선수 지식 - 통계
표준정규분포

표준정규분포 | 딥러닝의 기초가 되는 확률 개념 알아보기

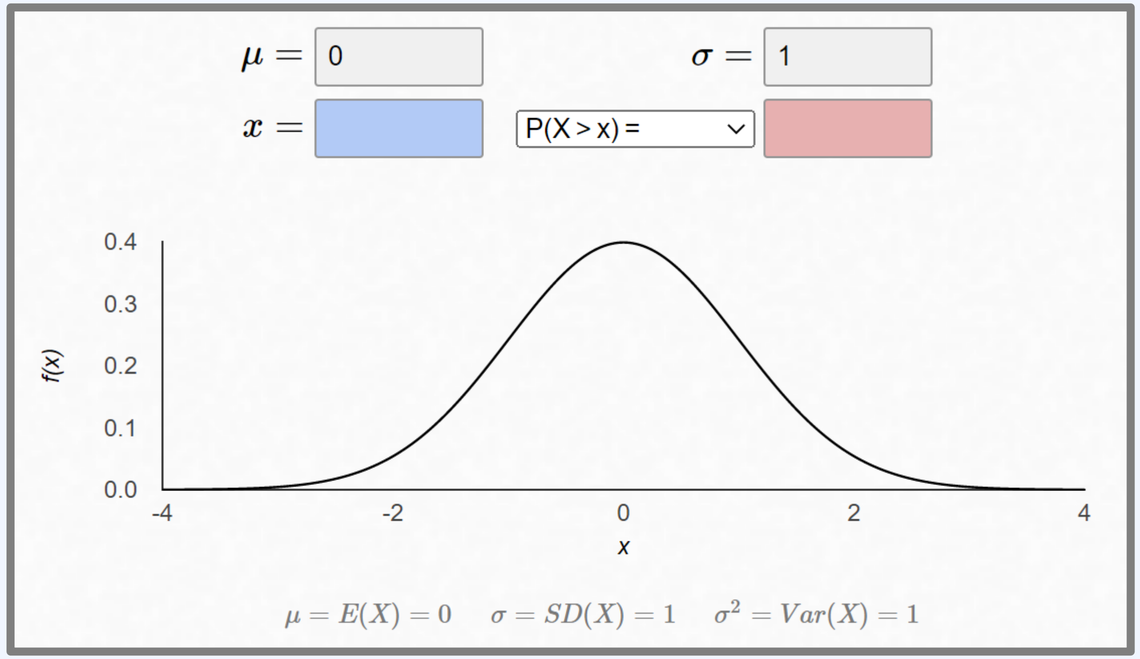
강사 나동빈

선수 지식 - 통계

표준정규분포

표준 정규 분포(Standard Normal Distribution)

- 표준 정규 분포(standard normal distribution)란?
- 평균이 0이고 분산이 1인 **표준화된 정규 분포**다.



* https://homepage.divms.uiowa.edu/~mbognar/applets/normal.html

표준 정규 분포(Standard Normal Distribution)

- 확률 변수 X가 $X \sim N(\mu, \sigma^2)$을 따를 때, 다음의 공식으로 표준화를 할 수 있다.

$$Z = \frac{X - \mu}{\sigma}$$

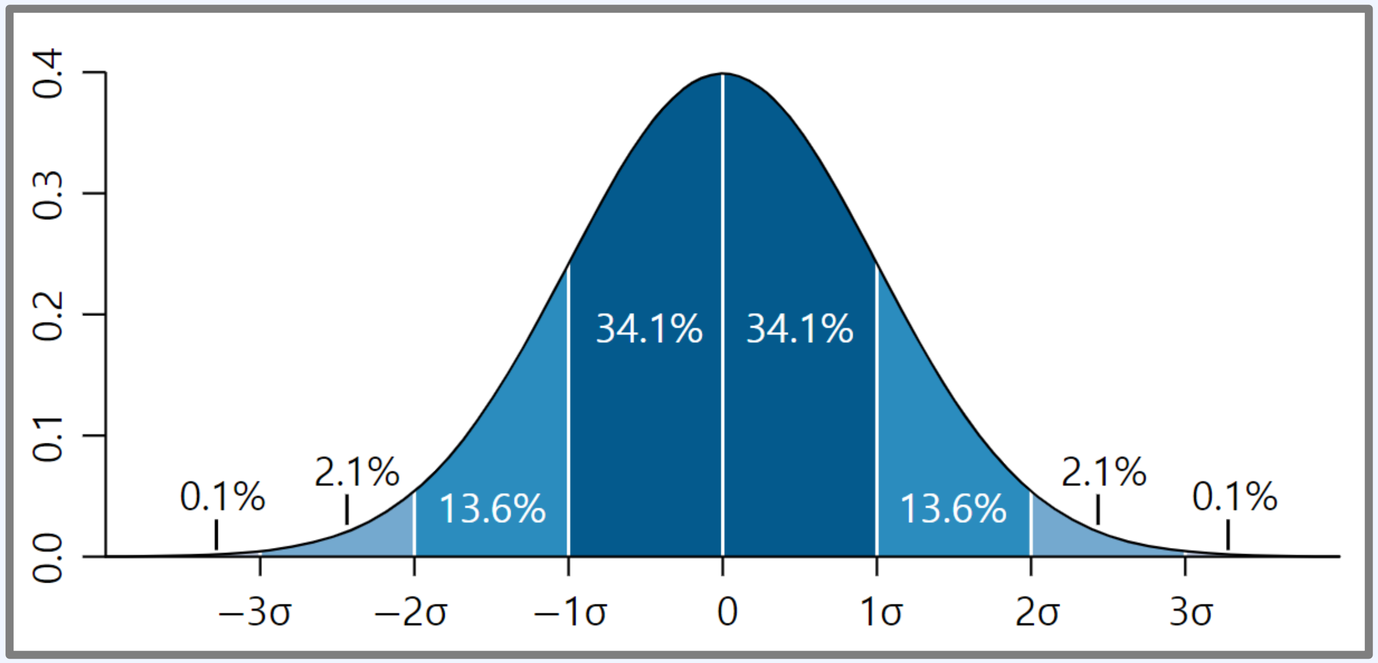
표준 정규 분포(Standard Normal Distribution)

