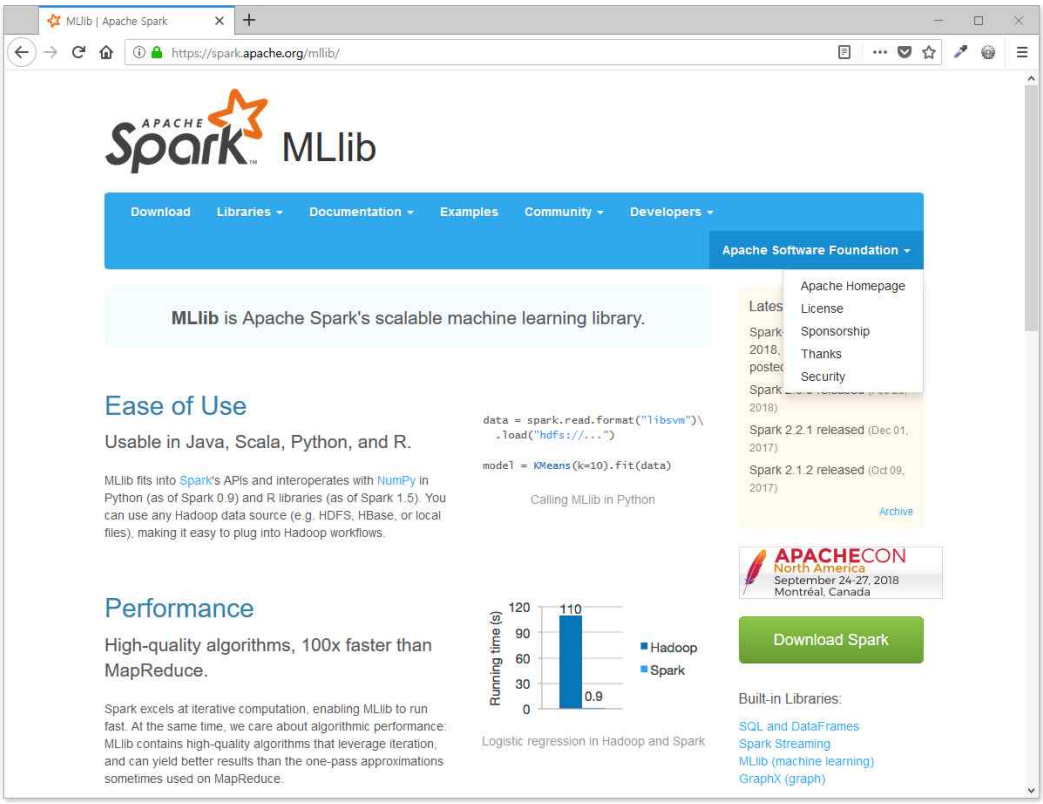
