

Week 11 • 소셜네트워크 데이터마이닝과 분석

Advanced Text Analysis Topics & Machine Learning 1

Joonhwan Lee
human-computer interaction + design lab.

오늘 다룰 내용

- Machine Learning의 기초
- Supervised Learning
- Unsupervised Learning

1. Machine Learning의 기초

Machine Learning

- ❖ 데이터를 기반으로 일어나지 않은 사실을 예측하거나 데이터를 분류하는 방법
- ❖ 데이터로부터 각종 패턴을 학습한다
- ❖ Rule Based vs Machine Learning
 - ❖ Rule Based
 - ❖ 컴퓨터에 여러 조건을 제시하고 해당되는 사건이 발생할 때 데이터를 처리
 - ❖ 수많은 if-else 문으로 구성
 - ❖ 조건문으로 제시되지 않은 경우는 처리할 수가 없다
 - ❖ Machine Learning
 - ❖ 기존의 데이터를 기반으로 (항상 그렇지는 않지만) 패턴을 학습
 - ❖ 새로운 데이터가 학습된 패턴에 해당될 확률을 계산

Machine Learning의 유형

- ❖ Supervised Learning (지도학습)
 - ❖ 정답이 있는 데이터셋의 학습을 통해 새로 수집된 데이터셋의 정답을 맞춘다.
 - ❖ 예: 스팸필터, 집값 예측
 - ❖ Prediction using Regression
- ❖ Unsupervised Learning (자율학습)
 - ❖ 정답이 없는 데이터 더미에서 패턴을 찾아 그룹을 만든다.
 - ❖ Clustering

2. Supervised Learning

통계로 튀기는 치킨

https://blog.naver.com/hi_nso/220489542903

가중치	1.8	1.42	1.73	2.5	1.34	1.26	예상량	실제량
3월	개절	날씨	이벤트1	이벤트2	이벤트3	이벤트4		
1여름							23	25
2여름	비				야구	42	45	
3여름		주말			야구	51	58	
4여름	비	주말			야구	72	66	
5여름	비					33	28	
6여름	비					33	32	
7여름						23	19	
8여름						23	21	
9여름						23	29	
10여름		주말				40	43	
11여름		주말				40	35	
12여름			말복			59	71	
13여름					야구	29	23	
14여름					야구	29	26	
15여름						23	16	
16여름	비					33	25	
17여름		주말			야구	51	43	
18여름	비	주말			야구	72	48	
19여름	비			기학열		45	37	
20여름				기학열		31	31	
21여름						23	27	
22여름					야구	29	19	

통계로 튀기는 치킨

https://blog.naver.com/hi_nso/220489542903

Features

가중치	1.8	1.42	1.73	2.5	1.34	1.26	0	상량	실제량
3월	개절	날씨	이벤트1	이벤트2	이벤트3	이벤트4			
1일							23	25	
2여름	비				야구		42	45	
3여름		주말			야구		51	58	
4여름	비	주말			야구		72	66	
5여름	비						33	28	
6여름	비						33	32	
7여름							23	19	
8여름							23	21	
9여름							23	29	
10여름		주말					40	43	
11여름		주말					40	35	
12여름			말복				59	71	
13여름					야구		29	23	
14여름					야구		29	26	
15여름							23	16	
16여름	비						33	25	
17여름		주말			야구		51	43	
18여름	비	주말			야구		72	48	
19여름	비			기학열			45	37	
20여름				기학열			31	31	
21여름							23	27	
22여름					야구		29	19	

