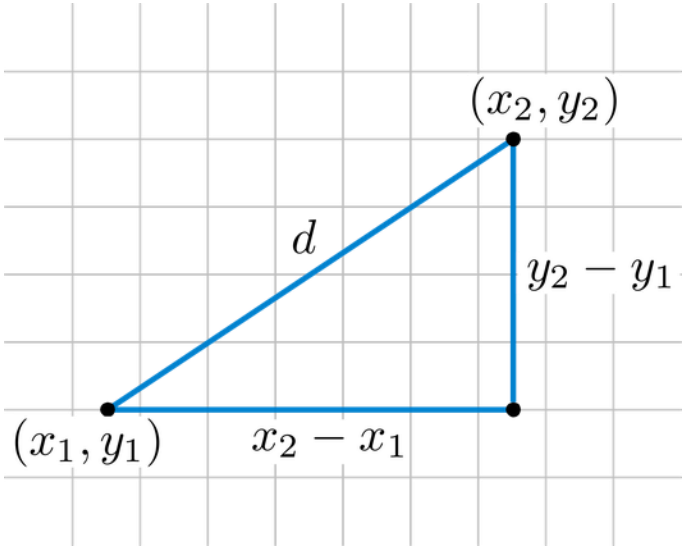
- k-NN Step 1: 참조 데이터Reference Data 준비
  - ✓분류 목적: 한 사람의 키, 몸무게, 체지방률을 이용해서 그 사람의 성별을 분류
  - ✓입력 변수 X: 키, 몸무게, 체지방률
  - ✓종속변수 Y: 성별 (M/F)
  - ✓각 범주로부터 충분한 수의 데이터 수집

개체	키	몸무게	체지방률	성별
1	187	93	15	M
2	165	51	25	F
3	174	68	14	M
4	156	48	29	F
…	…	…	…	…
N	168	59	12	M

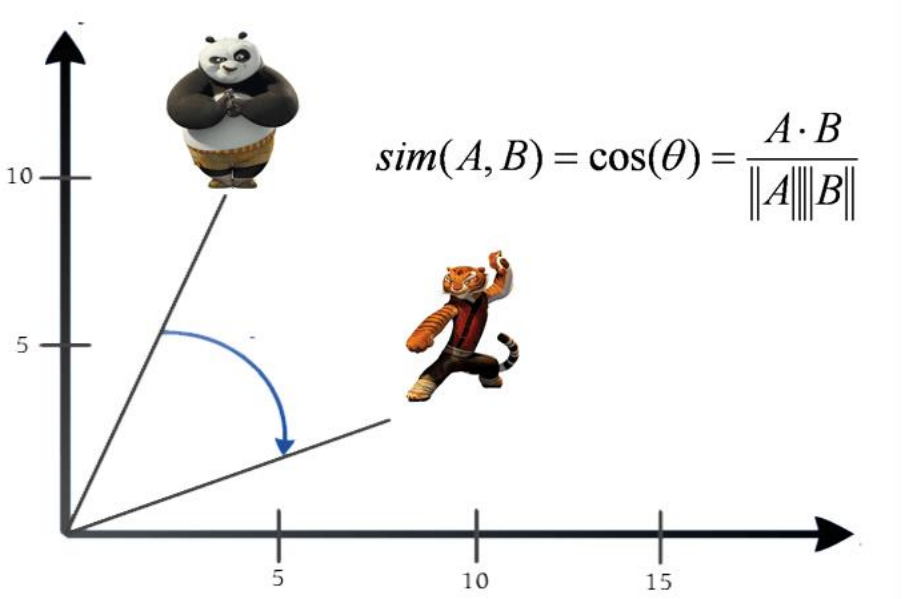
k-Nearest Neighbor: Classification

- k-NN Step 2: 유사도/거리 측정 지표 정의
  - ✓ 유사한 객체 = 거리가 가까운 객체

[Euclidean distance]



[Cosine similarity]



k-Nearest Neighbor: Classification

- k-NN Step 2: 유사도/거리 측정 지표 정의
  - ✓ 고려 사항: 정규화Normalization 또는 스케일링Scaling
    - 두 객체 사이의 거리를 계산할 때 정규화를 수행하지 않을 경우 측정 단위가 큰 변수가 측정 단위가 작은 변수보다 거리 계산에 더 큰 영향을 줌
    - 아래 예시에서는 키가 체지방률보다 두 객체 사이의 거리 계산에 큰 영향을 미침 (정규화 전)
    - 각 변수가 갖는 영향력을 동등하게 하기 위해 정규화/스케일링 수행

[정규화 전]

No.	키	몸무게	체지방률	성별
1	187	93	15	M
2	165	51	25	F
3	174	68	14	M
4	156	48	29	F
…	…	…	…	…
N	168	59	12	M
Avg.	165	65	20	-
Stdev.	15	10	5	-

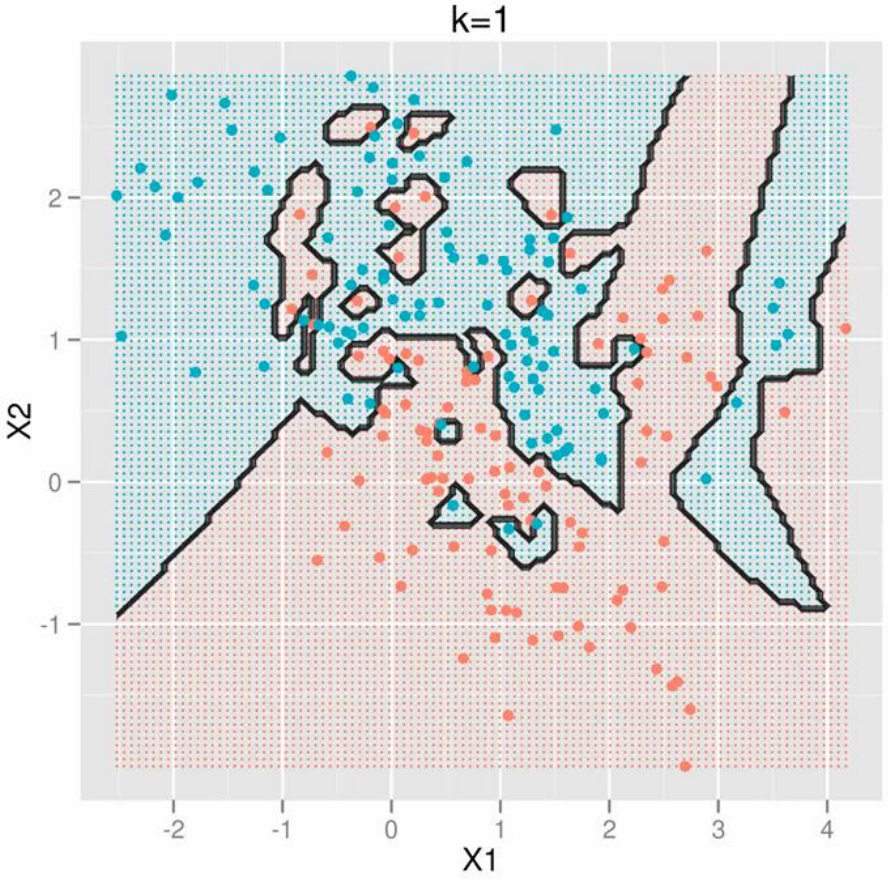
[정규화 후]

No.	키	몸무게	체지방률	성별
1	1.47	2.80	-1.00	M
2	0.00	-1.40	1.00	F
3	0.60	0.30	-1.20	M
4	-0.60	-1.70	1.80	F
…	…	…	…	…
N	0.20	-0.60	-1.60	M

k-Nearest Neighbor: Classification

- **k-NN Step 3: k의 후보 집합**Hyper-parameter Search Space**생성**

  ✓k 값의 변화에 따른 분류 경계면 예시 (k가 짝수인 경우는 무시: 동률 발생)



- k가 매우 작게 설정되면 분류 경계면이 노이즈Noise에 민감하게 되어 과적합Over-fitting의 우려가 있음

- k가 매우 크게 설정되면 적절한 지역적 구조Local Structure를 파악하는 능력을 잃고 부적합Under-fitting의 경향성을 보임

- 적절한 k값을 찾아내는 것이 k-NN의 성능 최적화에 가장 중요한 요소이며, 일반적으로 충분한 범위의 k 값들 중에서 검증 데이터 오류가 가장 낮은 값을 선택

k-Nearest Neighbor: Classification

- **k-NN Step 4: k개의 이웃들로부터 범주 정보를 취합하여 예측 수행**
  - ✓ 이웃들의 거리정보 반영 유무에 따라 다수결Majority Voting 방식과 가중합Weighted Voting 방식이 있음
    - 다수결 방식: k개의 이웃들은 최종 예측에 동등한 영향력을 발휘함
    - 가중합 방식: 이웃들이 최종 예측에 미치는 영향력은 각 이웃들의 유사도에 비례/거리에 반비례

For a new data

X

이웃	범주	거리	1/거리	가중치
N1	M	1	1.00	0.44
N2	F	2	0.50	0.22
N3	M	3	0.33	0.15
N4	F	4	0.25	0.11
N5	F	5	0.20	0.08

- 다수결 방식: P(X=M) = 2/5 = 0.4, 가중합 방식: P(X=F) = 0.59
- 분류 기준점을 0.5로 설정할 경우, 새로운 개체 X는 다수결에 의해 여성으로 판별되고 가중합에 의해서는 남성으로 판별됨

k-Nearest Neighbor: Classification

- k-NN Step 4: k개의 이웃들로부터 범주 정보를 취합하여 예측 수행
  - ✓ 고려 사항: 분류 기준값Cut-off/Threshold는 참조 데이터의 범주별 사전 확률Prior Probability을 고려할 필요가 있음
    - 만일 참조 데이터셋에 남성이 100명, 여성이 400명 존재했을 경우