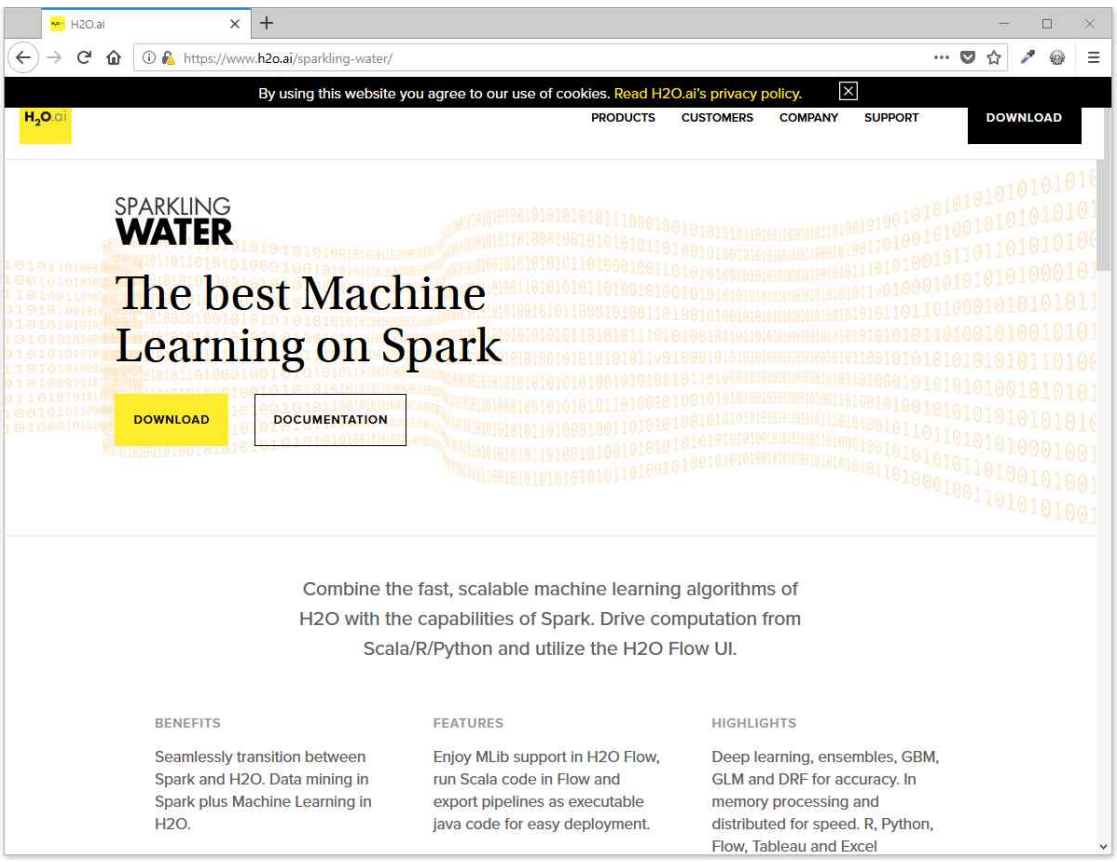
■ 우리 수업때 사용하게 될 머신러닝 / 딥러닝 API