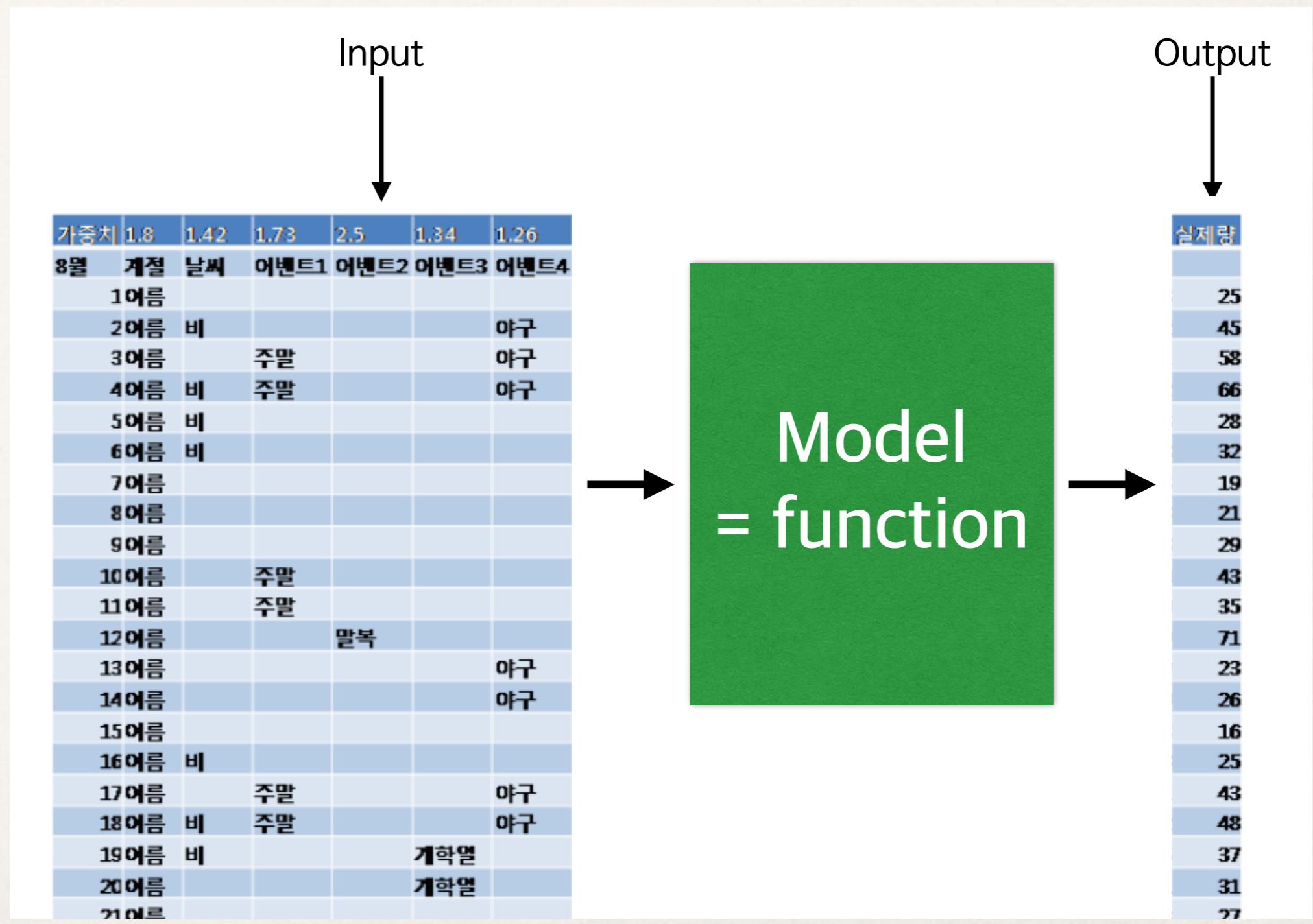
통계로 튀기는 치킨

<https://blog.naver.com/hinso/220489542903>

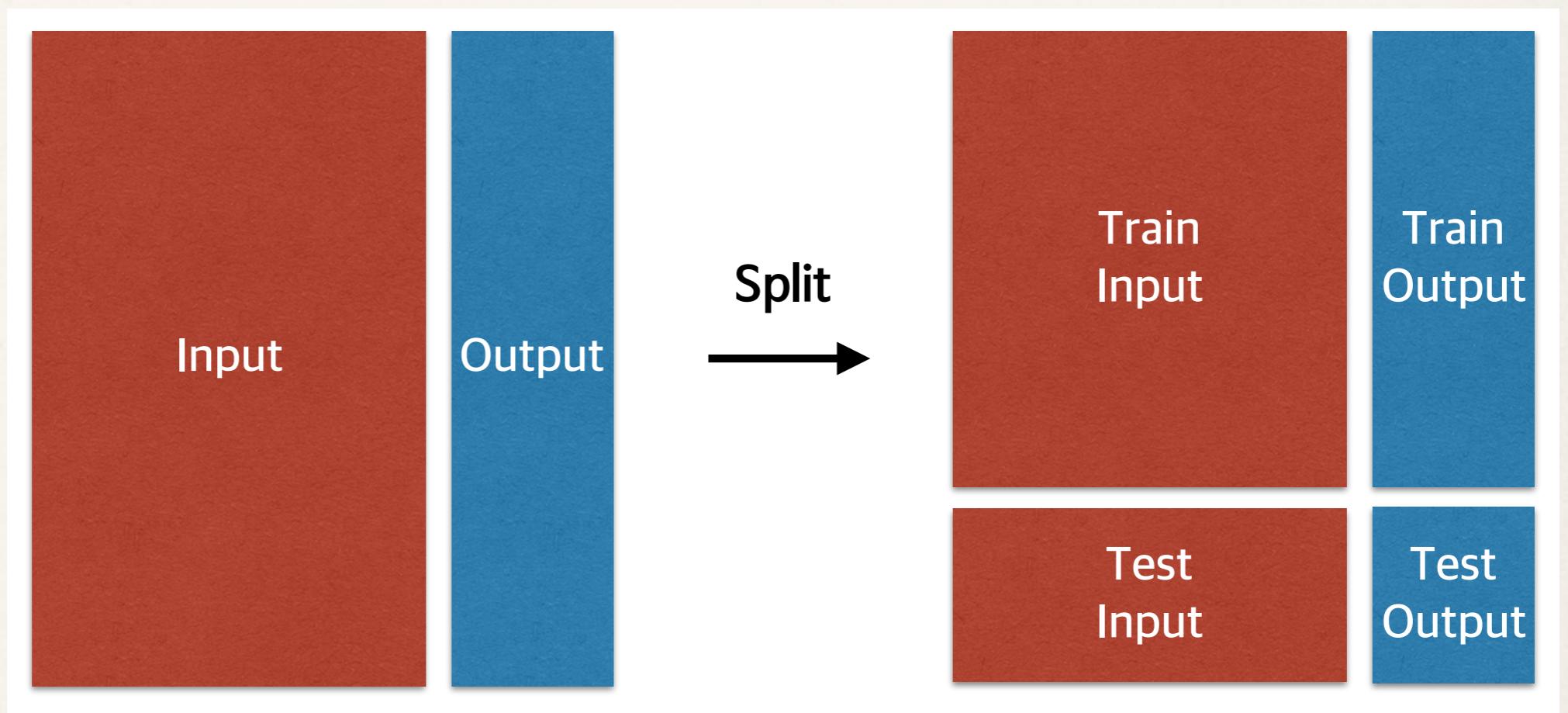
Features Output

가중치	1.8	1.42	1.73	2.5	1.34	1.26	0.3	상당	실제량
3월	개별	날씨	이벤트1	이벤트2	이벤트3	이벤트4			
1일							23	25	
2여름	비					야구	42	45	
3여름		주말				야구	51	58	
4여름	비	주말				야구	72	66	
5여름	비						33	28	
6여름	비						33	32	
7여름							23	19	
8여름							23	21	
9여름							23	29	
10여름		주말					40	43	
11여름		주말					40	35	
12여름			말복				59	71	
13여름					야구		29	23	
14여름					야구		29	26	
15여름							23	16	
16여름	비						33	25	
17여름		주말				야구	51	43	
18여름	비	주말				야구	72	48	
19여름	비				기학열		45	37	
20여름					기학열		31	31	
21여름							23	27	
22여름						야구	29	19	