For a new data

X

Neighbor	Class
N1	M
N2	F
N3	M
N4	F
N5	F

Majority voting

P(X=M)=0.4

    - 분류 기준값이 0.5로 설정된 경우 (범주간 사전 확률이 동일), X는 여성으로 분류됨
    - 분류 기준값이 0.2(남성의 사전 확률)로 설정된 경우, X는 남성으로 분류됨

k-Nearest Neighbor: Regression

- 이웃의 정보를 어떻게 결합할 것인가?
  - ✓ 예제: 한 사람의 키/몸무게/성별로부터 체지방률을 추정

개체	키	몸무게	성별(F=1)	체지방률
1	187	93	0	15
2	165	51	1	25
3	174	68	0	14
4	156	48	1	29
…	…	…	…	…
N	168	59	0	12

k-Nearest Neighbor: Regression

- 이웃의 정보를 어떻게 결합할 것인가?
  - ✓ 단순 평균Simple average vs. 가중 평균Weighted average

For a new data X

이웃	체지방률	거리	1/거리	가중치
N1	15.4	1	1.00	0.44
N2	17.2	2	0.50	0.22
N3	12.3	3	0.33	0.15
N4	11.5	4	0.25	0.11
N5	10.9	5	0.20	0.08

  - 단순 평균 이용
    - X의 체지방률 = (15.4+17.2+12.3+11.5+10.9)/5 = 13.46
  - 가중 평균 이용
    - X의 체지방률 = 0.44*15.4+0.22*17.2+0.15*12.3+0.11*11.5+0.08*10.9 = 14.54

k-Nearest Neighbor: Regression

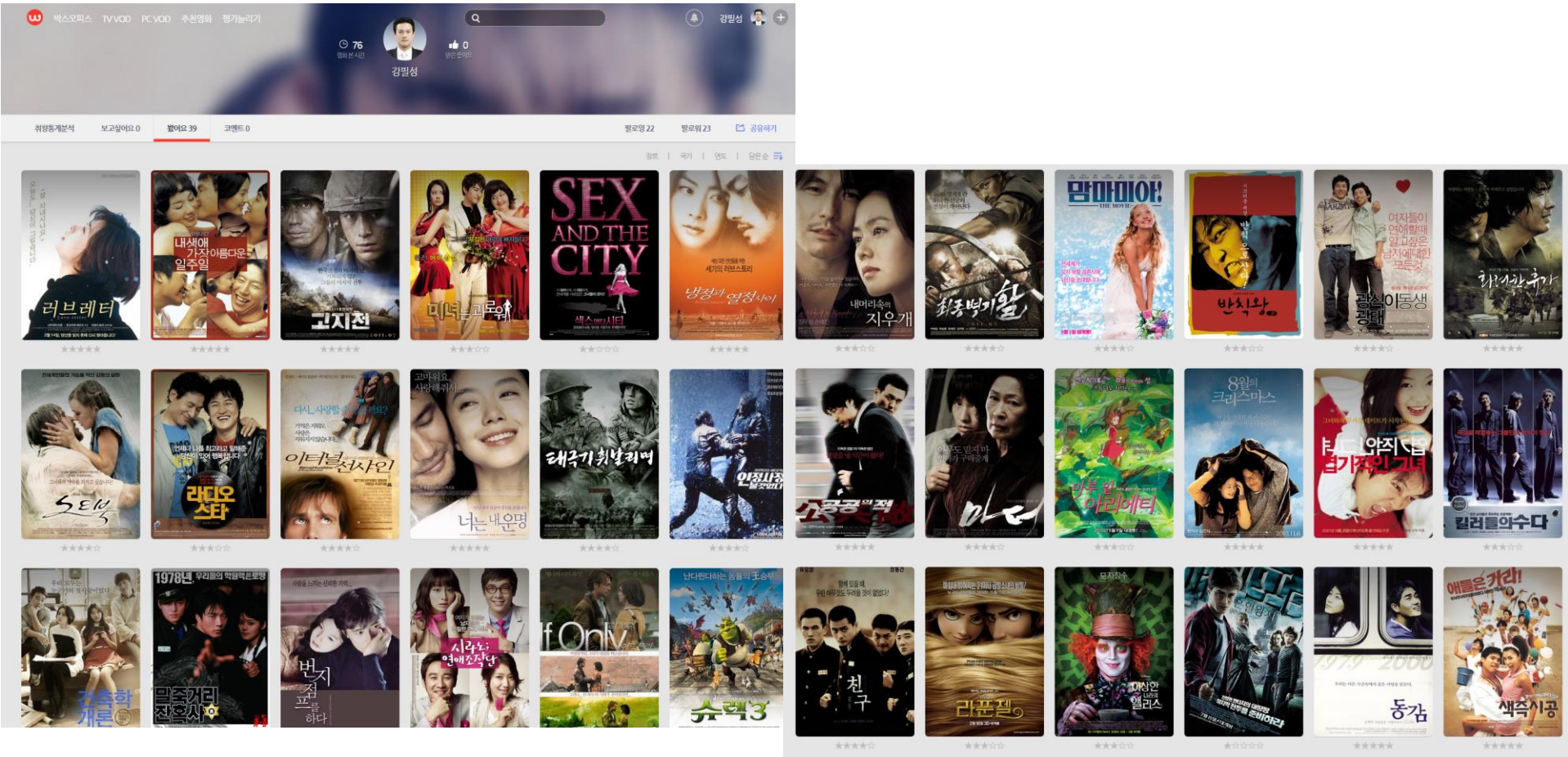
- 영화 추천 시스템: 협업 필터링

	영화 1	영화 2	영화 3	영화 4	영화 5	…	영화 D
강필성	10	9	5	6	9	…	? 9

User	영화 1	영화 2	영화 3	영화 4	영화 5	…	영화 D
1	10	8	4	7	10	…	10
2	8	5	7	9	4	…	5
3	10	9	6	5	8	…	9
4	4	2	10	10	5	…	3
5	7	4	6	8	5	…	3
6	5	2	10	10	10	…	6
7	10	8	6	6	8	…	8
…	…	…	…	…	…	…	…
N	5	7	1	5	4	…	7