1) 스파크 MLlib



- https://spark.apache.org/mllib/

2) 스파클링 워터(H20)

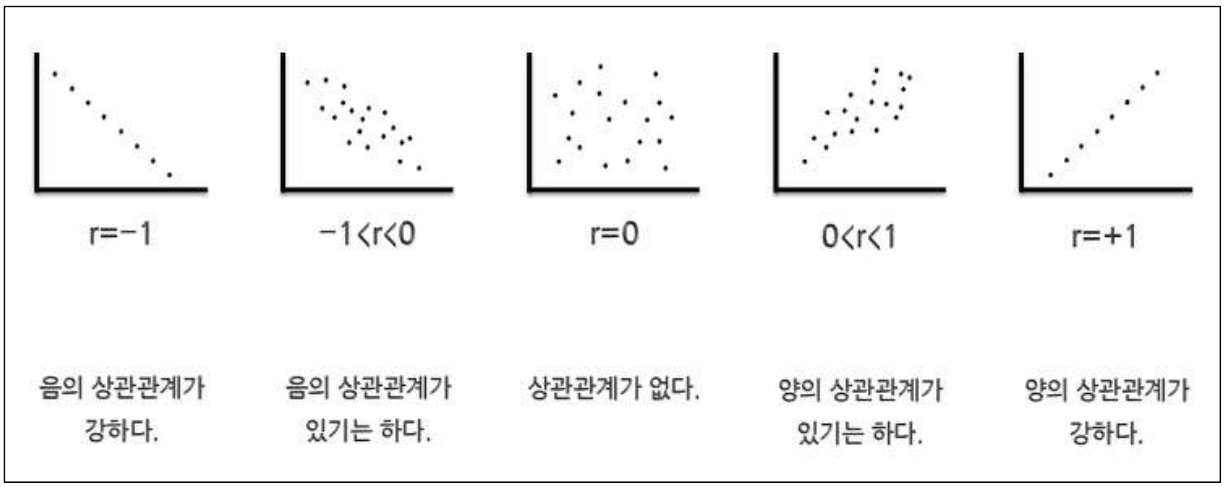


■ 상관분석

1) 상관분석은 독립변수와 종속변수간의 관계의 강도, 즉 얼마만큼 밀접하게 관련되어 있는지 분석하는 것

2) 이때 상관분석에서는 변수들 간에 상관성 유무만 확인 할 뿐, 서로 인과관계는 분석하지 않음

3) 상관분석의 핵심은 상관계수를 구하는 것



4) 여러 독립변수와 종속변수의 관계를 함수식으로 설명하는 방법

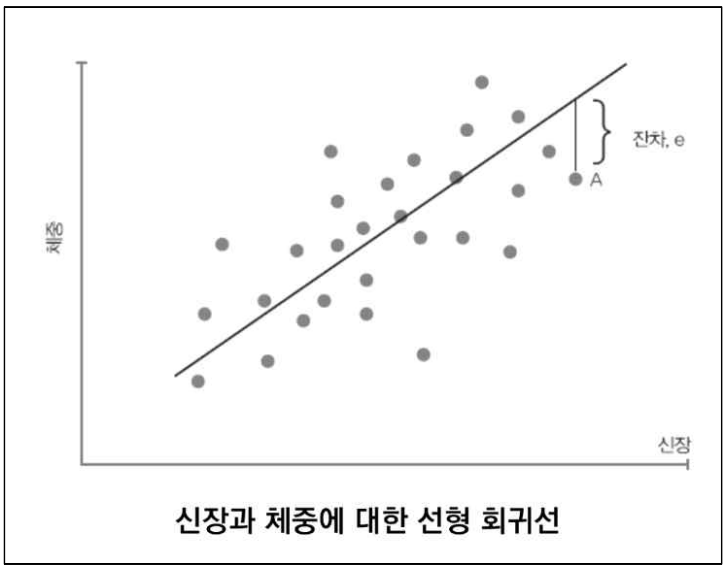
5) 종속변수란?

- 사실 우리가 알고 싶어하는 결과값(기대값 or 예상값이라고도 함)

6) 독립변수란?