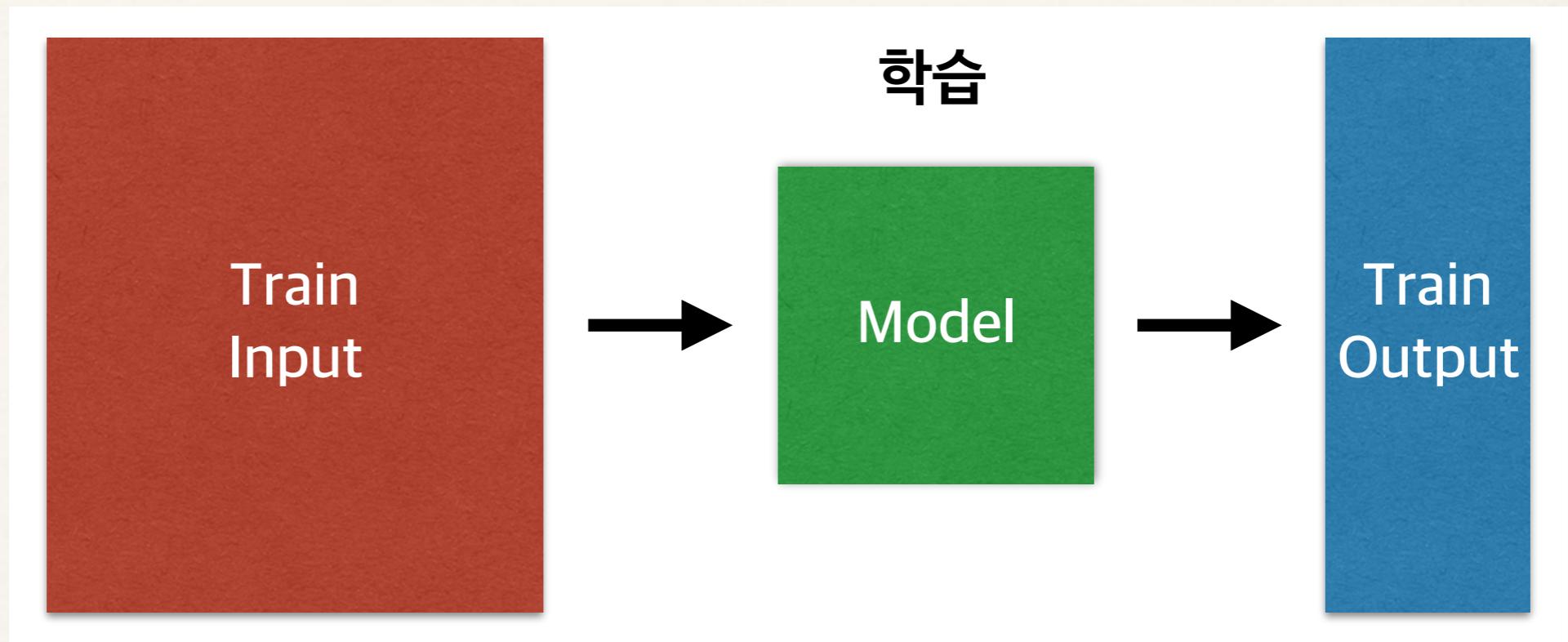
Supervised Learning - Model Training



Supervised Learning - Train / Test Set

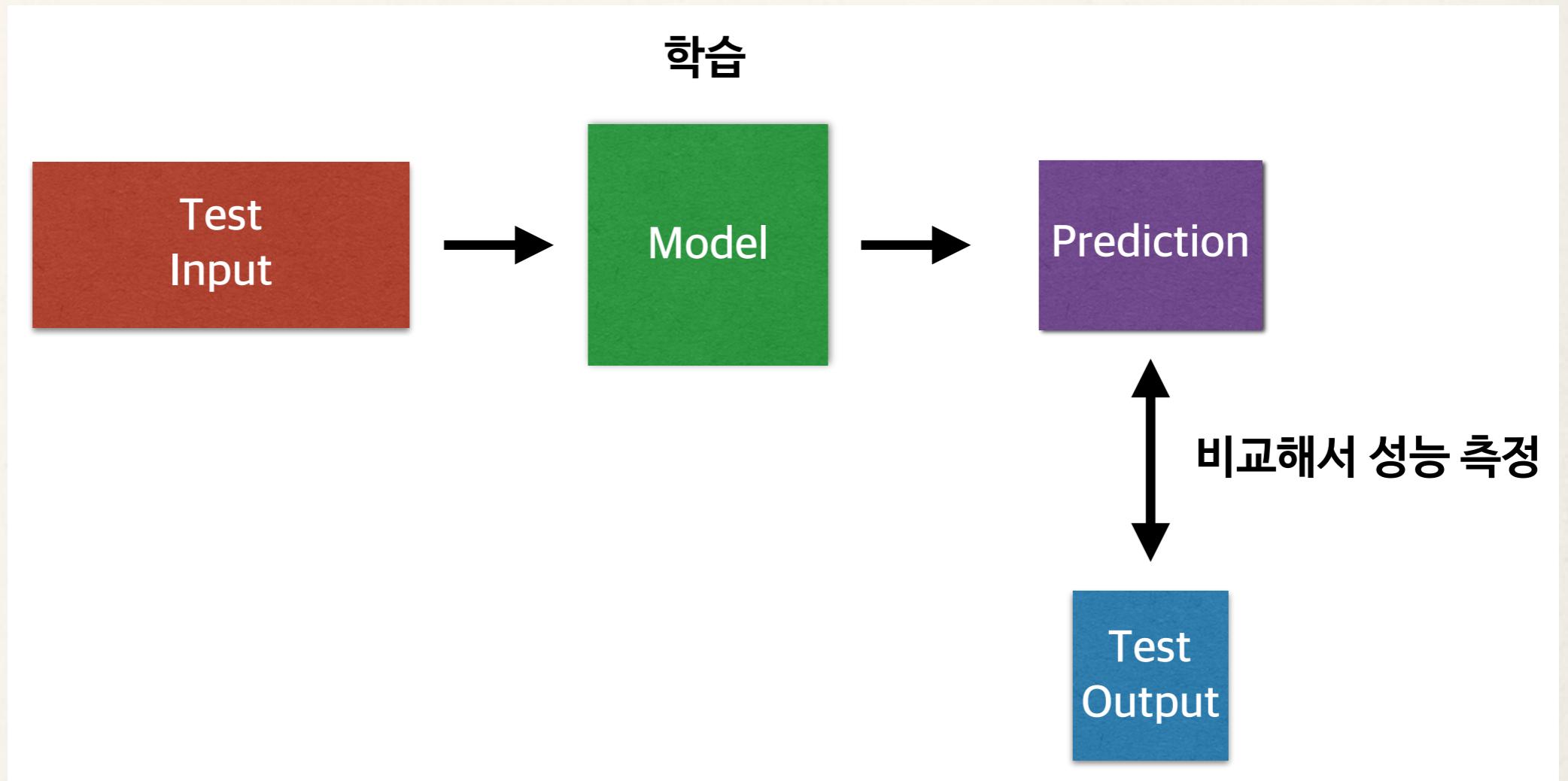


Supervised Learning - Train / Test Set



Training Set으로 모델을 학습

Supervised Learning - Train / Test Set



실제 데이터를 통해 얼마나 학습을 잘 했는지 판단한다.

Supervised Learning - Linear Regression

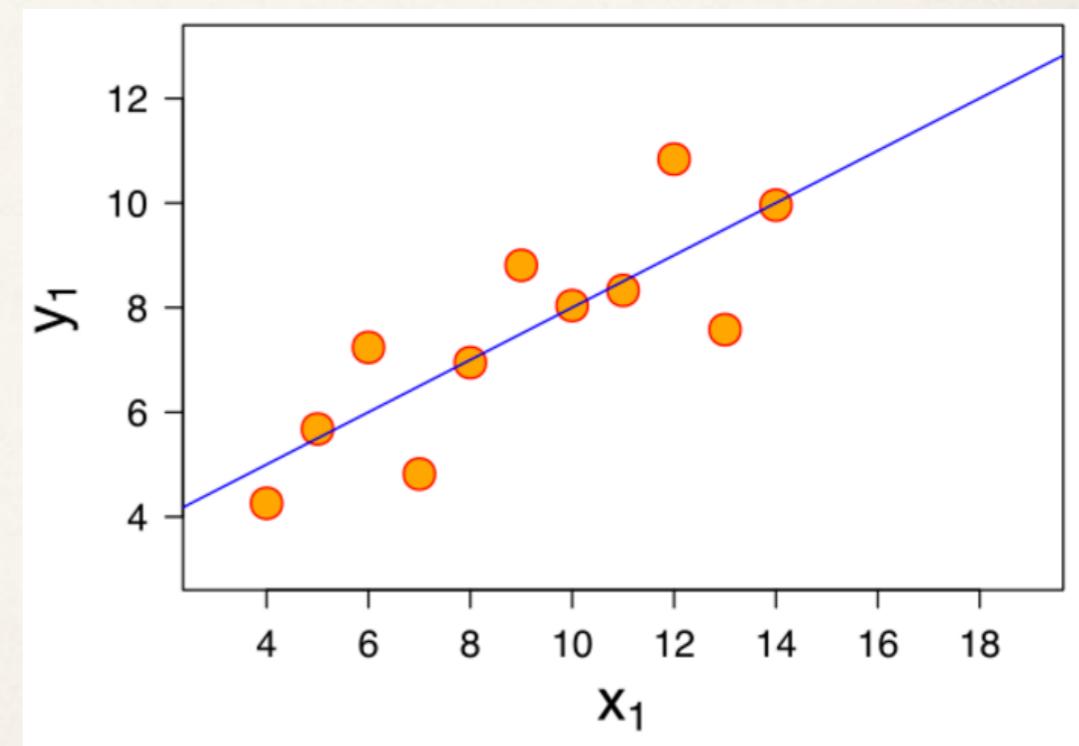
- ◆ 지도 학습은 크게 두 종류로 구분
 - ◆ Regression : 실수 범위의 연속적인 값을 예측
(예: 내일의 기온, 주가 등)
 - ◆ Classification : 데이터의 종류를 구분
(예: 사진의 개/고양이 구분)

- ◆ Linear Regression
(선형 회귀 분석)