k-Nearest Neighbor: Regression

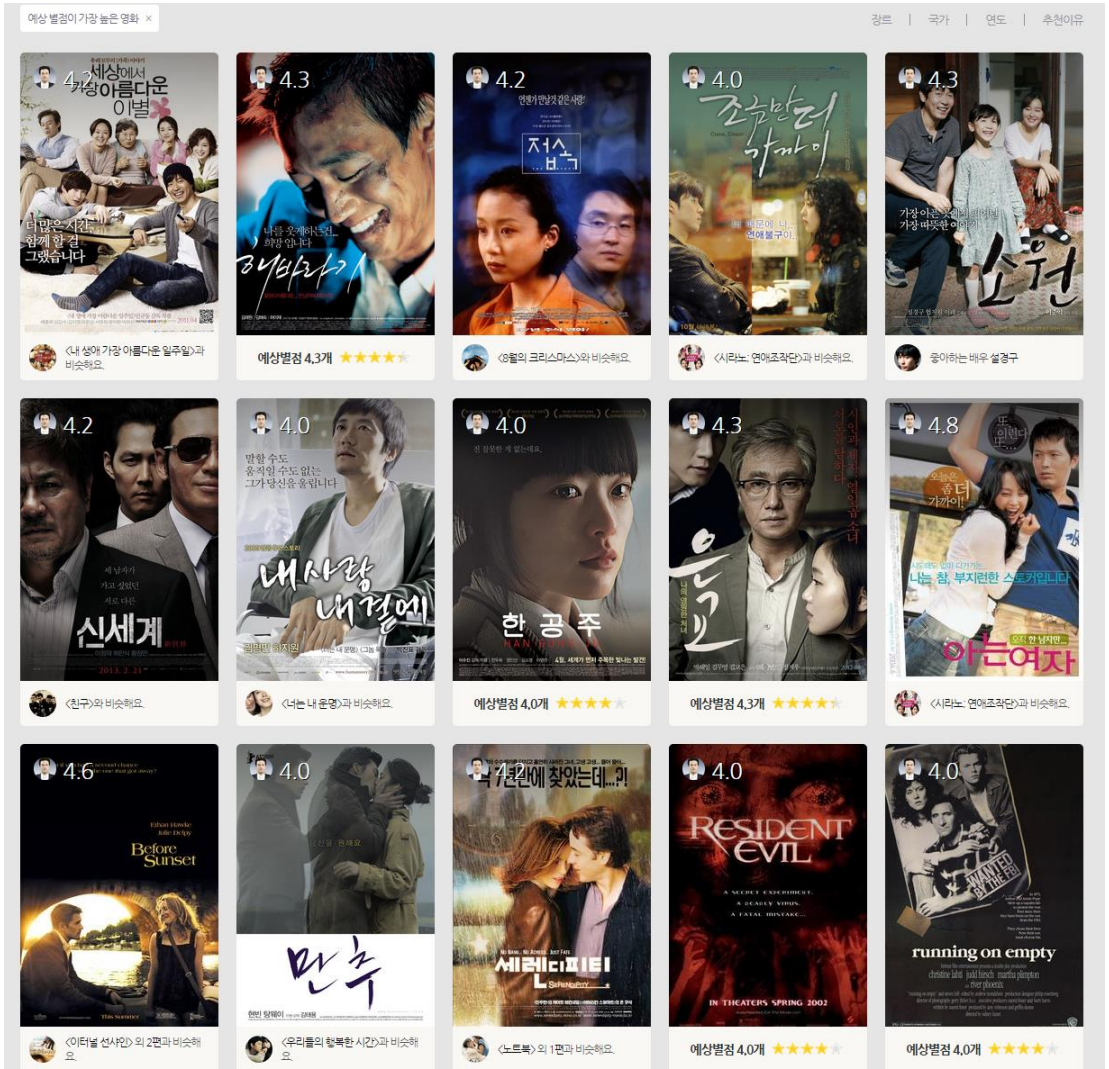
- 예시: 사용자 선호도 기반의 영화 추천 시스템



http://watcha.net

k-Nearest Neighbor: Regression

- 추천 영화 리스트



AGENDA

01 k-Nearest Neighbor

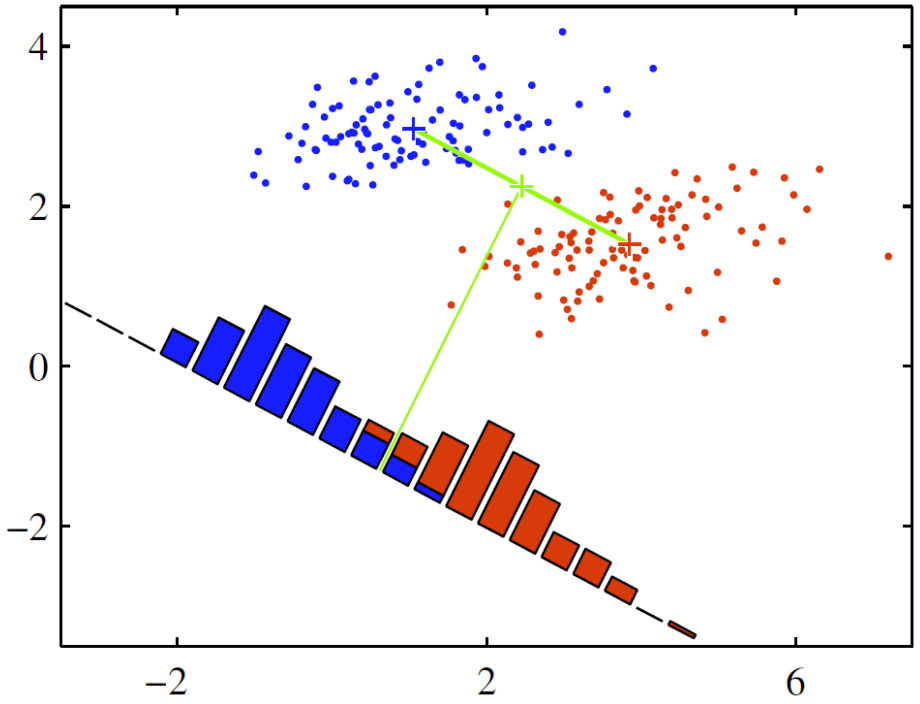
02 Linear Discriminant Analysis

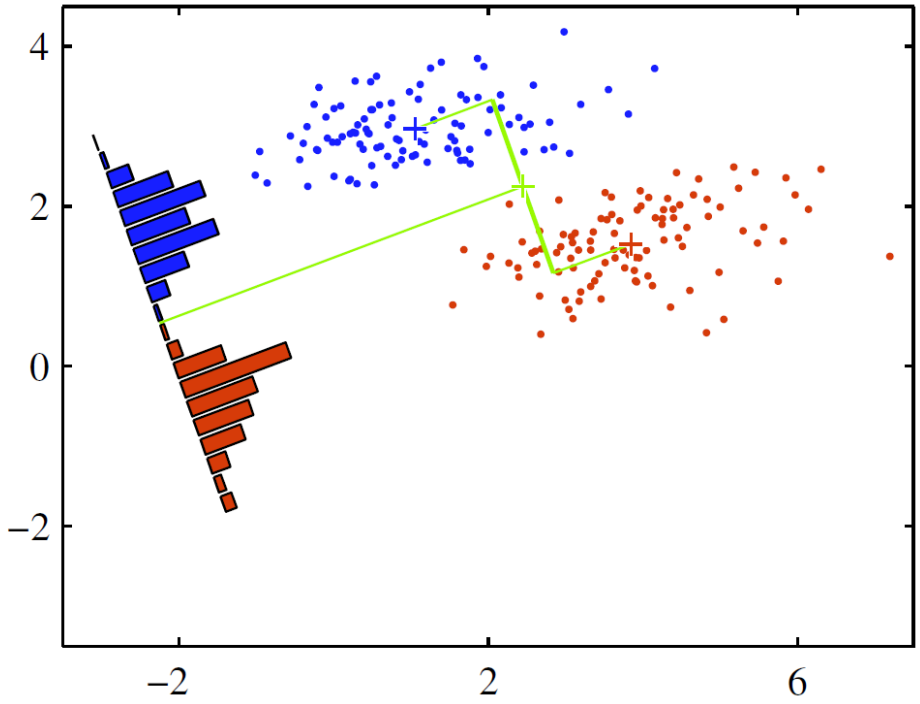
03 Naïve Bayesian Classifier

Linear Discriminant Analysis: LDA

- 선형판별분석: 배경

  ✓아래 두 직선 중 사영projection 후에 두 범주를 보다 잘 구분할 수 있는 직선은?



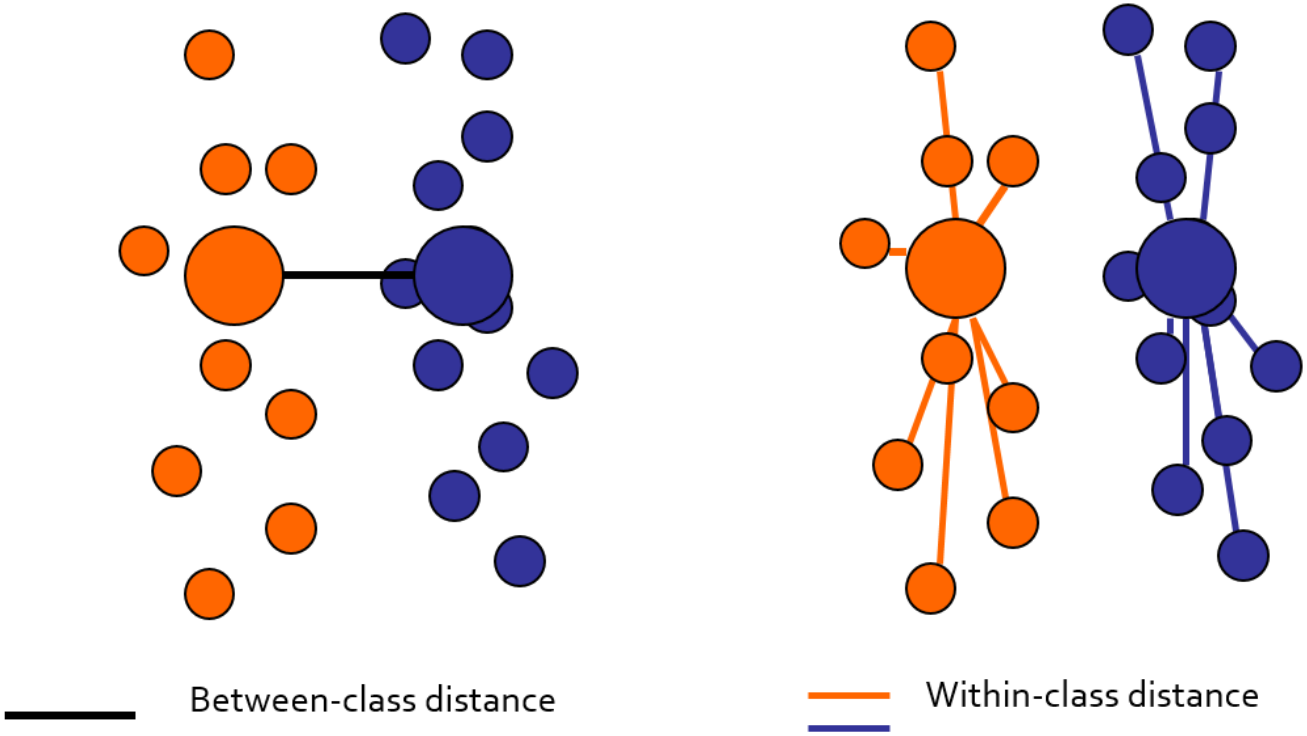


(Source: Bishop (2006))

- 목적: 사영 후 두 범주를 가장 잘 분류할 수 있는 하나의 직선을 찾자!