- 결과값에 영향을 주는 입력값

7) 종속변수와 독립변수의 예



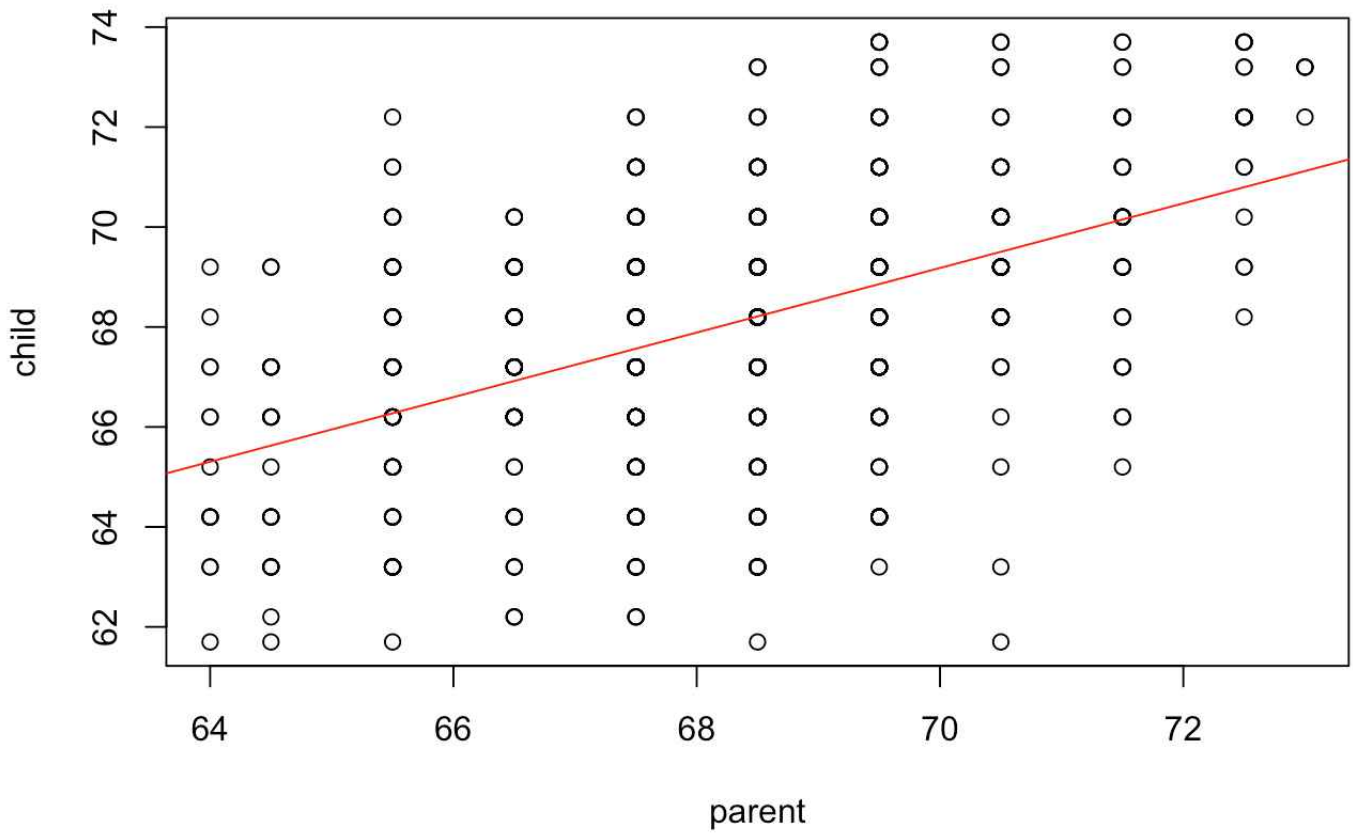
- 이때 상관관계를 함수식으로 규명하는 것이 회귀분석

■ 회귀분석

1) 회귀(regression)이라는 것은 말 그대로 '원 위치로 돌아간다'라는 뜻으로 상관과 밀접한 연관되어 있음

2) 회귀의 어원은 영국의 프랜시스 골턴이 세대별 키의 상관관계를 연구하다가 다윈의 진화론에 문제가 있음을 지적함

성인 자녀	부모 키										
	<64.0	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	>73.0	Totals
>73.7	—	—	—	—	—	—	5	3	2	4	—	14
73.2	—	—	—	—	—	3	4	3	2	2	3	17
72.2	—	—	1	—	4	4	11	4	9	7	1	41
71.2	—	—	2	—	11	18	20	7	4	2	—	64
70.2	—	—	5	4	19	21	25	14	10	1	—	99
69.2	1	2	7	13	38	48	33	18	5	2	—	167
68.2	1	—	7	14	28	34	20	12	3	1	—	120
67.2	2	5	11	17	38	31	27	3	4	—	—	138
66.2	2	5	11	17	36	25	17	1	3	—	—	117
65.2	1	1	7	2	15	16	4	1	1	—	—	48
64.2	4	4	5	5	14	11	16	—	—	—	—	59
63.2	2	4	9	3	5	7	1	1	—	—	—	32
62.2	—	1	—	3	3	—	—	—	—	—	—	7
<61.7	1	1	1	—	—	1	—	1	—	—	—	5
Totals	14	23	66	78	211	219	183	68	43	19	4	928



3) 골턴은 부모와 자식 간 키와 몸무게의 상관 관계를 분석했는데, 키가 큰 아버지의 아들은 아버지보다 작은 경향이 있고, 반대로 키가 작은 아버지의 아들은 키가 큰 경향이 있다는 사실을 발견함