- ◆ 가장 단순한 regression model

$$y = ax + b$$

$$y = a_1x_1 + a_2x_2 + \dots + a_nx_n$$

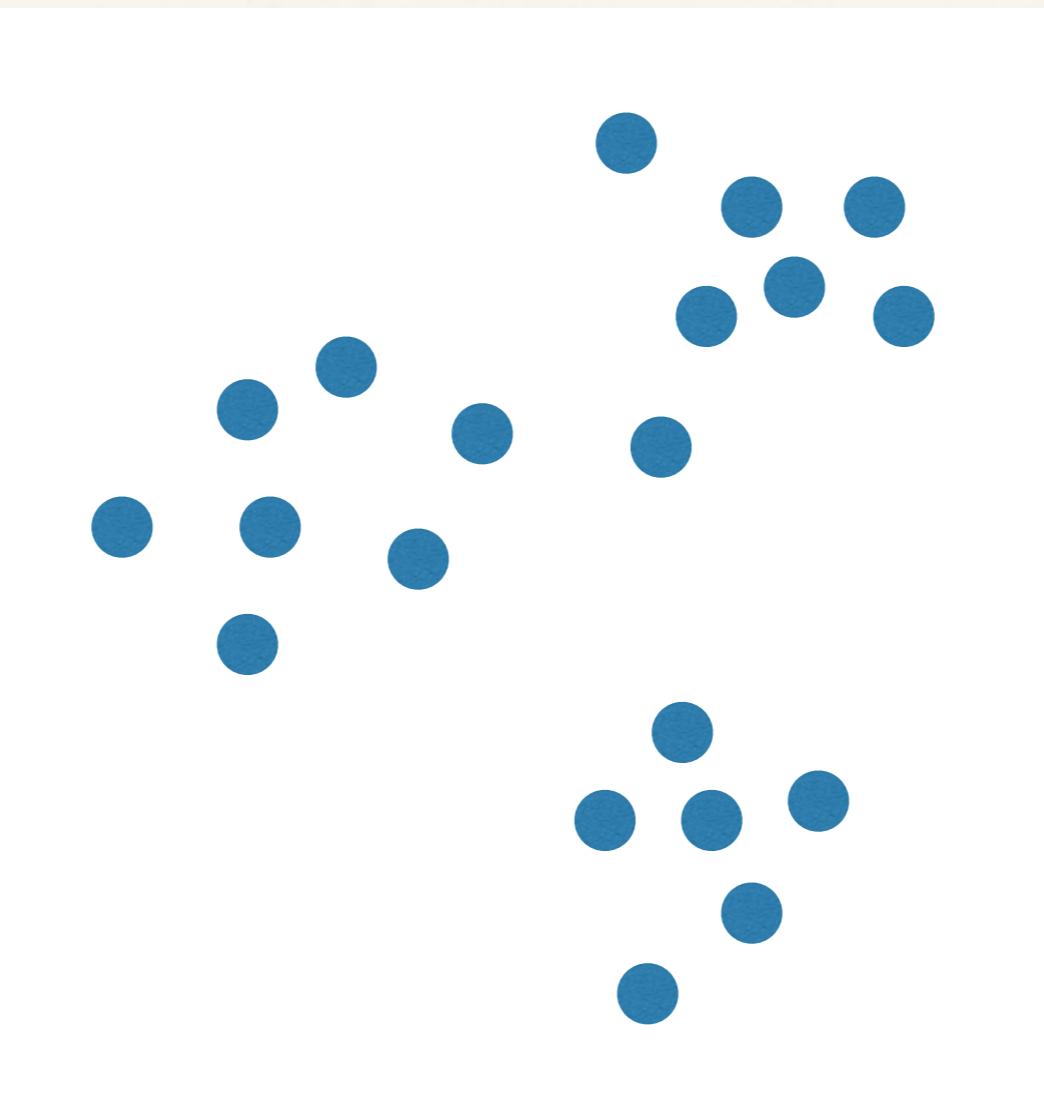


3. Unsupervised Learning

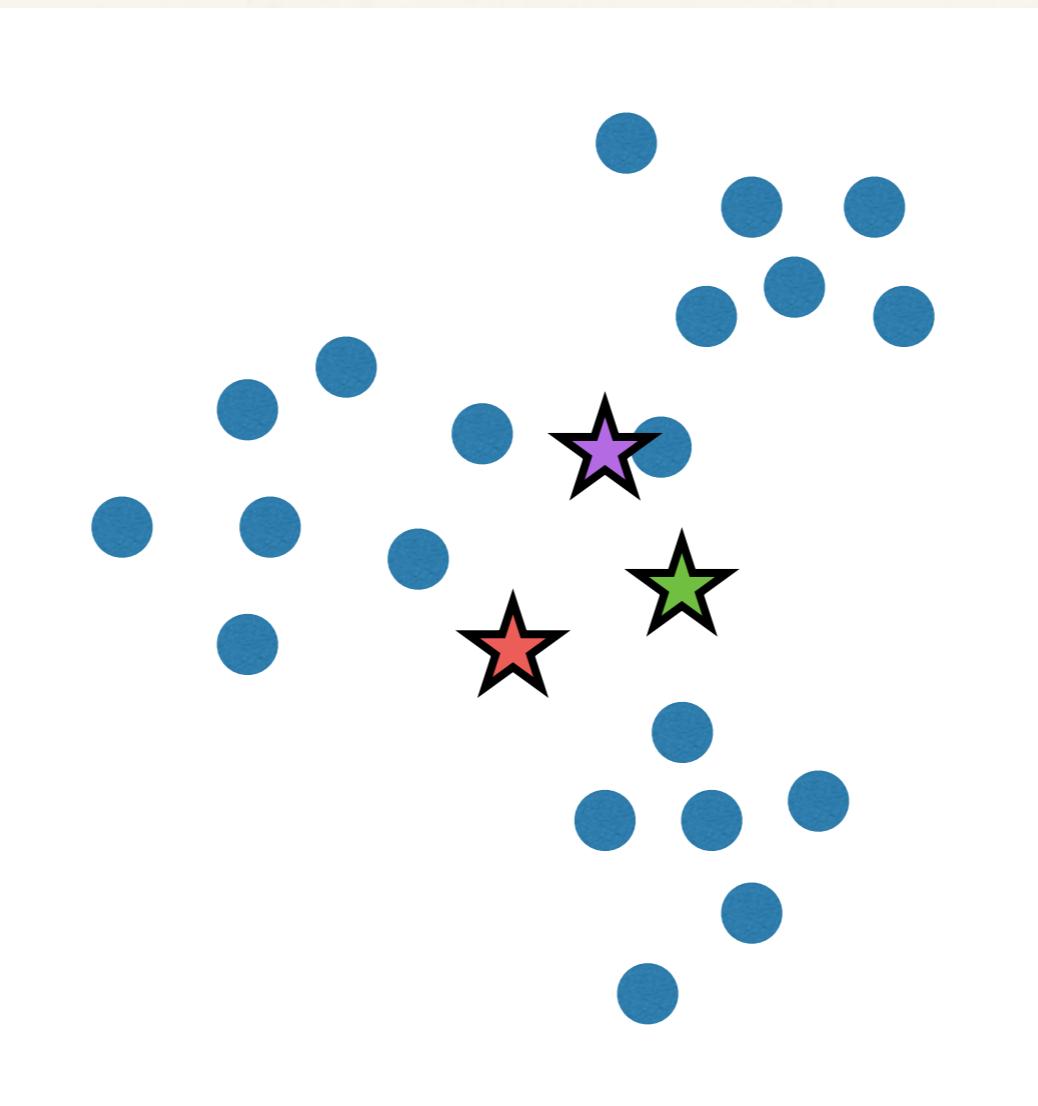
Unsupervised Learning - K-Means Clustering

- ❖ K 개의 클러스터를 찾는 알고리즘
- ❖ K 개의 ‘클러스터 중심점’ 을 데이터 공간에 뿌린다.
 - ❖ 1. 각 중심점과 가까운 데이터 점들을 해당 중심점의 클러스터로 할당한다.
 - ❖ 2. 각 클러스터의 데이터 점들을 각각 평균내서 새로운 중심점을 찾는다.
- ❖ 변화가 없을 때까지 1, 2를 반복한다.