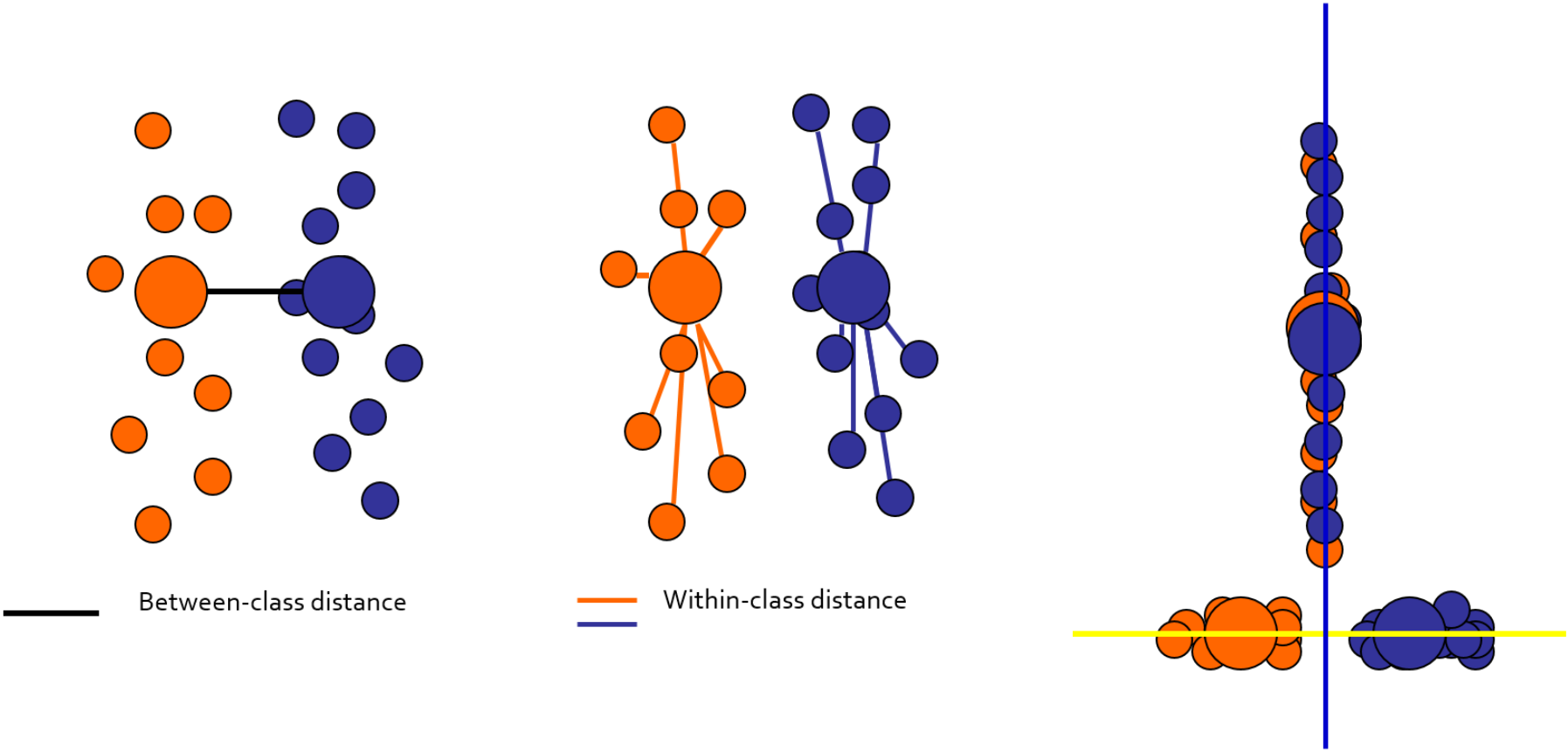
Linear Discriminant Analysis: LDA

- 두 종류의 Class distance
  - ✓ Between-class distance: 범주 중심 간 거리
  - ✓ Within-class distance: 범주 내 개체들의 평균 거리



Linear Discriminant Analysis: LDA

- (Fisher's) Linear Discriminant Analysis
  - ✓ Find most discriminant projection by maximizing between-class distance (variance) and minimizing within-class distance (variance)



Linear Discriminant Analysis: LDA

- **선형판별분석: 절차**
  - ✓ d차원의 입력 벡터 x를 w라는 벡터에 사영시킨 후 생성되는 1차원상의 좌표 값

$$y = \mathbf{w}^T \mathbf{x}$$

  - ✓ N_1 개와 N_2 개의 관측치를 갖는 C_1 과 C_2 두 범주에 대해 원래 입력 공간에서 각 범주의 중심Mean

$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{n \in C_1} \mathbf{x}_n, \quad \mathbf{m}_2 = \frac{1}{N_2} \sum_{n \in C_2} \mathbf{x}_n$$

  - ✓ 목적 1: 사영 후 두 범주의 중심이 멀리 떨어져 위치하는 벡터 w를 찾자: maximize between class variance

$$m_2 - m_1 = \mathbf{w}^T (\mathbf{m}_2 - \mathbf{m}_1), \quad m_k = \mathbf{w}^T \mathbf{m}_k$$

Linear Discriminant Analysis: LDA

- 선형판별분석: 절차
  - ✓목적 2: 사영 후 각 범주에 속한 관측치들은 해당 범주의 중심에 가까이 위치하는 벡터 w를 찾자: minimize within class variance

$$s_k^2 = \sum_{n \in C_k} (y_n - m_k)^2$$

  - ✓최적화 문제로 변환
    - 두 개의 목적(범주간 분산 최대화, 범주내 분산 최소화)를 동시에 만족시키기 위한 목적함수

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T$$

$$\mathbf{S}_W = \sum_{n \in C_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T + \sum_{n \in C_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T$$

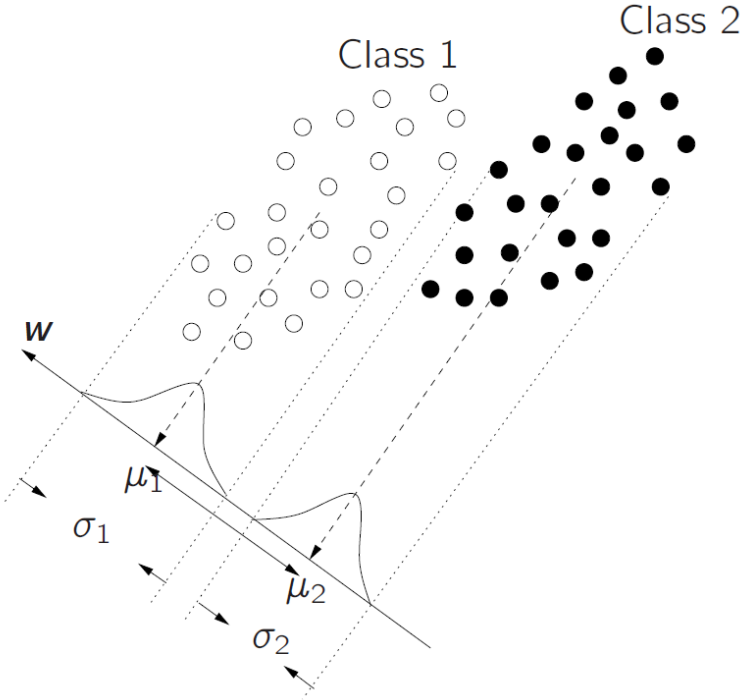
Linear Discriminant Analysis: LDA

- **선형판별분석: 절차**
  - ✓사영 벡터 **w**를 어떻게 찾을 수 있는가?
    - 목적함수 ***J*(w)** 는 **w**에 대해 1차 미분한 값이 0이 되는 지점에서 최대값을 가짐

$$(\mathbf{w}^T\mathbf{S}_B\mathbf{w})\mathbf{S}_W\mathbf{w} = (\mathbf{w}^T\mathbf{S}_W\mathbf{w})\mathbf{S}_B\mathbf{w}$$

    - $\mathbf{S}_B\mathbf{w}$ 는 항상 $(\mathbf{m}_2\text{-}\mathbf{m}_1)$의 방향에 위치
    - 상수항인 $(\mathbf{w}^T\mathbf{S}_B\mathbf{w})$ 와 $(\mathbf{w}^T\mathbf{S}_W\mathbf{w})$를 소거할 수 있음
    - 이를 통해 다음과 같은 사영벡터 **w**를 구할 수 있음

$$\mathbf{w} \propto \mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$



AGENDA

01 k-Nearest Neighbor

02 Linear Discriminant Analysis

03 Naïve Bayesian Classifier

Naïve Bayesian Classifier

- 다음 사람들의 성별은 무엇인가?