4) 그래서, '아버지 + 아들의 키는 평균으로 회귀한다' 라는 표현을 사용했고, 그 이후 이러한 형태 분석법을 '회귀분석'이라고 부르게 되었음

5) 즉, 꼭 결과가 회귀해야만 '회귀분석'은 절대로 아님

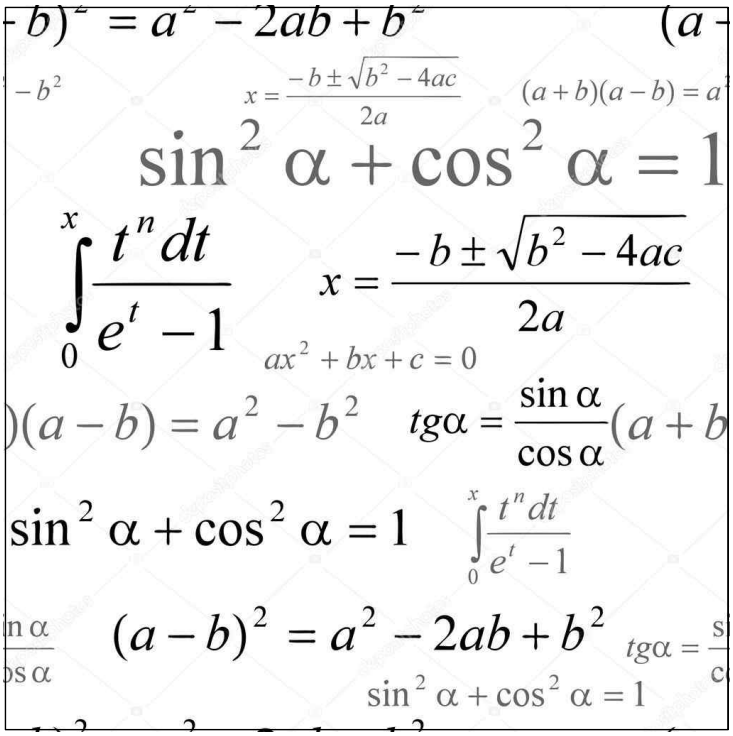
6) 회귀의 종류

- 선형 회귀
- 로지스틱 회귀
- 다항 회귀
- 단계적 회귀
- 리지 회귀
- 라소 회귀
- 엘라스티뉴 회귀

7) API를 활용하여 분석이 가능함

■ 신경망의 이해와 딥러닝

1) 인간과 컴퓨터



문제	컴퓨터	사람
수학 공식(수천개의 숫자 곱하기)	쉬 움	어려움
사진에서 강아지 찾기	어려움	쉬 움

2) 1과 0으로 구성된 차갑고 딱딱한 논리 기반(if~then)의 기계가 미묘하고 모호한 생물학적인 두뇌의 사고 체계를 달성하는 것은 불가능한 일이라고 믿게 됨

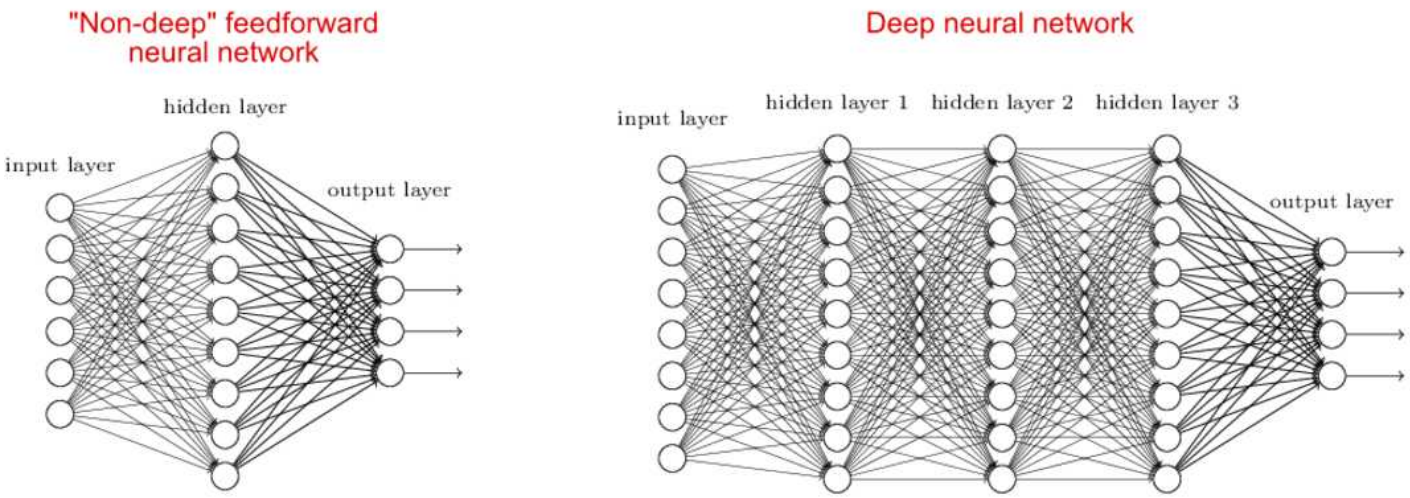
3) 기발한 아이디어가 등장 : 꿀벌이나 비둘기가 매우 단순한 뇌 구조를 가졌음에도 복잡한 업무를 수행할 수 있음 / 즉 동물의 뇌를 그대로 복제함으로써 인공두뇌를 만들 아이디어가 나타남