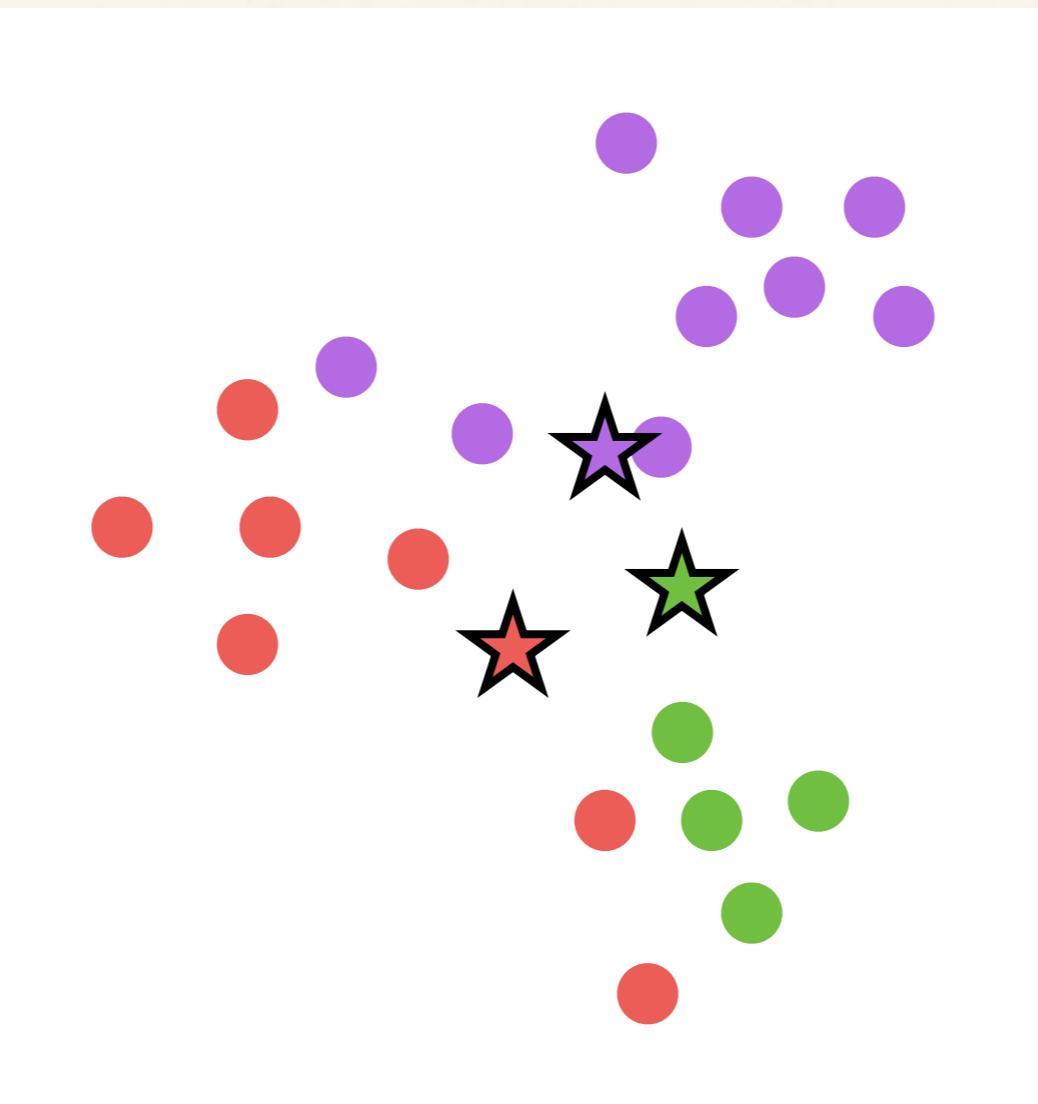
Unsupervised Learning - K-Means Clustering



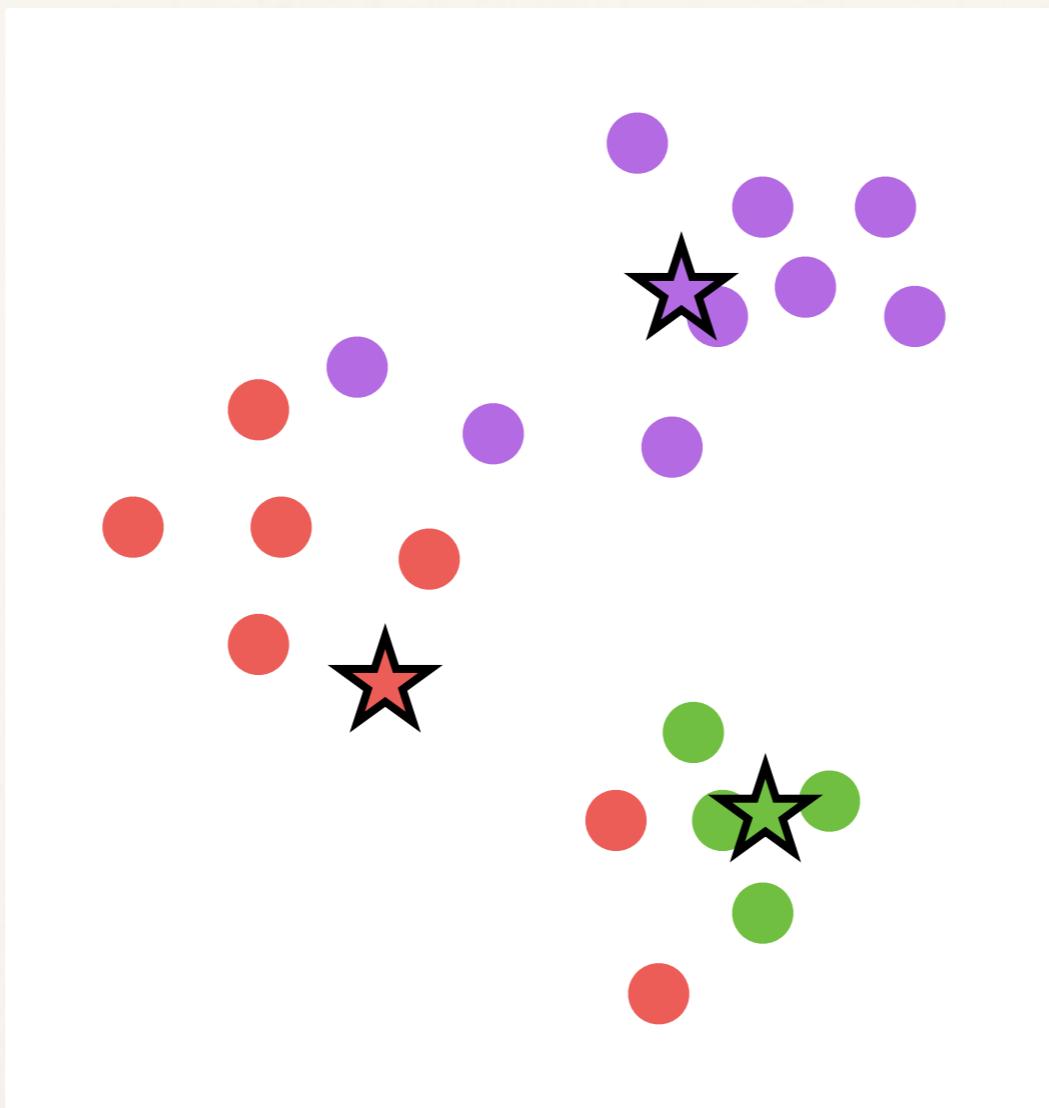
Unsupervised Learning - K-Means Clustering