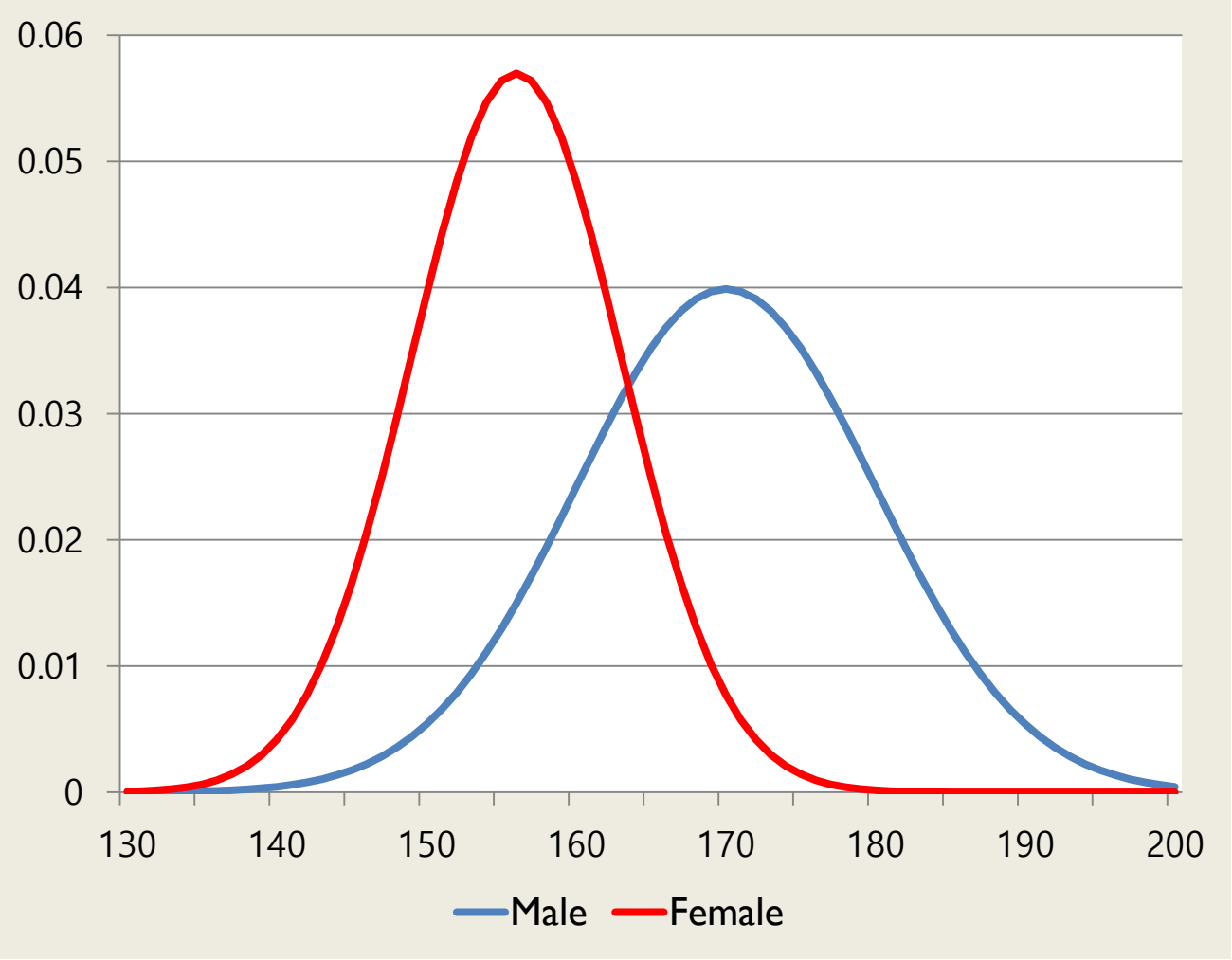
Men

Vs.

Women

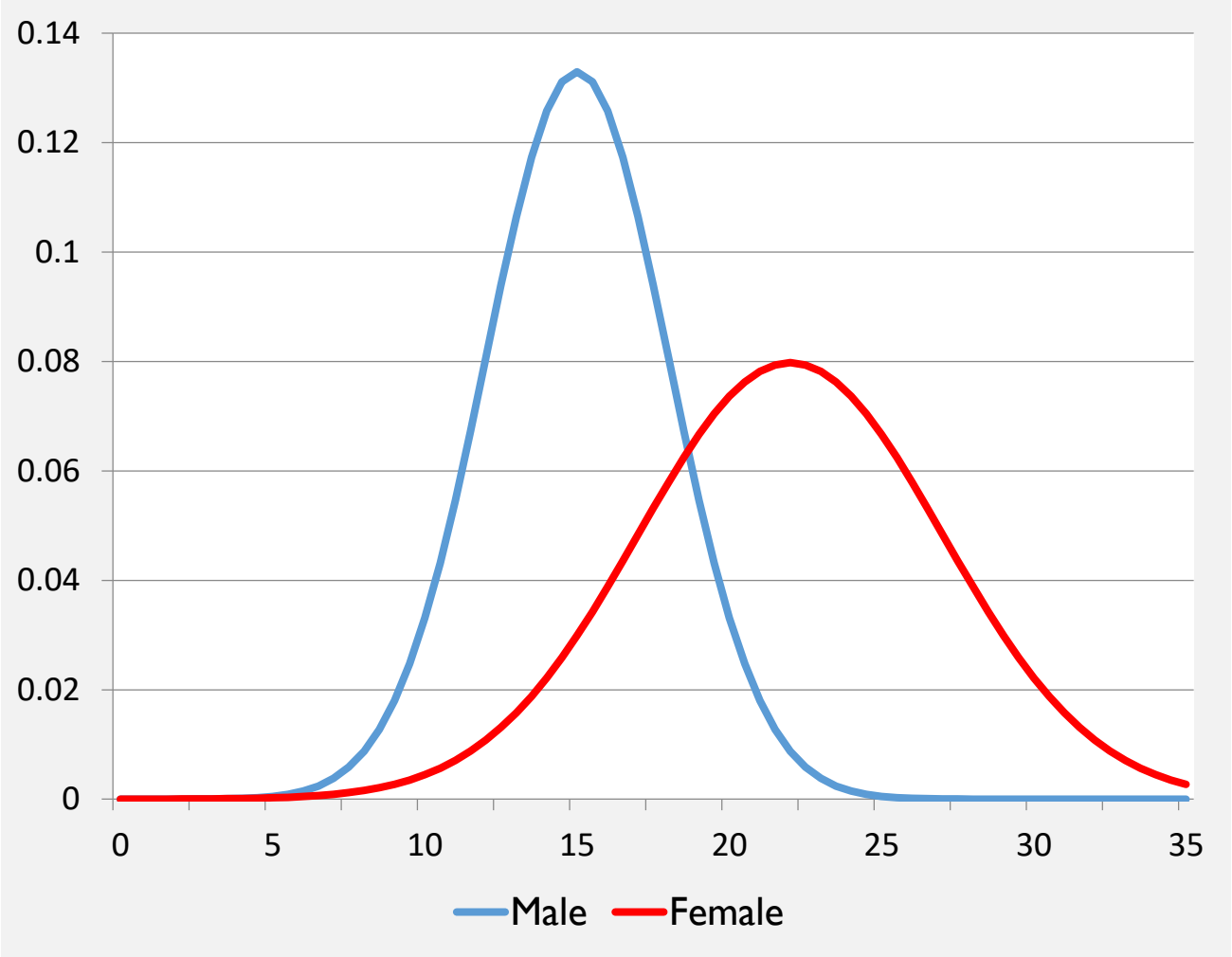
Naïve Bayesian Classifier

- 만약 남자와 여자의 키에 대한 사전 분포를 미리 알고 있다면…



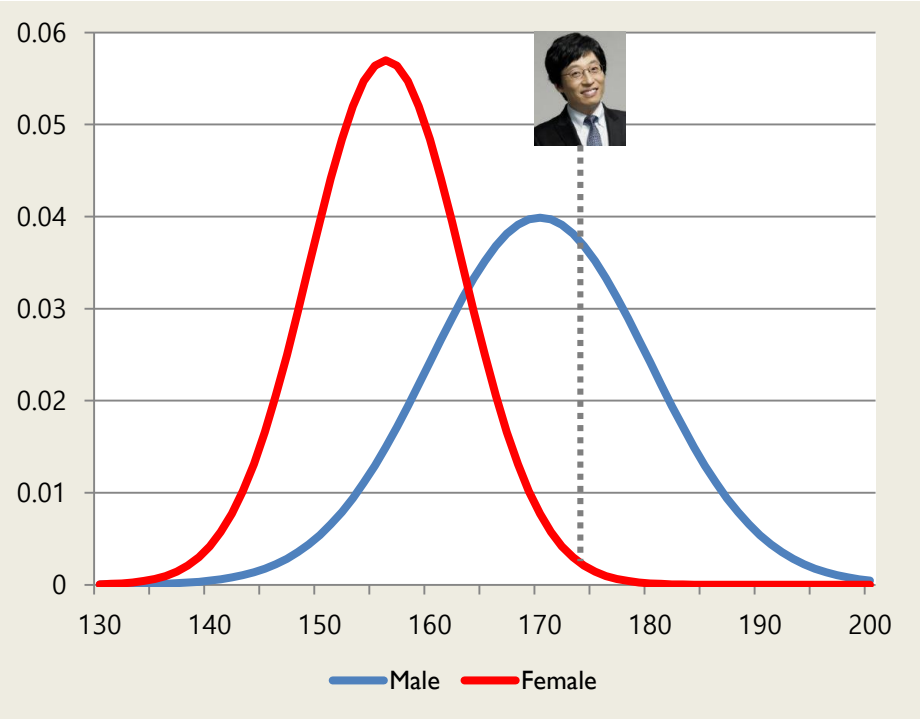
Naïve Bayesian Classifier

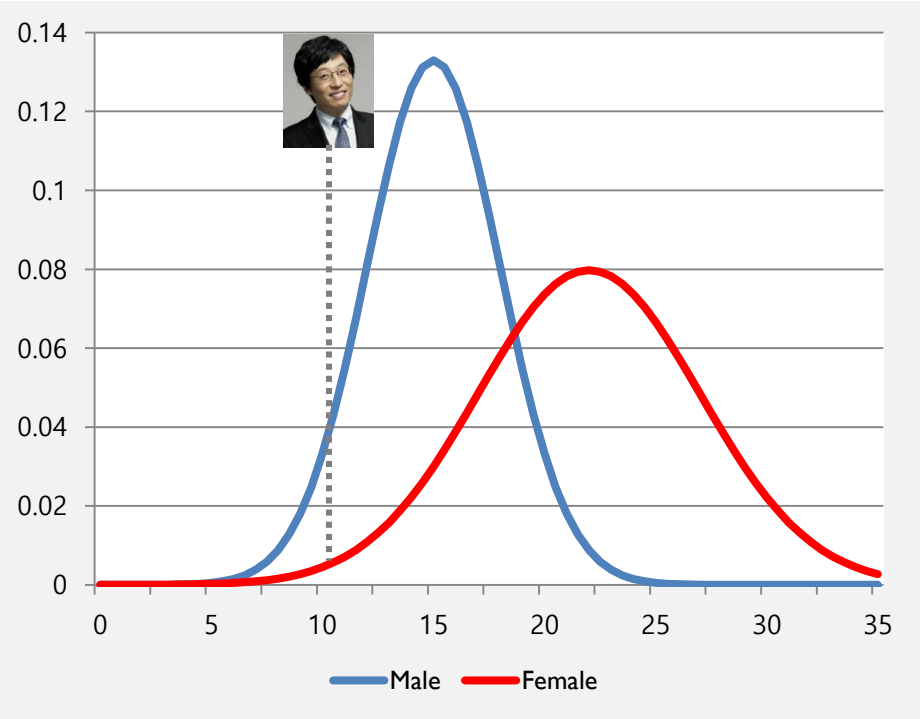
- 만약 남자와 여자의 체지방률에 대한 사전 분포도 미리 알고 있다면…



Naïve Bayesian Classifier

- 유재석은 남자일까 여자일까?