- 딱딱한 흑백논리를 가진 알고리즘보다 부드럽고 유기적인 사고 체계를 도입하자는 아이디어

4) 인간의 뉴런을 본따서 '신경망'이 등장함 : 신경망은 오래전부터 있던 아이디어

5) 2010년 캐나다 토론토대학의 제프리 힌튼 교수 연구팀이 '세계 최대 이미지 인식 경연대회'에 출전하여, 옥스퍼드, 도쿄대, 독일 예나대, 제록스 연구소 등의 유명 연구기관이 개발한 인공지능을 압도적인 차이로 누르며 우승함

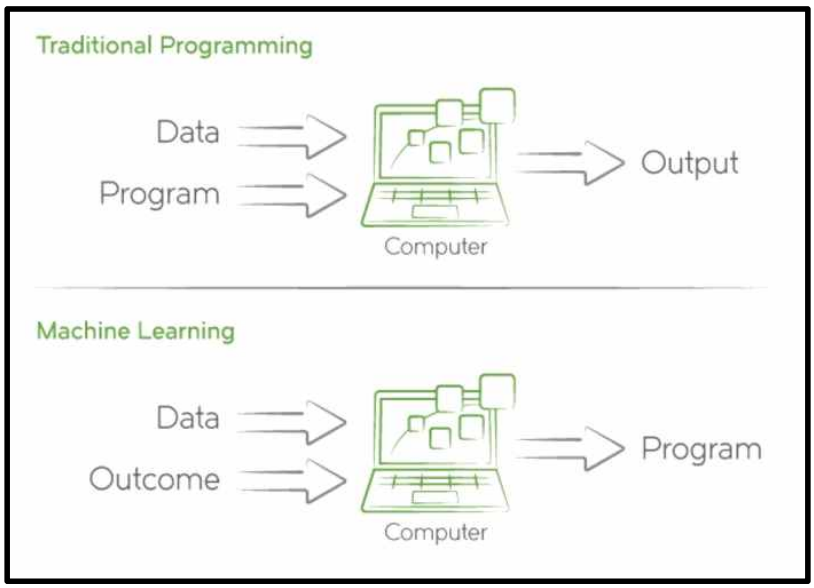
- 다른 팀이 오답률 26%대의 소수점 공방을 벌일때 '딥러닝'이라는 기법을 사용한 '슈퍼비전'팀은 15%의 오답률을 기록함



6) 신경망을 촘촘히 여러층으로(deep) 구성하여 인공지능 분야의 엄청난 발전이 일어남

■ 간단한 예측자

1) 머신러닝 / 딥러닝과 다른 프로그래밍의 차이는?



- 기존 프로그래밍 '알고리즘'을 만들고 입력값을 넣어서 '출력'을 얻음
- 머신러닝은 '입력'과 '출력'을 빅데이터로 넣어줘서 학습시켜 '알고리즘'을 만들어 냄

2) 가장 기초적인 신경망의 이해(간단한 예측자 이해)

- 선형 관계의 예측자 만들기
- 킬로미터를 마일로 변환하는 예측자

마일 = 킬로미터 x c(c는 상수)

- 결국 상수 C를 구하는 방법
- 빅데이터(사실 스몰데이터)가 필요함

km	mi
0.1	0.0621371192
0.2	0.1242742384
0.3	0.1864113577
0.4	0.2485484769
0.5	0.3106855961
0.6	0.3728227153
0.7	0.4349598346
0.8	0.4970969538
0.9	0.5592340730
1	0.6213711922
1.1	0.6835083115
1.2	0.7456454307

- 만약 0.5를 대입해본다면

100 -> 100 x 0.5 -> 50
오차는 12.137...

- 만약 0.6을 대입한다면

100 -> 100 x 0.6 -> 60
오차는 2.137...

- 만약 0.7을 대입한다면

100 -> 100 x 0.7 -> 70
오차는 -7.863...

- 오버슈팅 발생
- 다시 0.61을 대입

100 -> 100 x 0.61 -> 61
오차는 1.137...

- 꽤 정확한 값을 얻음

3) 위의 방식으로 인공 신경망을 학습시키는 핵심과정 : 강화학습

- 반복을 통해 점점 더 정답에 가까운 값을 스스로 얻음