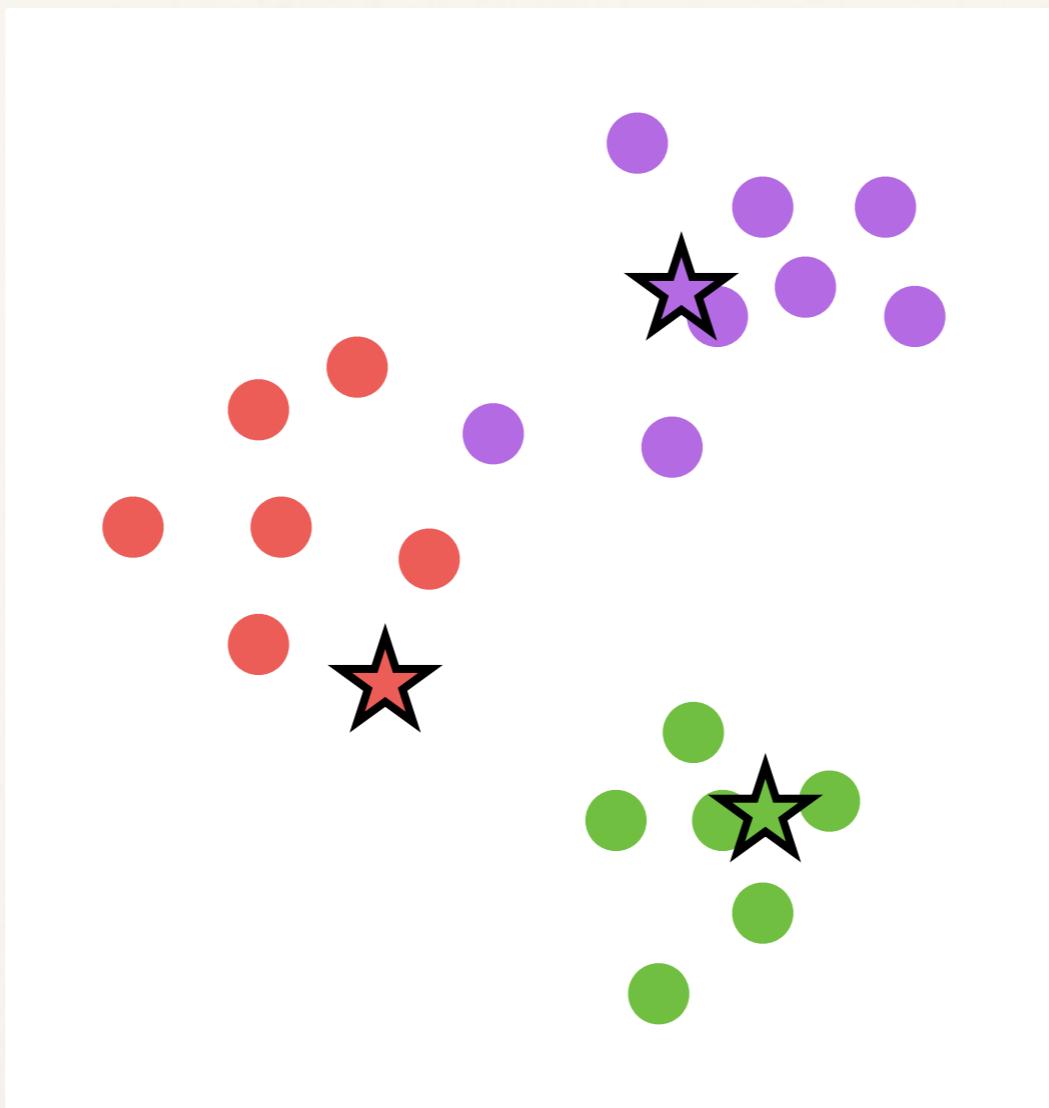
Unsupervised Learning - K-Means Clustering



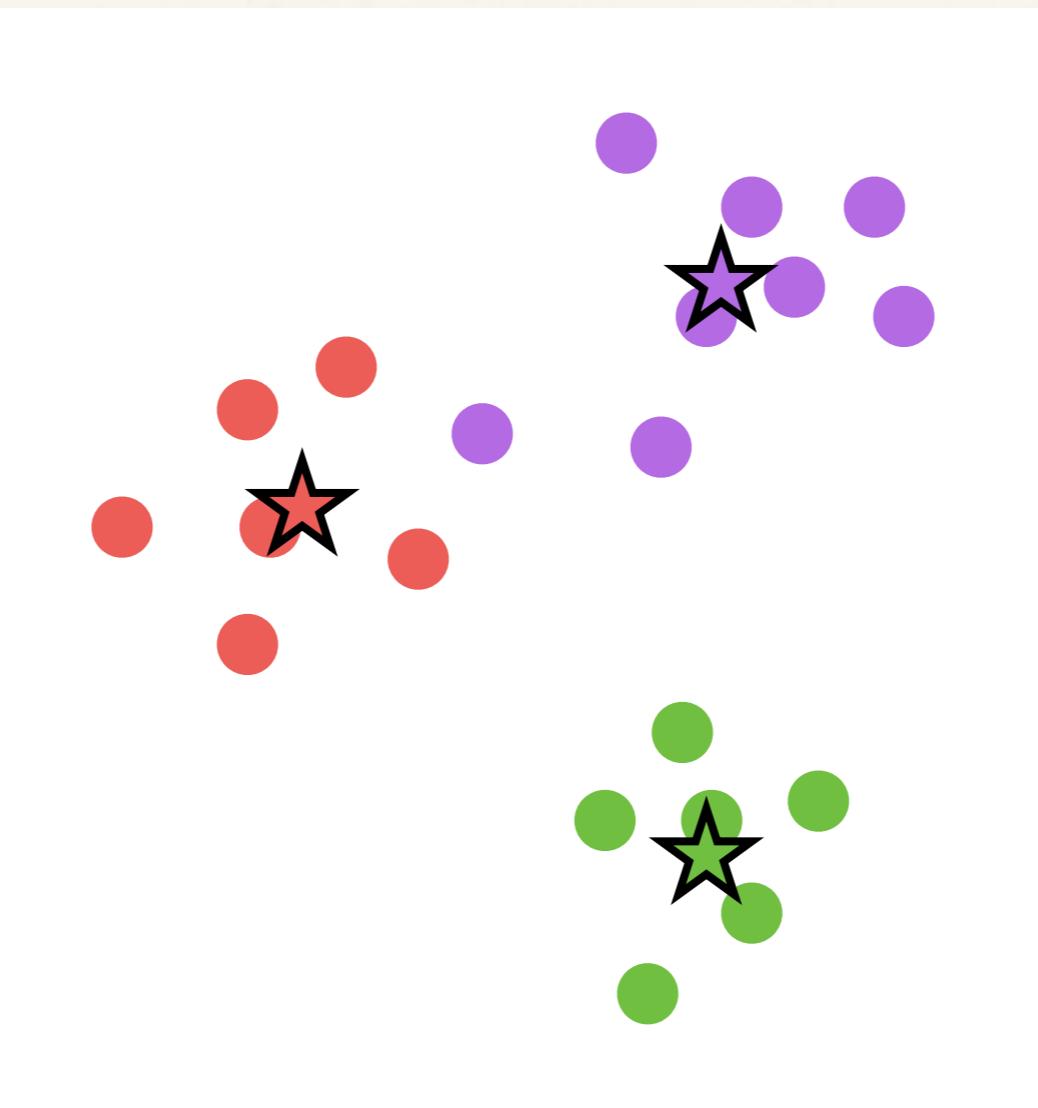
Unsupervised Learning - K-Means Clustering