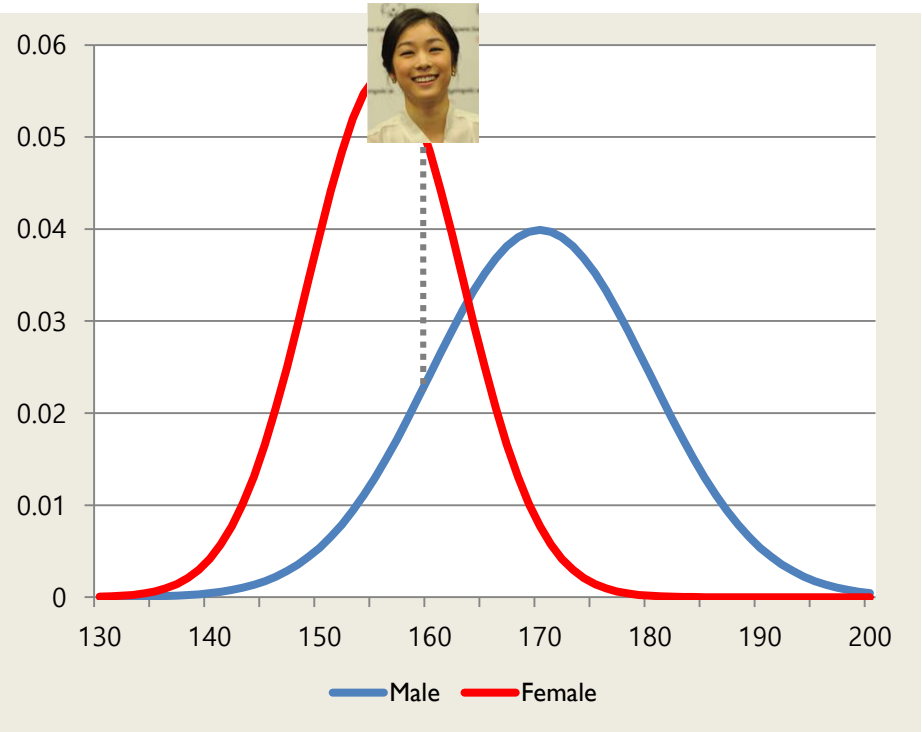


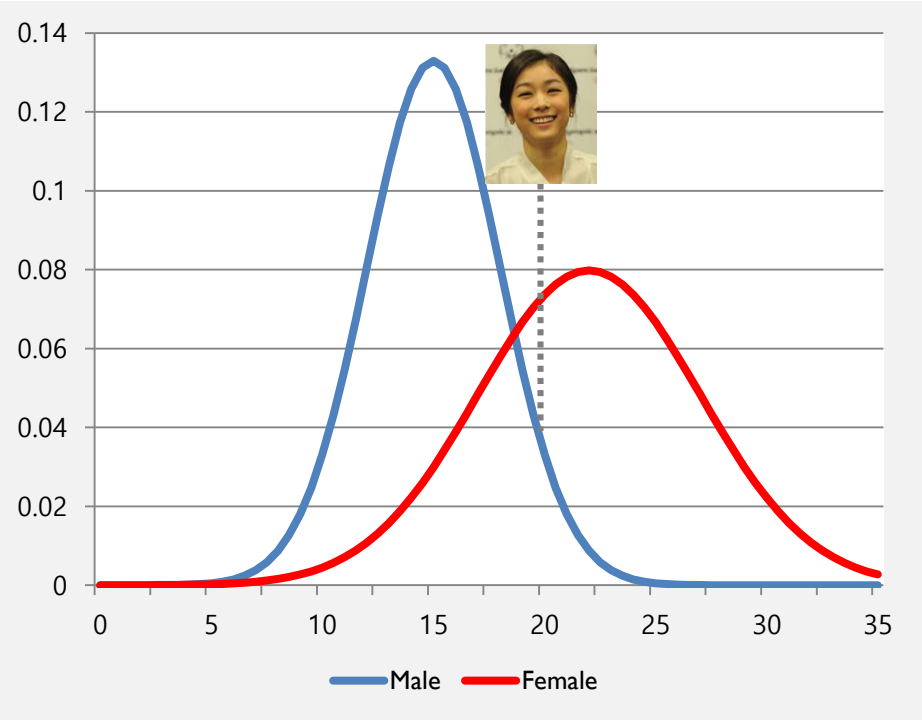


✓키로 보나 체지방률로 보나 남자일 가능성이 높음

Naïve Bayesian Classifier

- 김연아는 남자일까 여자일까?

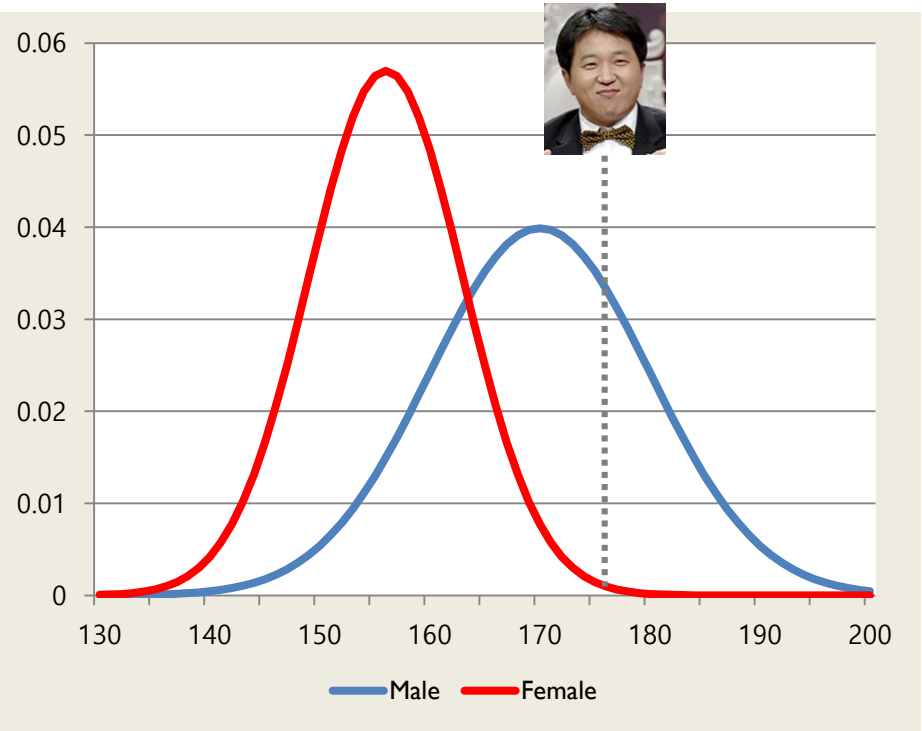


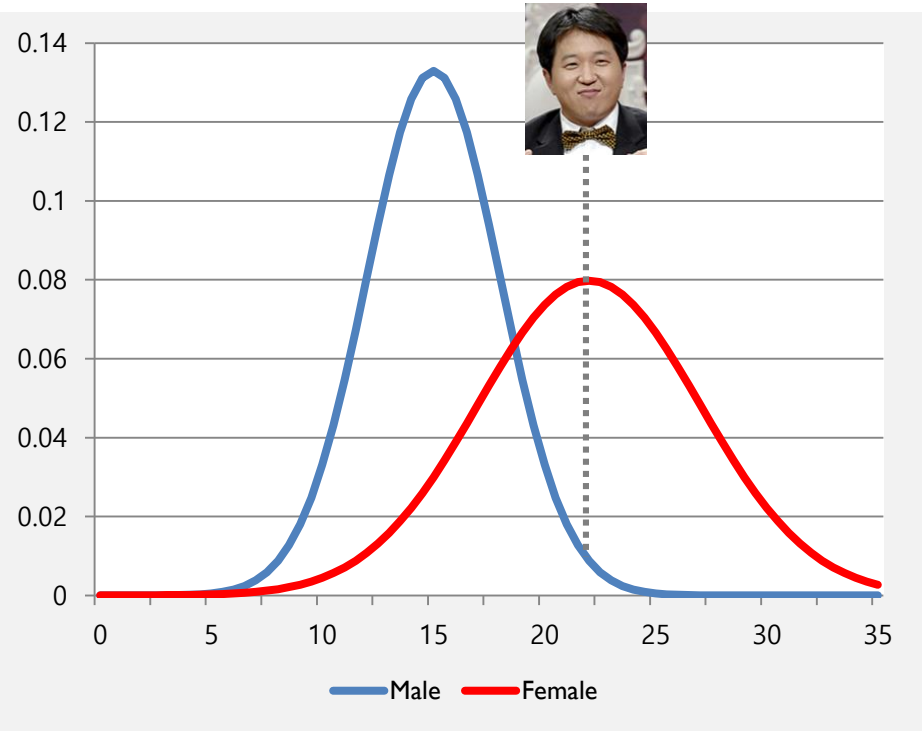


✓키로 보나 체지방률로 보나 여자일 확률이 높음

Naïve Bayesian Classifier

- 그렇다면 정형돈은?





✓키로 보면 남자인데 체지방률로 보면 여자… 어느 범주로 분류를 해야 하지???

Naïve Bayesian Classifier

- 베이즈 규칙(Baye's Rule)

Posterior probability (사후확률)

Likelihood (우도)

Prior probability (사전확률)

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A) \times P(A)}{P(B)}$$

Evidence (증거)

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{\frac{P(B,A)}{P(A)} \times P(A)}{P(B)}$$

Naïve Bayesian Classifier

- 변수가 두 개일 때의 Bayes's Rule

$$P(C_i|x_1, x_2) = \frac{P(x_1, x_2|C_i) \cdot P(C_i)}{P(x_1, x_2)}$$

- Naïve: 모든 변수들은 모두 통계적으로 독립Statistically Independent이라고 가정해보자

$$P(C_i|x_1, x_2) = \frac{P(x_1|C_i) \cdot P(x_2|C_i) \cdot P(C_i)}{P(x_1, x_2)}$$

Naïve Bayesian Classifier

- 이를 일반화 해보면…

$$P(C_i \mid x_1, x_2, ..., x_d) = \frac{P(x_1, x_2, ..., x_d \mid C_i) P(C_i)}{P(x)}$$ Baye's Rule

$$= \frac{\left(P(x_1 \mid C_i) \cdot P(x_2 \mid C_i) \cdot ... \cdot P(x_n \mid C_i)\right) P(C_i)}{P(x)}$$

Naïve: Variables are statistically independent!

Naïve Bayesian Classifier

- **분류 범주 예측**
  - ✓각 범주에 대한 사후 확률 계산

$$P(C_1 \mid x_1, x_2, ..., x_d) = \frac{\left(P(x_1 \mid C_1) \cdot P(x_2 \mid C_1) \cdot ... \cdot P(x_n \mid C_1)\right) P(C_1)}{P(x)}$$

$$P(C_2 \mid x_1, x_2, ..., x_d) = \frac{\left(P(x_1 \mid C_2) \cdot P(x_2 \mid C_2) \cdot ... \cdot P(x_n \mid C_2)\right) P(C_2)}{P(x)}$$

  - 위 두 식에서 분모인 P(x)가 같으므로 굳이 계산할 필요는 없음
  - 분자만 계산한 뒤 그 합으로 각각을 나눠주면 사후 확률 계산 가능

Naïve Bayesian Classifier

- **앞선 예시로 돌아가보면**
  - ✓우리가 구하고자 하는 것은 다음의 두 확률
    - P(정형돈 Height, 정형돈 BFP | Male)*P(Male) vs.
    - P(정형돈 Height, 정형돈 BFP | Female)*P(Female)
  - ✓만일 두 속성인 height 와 BFP가 통계적으로 독립이라는 가정을 할 수 있고 남자와 여자의 비율이 같다면
    - P(정형돈 Height, 정형돈 BFP | Male) *P(Male) =

    P(정형돈 Height | Male)* P(정형돈 BFP | Male) *P(Male) = 0.035*0.01*= 0.000175
    - P(정형돈 Height, 정형돈 BFP | Female) P(Female) =

    P(정형돈 Height | Female)* P(정형돈 BFP | Female)* P(Female) = 0.001*0.08*0.5 = 0.00004
  - ✓0.000175 > 0.00004 이므로 정형돈은 남자로 분류!

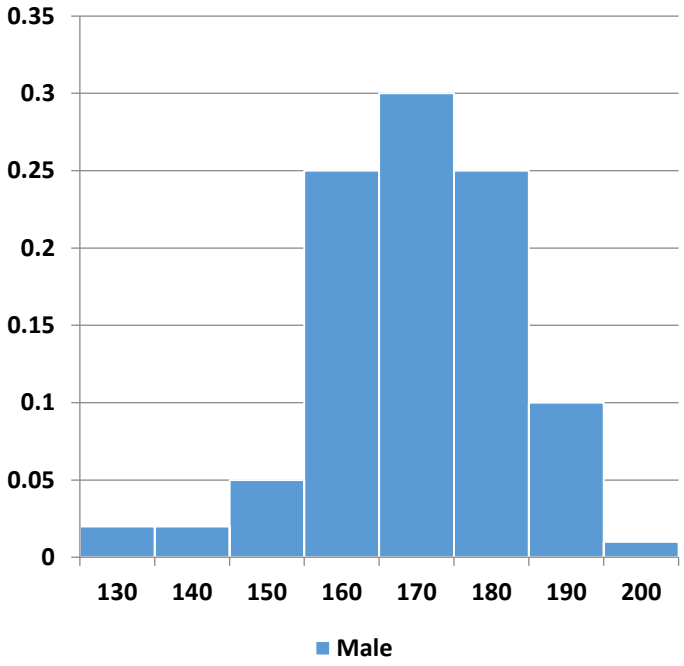
Naïve Bayesian Classifier

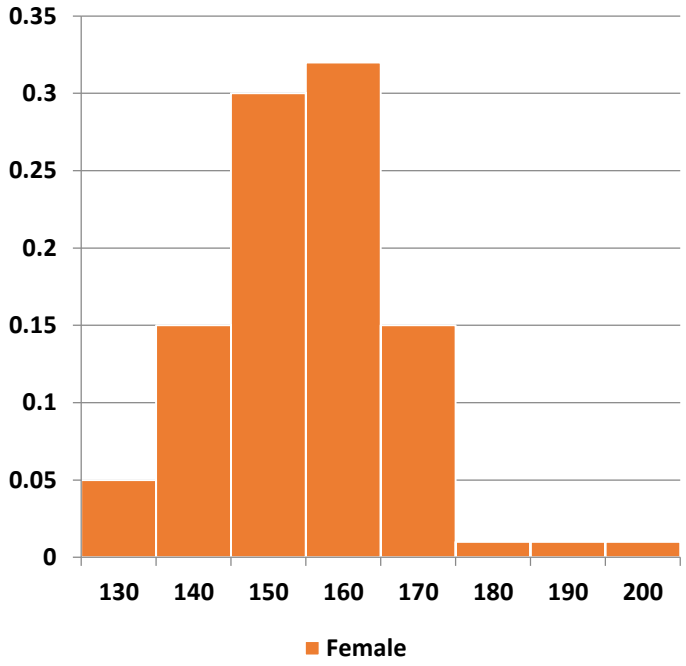
- Naïve Bayesian Classifier 절차
  - ✓ Step 1: 학습 데이터 준비
    - 설명변수를 정의하고 필요한 학습 데이터 수집
    - 학습 데이터 총 개체 수: 200 (남성 100명, 여성 100명)
    - 설명변수: 키(Height), 체지방률(BFS)

Record	Height	BFS	Class
1	187	15	M
2	165	25	F
3	174	14	M
4	156	29	F
…	…	…	…
N	168	12	M

Naïve Bayesian Classifier

- Naïve Bayesian Classifier 절차
  - ✓ Step 2: 범주-변수별 확률분포 추정
    - 각 범주의 모든 변수에 대해 확률 분포 추정 (즉, 이 예시에서는 총 4개의 확률분포 추정이 필요함)
    - 일반적으로 정규분포를 가정하는 경우가 더 많으나, 이해를 돕기 위해 히스토그램으로 추정

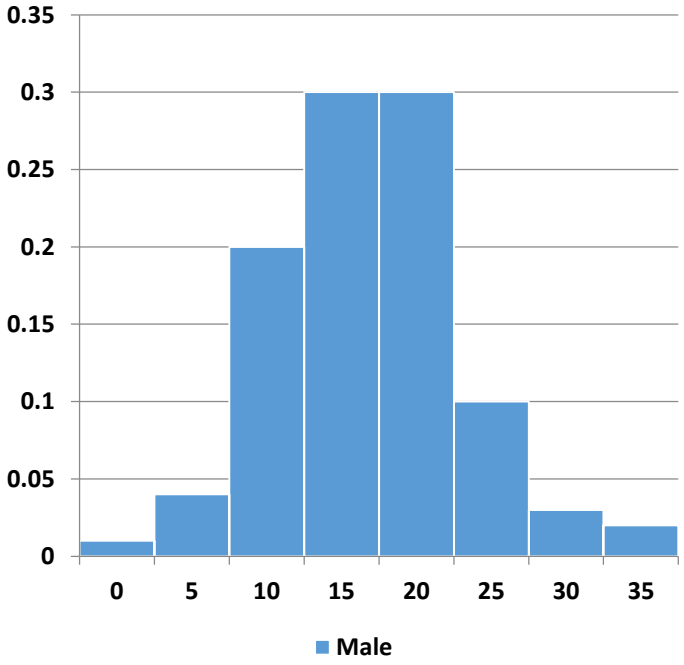


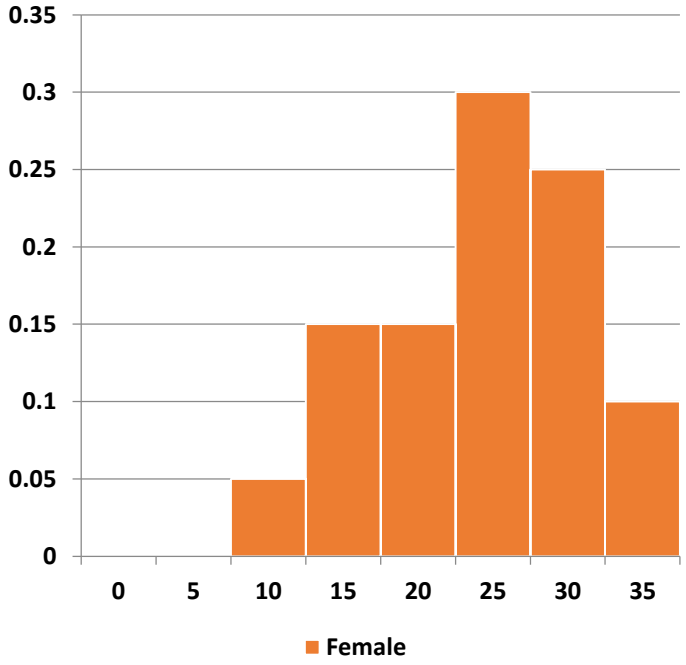


키

Naïve Bayesian Classifier

- Naïve Bayesian Classifier 절차
  - ✓ Step 2: 범주-변수별 확률분포 추정
    - 각 범주의 모든 변수에 대해 확률 분포 추정 (즉, 이 예시에서는 총 4개의 확률분포 추정이 필요함)
    - 일반적으로 정규분포를 가정하는 경우가 더 많으나, 이해를 돕기 위해 히스토그램으로 추정