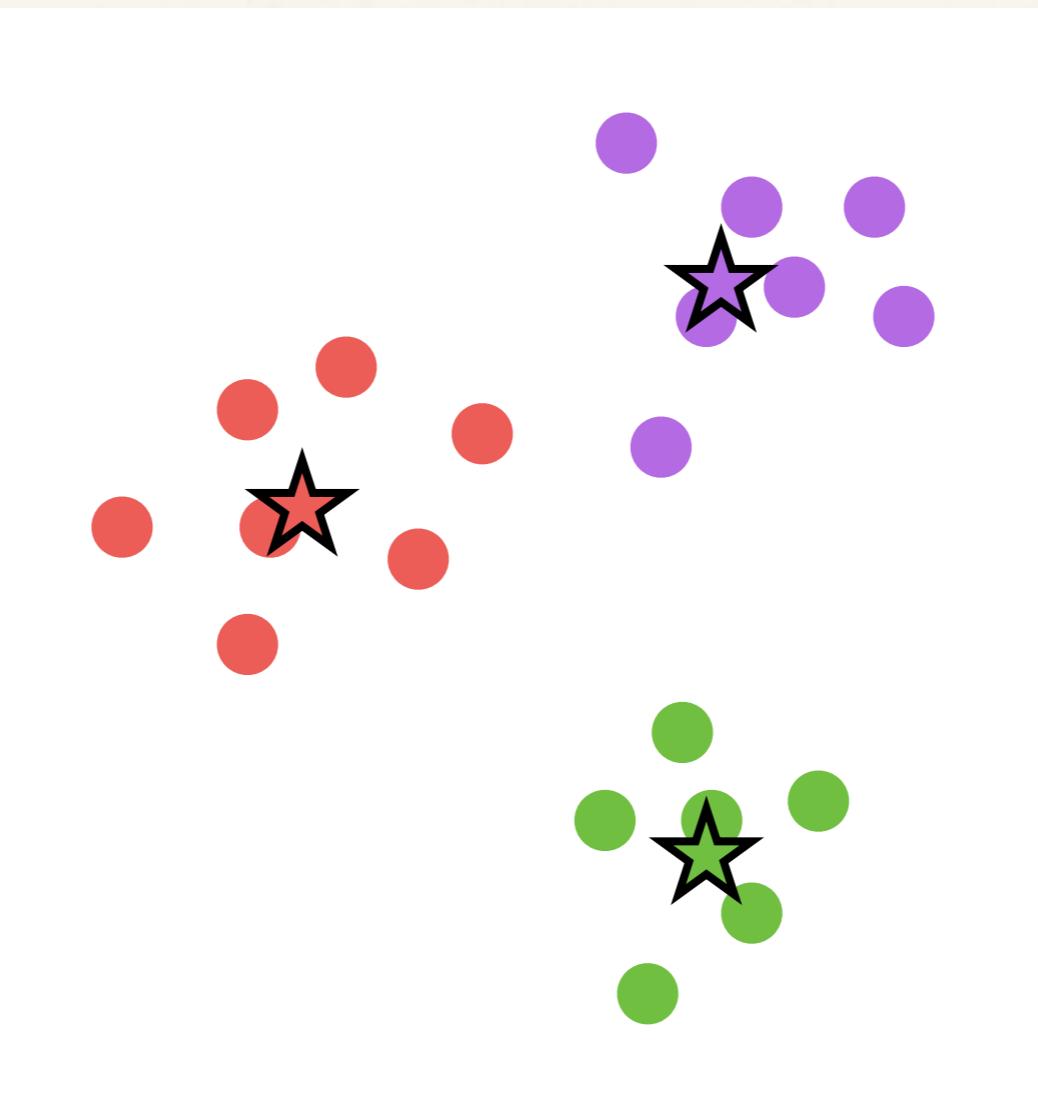
Unsupervised Learning - K-Means Clustering



Unsupervised Learning - K-Means Clustering



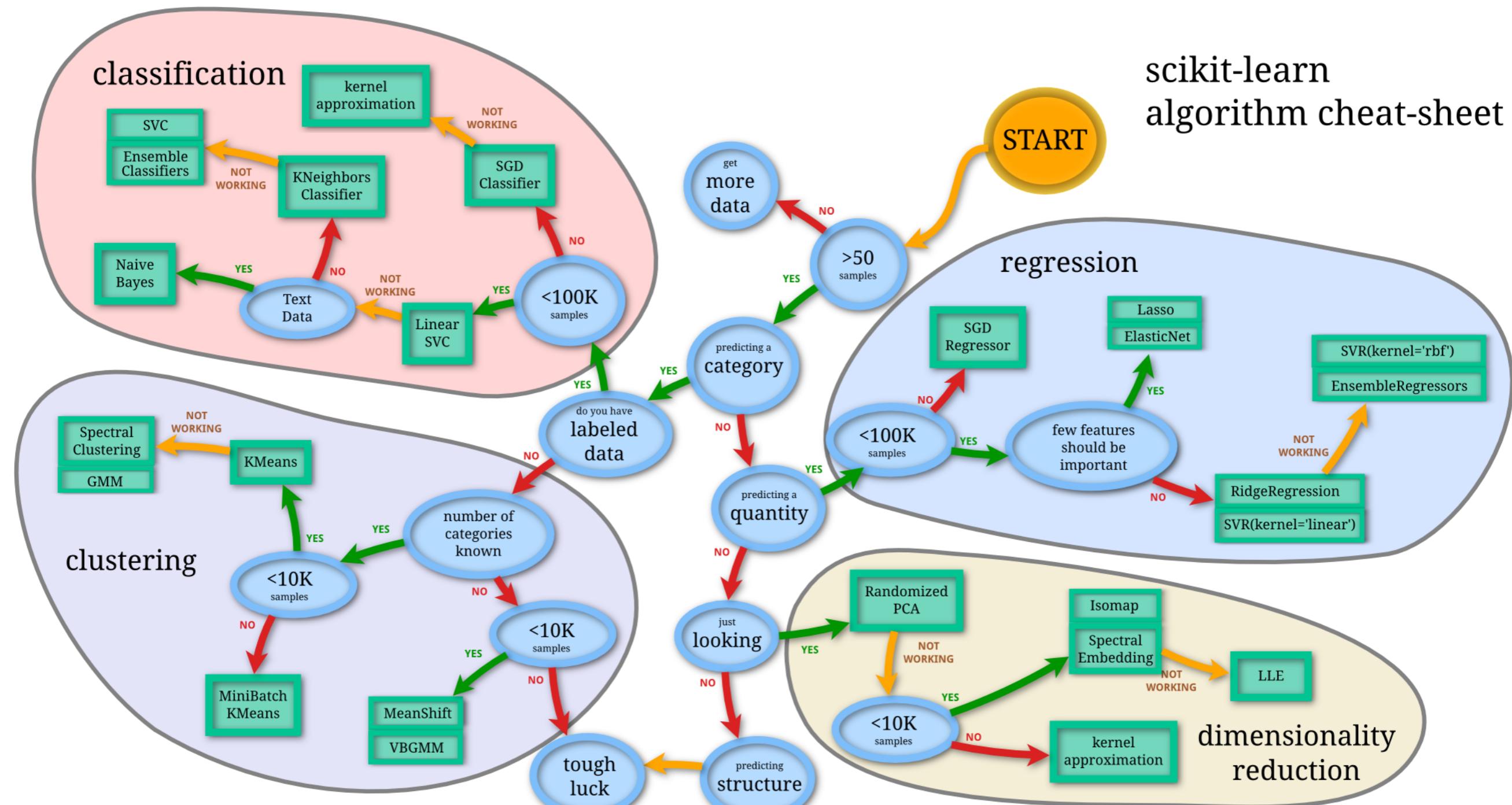
Unsupervised Learning - K-Means Clustering



Unsupervised Learning - K-Means Clustering

- ◆ 와인 성분/속성 데이터
 - ◆ 178종류의 와인
 - ◆ 13종류의 성분/속성
 - ◆ 알코올 도수
 - ◆ 사과산 농도
 - ◆ 페놀
 - ◆ 색깔
 - ◆ 채도
 - ◆ ...

scikit-learn algorithm cheat-sheet



Back

scikit
learn

Text Data Vectorization

Text Data Vectorization

- ◆ 기계학습은 데이터의 매트릭스 계산 등을 통해 이루어짐.
- ◆ 그러나 우리가 수집하는 텍스트 데이터 (string)은 수학적 연산을 할 수 있는 형식의 데이터가 아님.
- ◆ 따라서 텍스트 데이터를 수치형 데이터로 변환하는 과정이 필요.

Bag of words

- ◆ 자연어 처리에서 기본적으로 사용되는 워드벡터 모델.
- ◆ 단어가 문장에서 출현하는 순서 등을 고려하지 않은채 리스트에 넣어 처리한다.
 - ◆ John is quicker than Mary.
 - ◆ Mary is quicker than John.

=> [John , is , Mary , quicker , than]

TF-IDF

- ❖ 어떤 ‘단어’가 특정 ‘문서’에서 얼마나 중요한지를 나타내는 가중치
 - ❖ TF(Term Frequency): 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값
“**red** cars and **red** trucks are...” = (0, 1, 0, 1, **2**, 0, ...)
 - ❖ IDF(Inverse Document Frequency): 한 단어가 문서 집합 전체에서 얼마나 공통적으로 나타나는지를 나타내는 값
“arachnocentric universe”

Term Frequency (tf)

- ❖ 문서에서 해당 단어가 얼마나 나타났는가?를 의미한다. 예를 들어 문서에 “강아지”라는 단어가 10번 나오면 TF값은 10이 된다.

Document Frequency (df)

- ♦ DF는 전체 문서들에서 몇개의 문서에 해당 단어가 나타나 있는지에 대한 값이다.
- ♦ DF = 해당 단어가 나타난 문서수 / 전체 문서 수

Inverse Document Frequency (idf)

- ♦ IDF는 DF의 역수이다. 가장 DF가 큰 값을 1로 만들기 위해 IDF를 사용한다.
- ♦ $IDF = \text{전체 문서 수} / \text{해당 단어가 나타난 문서 수}$
- ♦ 그러나 보통 값이 크게 나타나기 때문에 log를 씌운다.
- ♦ $IDF = \log(\text{전체 문서 수} / \text{해당 단어가 나타난 문서수})$

TF-IDF

- ◆ TF값과 IDF값을 곱한 값이다.
- ◆ 이 값이 높을 수록 해당 문서에서 자주 등장한다는 뜻이고, 다른 문서에서 등장하면 단어의 중요성이 하락한다는 뜻이 된다.
- ◆ 예:
 - docA = "The cat sat on my face"
 - docB = "The dog sat on my bed"

TF-IDF 계산의 예

<https://medium.com/@nsh235482/tf-idf-term-frequency-inverse-document-frequency-algorithm-55f64714880d>

- ♦ 주어진 문장
 - ♦ Tom plays soccer. (Doc1)
 - ♦ Tom loves soccer and baseball. (Doc2)
 - ♦ baseball is his hobby and his job. (Doc3)
- ♦ BoW => [Tom, plays, soccer, loves, and, baseball, is, his, hobby, job]

TF-IDF 계산의 예

<https://medium.com/@nsh235482/tf-idf-term-frequency-inverse-document-frequency-algorithm-55f64714880d>

- ♦ TF 계산

- ♦ 각 문서에 출현한 단어의 숫자를 카운트

	Tom	plays	soccer	loves	and	baseball	is	his	hobby	job
Doc1	1	1	1							
Doc2	1		1	1	1	1	1			
Doc3						1	1	1	2	1

TF-IDF 계산의 예

<https://medium.com/@nsh235482/tf-idf-term-frequency-inverse-document-frequency-algorithm-55f64714880d>

- ♦ IDF 계산

- ♦ IDF값은 $\log(\text{전체 문서 수} / \text{해당 단어가 나타난 문서수})$
- ♦ Tom을 계산해보면 해당 단어가 나타난 문서수는 2이고 현재 전체 문서의 수는 3이므로 $\log(3/2) \approx 0.18$
- ♦ 즉, 한 단어가 여러 문서에 출현할수록 한 문서에서 그 단어의 중요도는 떨어짐

	Tom	plays	soccer	loves	and	baseball	is	his	hobby	job
Doc1	0.18	0.48	0.18							
Doc2	0.18		0.18	0.48	0.18	0.18				
Doc3					0.18	0.18	0.48	0.48	0.48	0.48

TF-IDF 계산의 예

<https://medium.com/@nsh235482/tf-idf-term-frequency-inverse-document-frequency-algorithm-55f64714880d>

- ♦ TF-IDF 계산

- ♦ TF-IDF값은 TF값과 IDF값을 곱한 값이다.
- ♦ ‘his’ 단어의 경우 TF의 값이 2, IDF값이 0.48이므로 두개를 곱한 $2 \times 0.48 = 0.96$ 이 TF-IDF값이 된다.
- ♦ 아래 표를 보면 Doc1에서는 ‘plays’, Doc2에서는 ‘loves’, Doc3에서는 ‘his’가 가장 중요한 단어가 된다.

	Tom	plays	soccer	loves	and	baseball	is	his	hobby	job
Doc1	0.18	0.48	0.18							
Doc2	0.18		0.18	0.48	0.18	0.18				
Doc3					0.18	0.18	0.48	0.96	0.48	0.48

Cosine Similarity

- ◆ 두 벡터 간의 코사인 각도를 이용하여 구할 수 있는 두 벡터의 유사도를 의미



- ◆ 두 벡터의 방향이 완전히 동일한 경우에 1의 값을 가진다. 따라서 두 벡터가 유사한 방향성을 갖는지를 계산함.
- ◆ 유사도가 1에 가까울 수록 비슷한 벡터.
- ◆ 입력한 문장과 준비된 답변 문장이 얼마나 유사한지를 계산
예: [서울, 날씨] ~ [서울, 날씨, 미세먼지] => 0.816

Final Project

Final Project: Team Project

- ◆ 자유주제
 - ◆ 연구프로젝트
 - ◆ 연구문제를 설정하고 데이터를 수집한 후 분석하여 페이퍼 제출
 - ◆ Data Analysis Project & Paper (70 points)
 - ◆ Peer Review (30 points)

Final Project: Team Project

- ❖ 과제 사례

- ❖ 기후변화 프레임에 따른 위험인식 차이
- ❖ 온라인 커뮤니티의 커뮤니케이션 적응 (communication accommodation)에 대한 연구: ‘루리웹 유머 게시판’을 중심으로
- ❖ 날씨 및 계절이 한국인의 음악 청취에 미치는 영향
- ❖ 포털뉴스 댓글로 본 여성혐오 논쟁의 현주소

Final Project: Team Project

- ◆ 일정
 - ◆ 6/10: 최종 발표
 - ◆ 6/20: 최종 보고서 제출

Questions...?