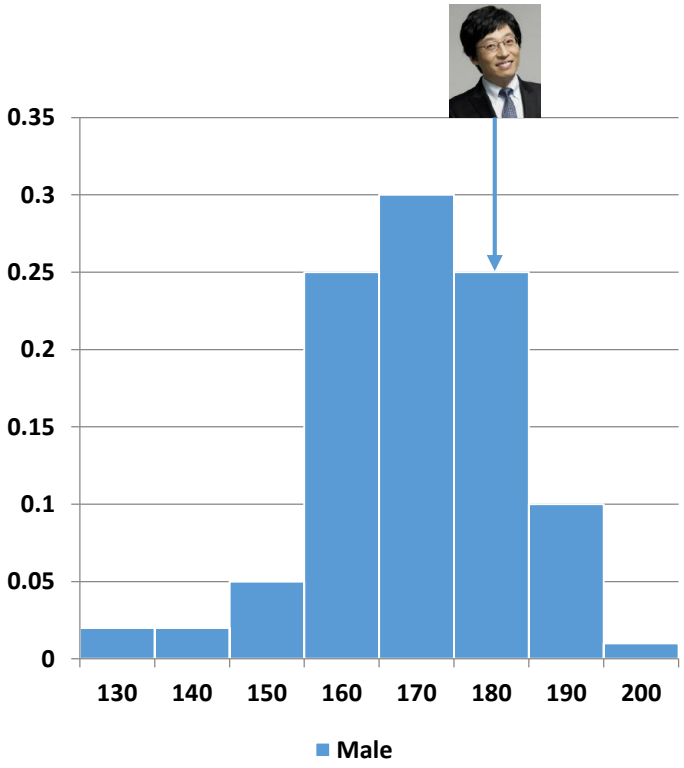


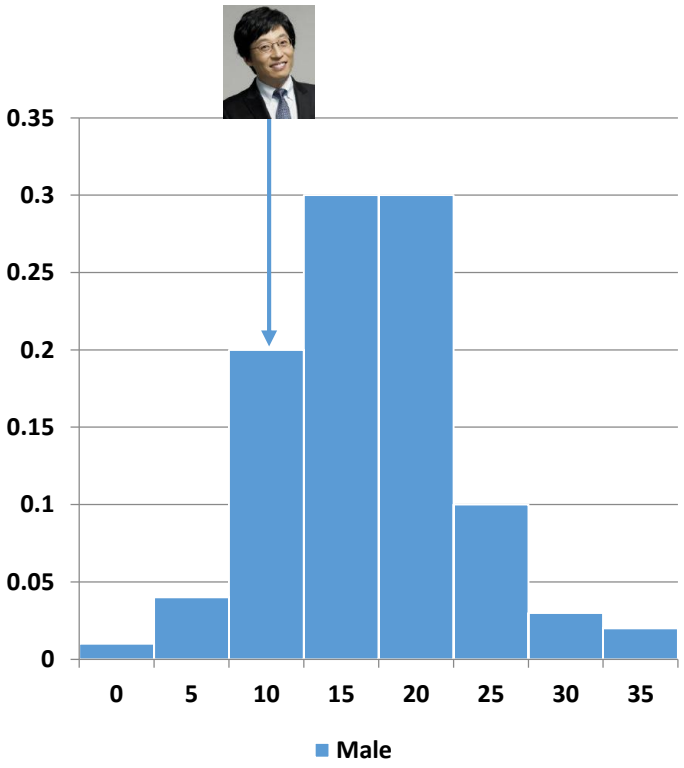


체지방률

Naïve Bayesian Classifier

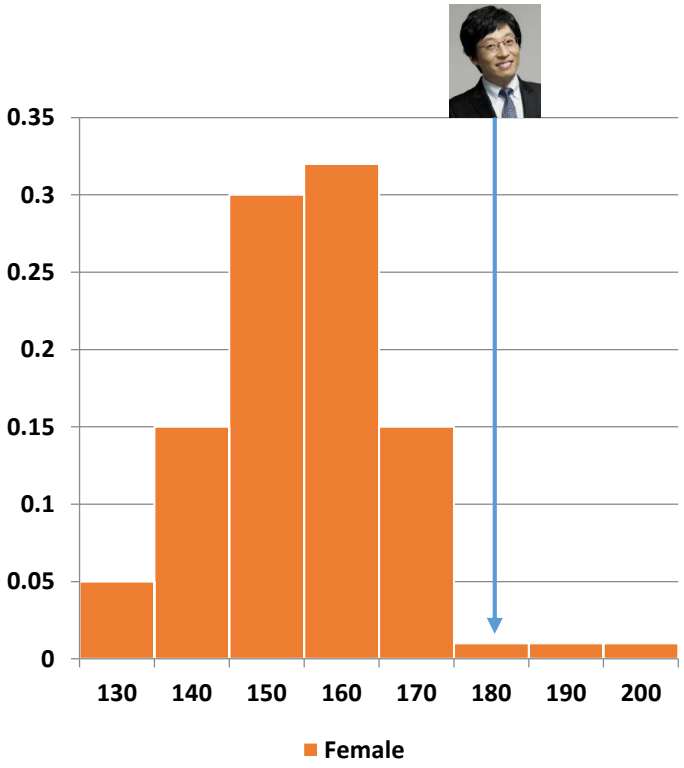
- Naïve Bayesian Classifier 절차
  - ✓ Step 3: 각 변수에 대한 조건부 확률 추정
    - P(Height = 178 | Male) = 0.25, P(BFS = 11 | Male) = 0.2

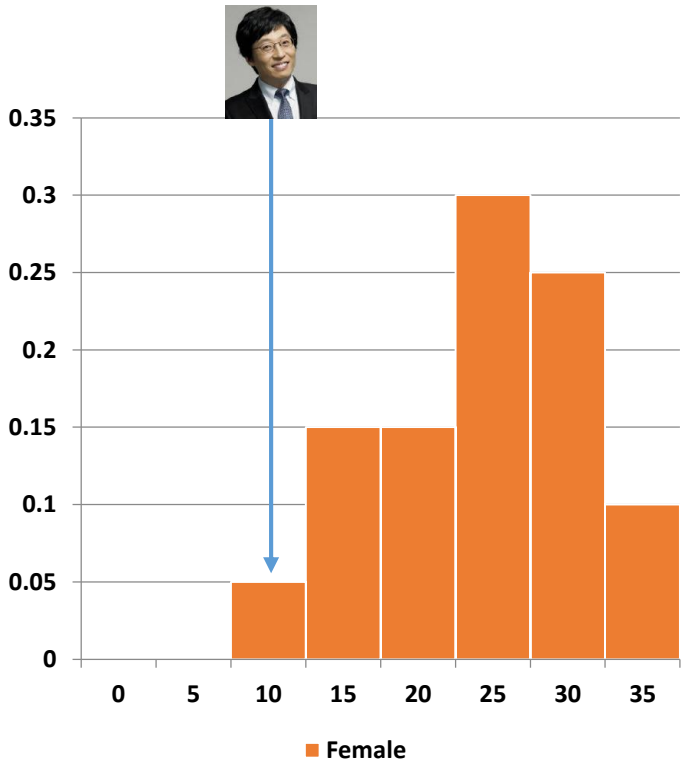




Naïve Bayesian Classifier

- Naïve Bayesian Classifier 절차
  - ✓ Step 3: 각 변수에 대한 조건부 확률 추정
    - P(Height = 178 | Female) = 0.01, P(BFS = 11 | Female) = 0.05





Naïve Bayesian Classifier

- Naïve Bayesian Classifier 절차
  - ✓ Step 4: 각 범주에 대한 사후 확률 추정 및 범주 분류

    - P(Height = 178, BFS = 11 | Male) * P(Male)

    = P(Height = 178 | Male)* P(BFS = 11 | Male) * P(Male)

    = 0.25*0.2*0.5 = 0.025

    - P(Height = 178, BFS = 11 | Female) * P(Female)

    = P(Height = 178 | Female)* P(BFS = 11 | Female) * P(Female)

    = 0.01*0.05*0.5 = 0.00025

    남성일 확률 = 99% (0.025/(0.025+0.00025))

    여성일 확률 = 1% (0.00025/(0.025+0.00025))

    남성으로 분류

Naïve Bayesian Classifier

- Naïve Bayesian Classifier 절차
  - ✓ Step 4: 각 범주에 대한 사후 확률 추정 및 범주 분류
    - 만약 학습 데이터가 100명의 남성과 400명의 여성으로 구성되어 있다면?
    - P(Height = 178, BFS = 11 | Male) * P(Male)

= P(Height = 178 | Male)* P(BFS = 11 | Male) * P(Male)

= 0.25*0.2*0.2 = 0.01

    - P(Height = 178, BFS = 11 | Female) * P(Female)

= P(Height = 178 | Female)* P(BFS = 11 | Female) * P(Female)

= 0.01*0.05*0.8 = 0.0004

남성일 확률 = 96% (0.01/(0.01+0.0004))

여성일 확률 = 4% (0.0004/(0.01+0.0004))

그래도 남성으로 분류


ANY
questions?