- 확률 변수 Z가 평균이 0이고, 분산이 1인 정규분포를 따르는 상황을 생각해 보자.
- 이때 Z는 표준정규분포를 따른다고 말한다. 즉, $Z \sim N(0,1)$로 표현한다.
- 확률 변수 Z의 확률밀도함수는 다음과 같다. (단순히 정규 분포의 PDF에서 유도)

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

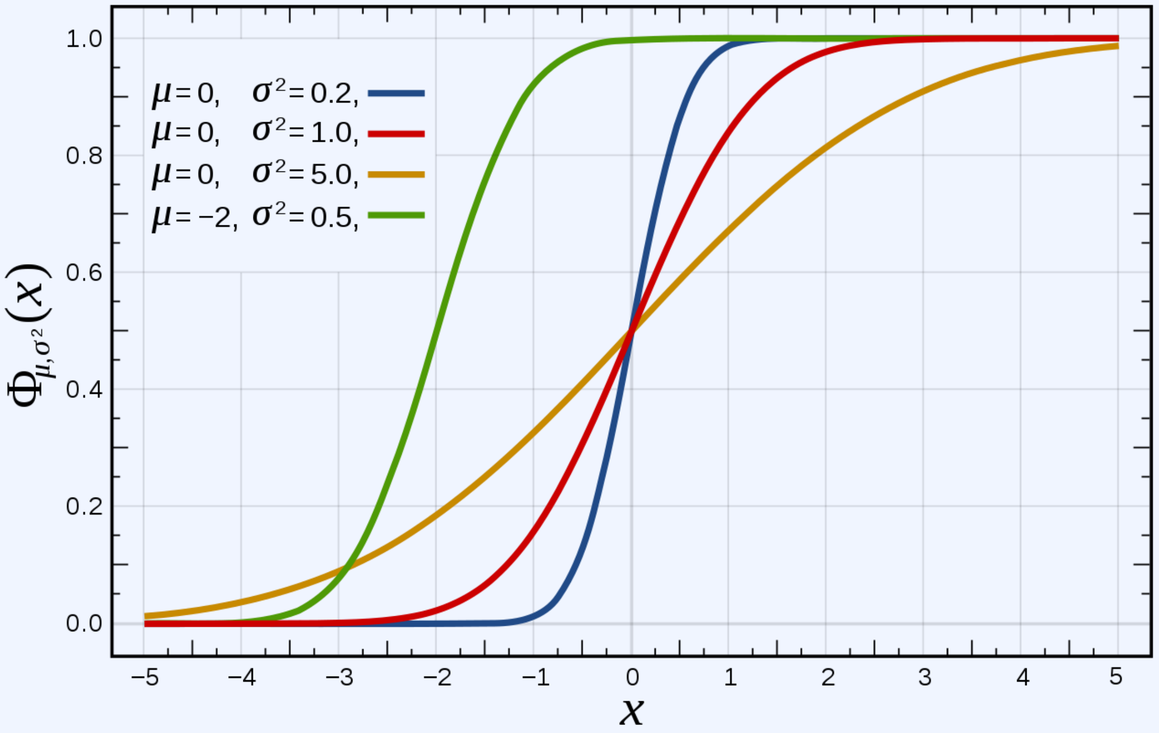
표준 정규 분포(Standard Normal Distribution)

- **정규 분포**의 확률을 계산할 때는 다음의 그림을 참고한다.
- 표준 정규 분포의 경우 σ의 값이 1이므로, $P(Z \leq 1)$을 **약 84.1%**로 볼 수 있다.



표준 정규 분포의 누적 분포 함수

- 아래 그래프에서 정규 분포의 누적 분포 함수를 확인할 수 있다.



* https://ko.wikipedia.org/wiki/%EC%A0%95%EA%B7%9C_%EB%B6%84%ED%8F%AC

표준 정규 분포 표

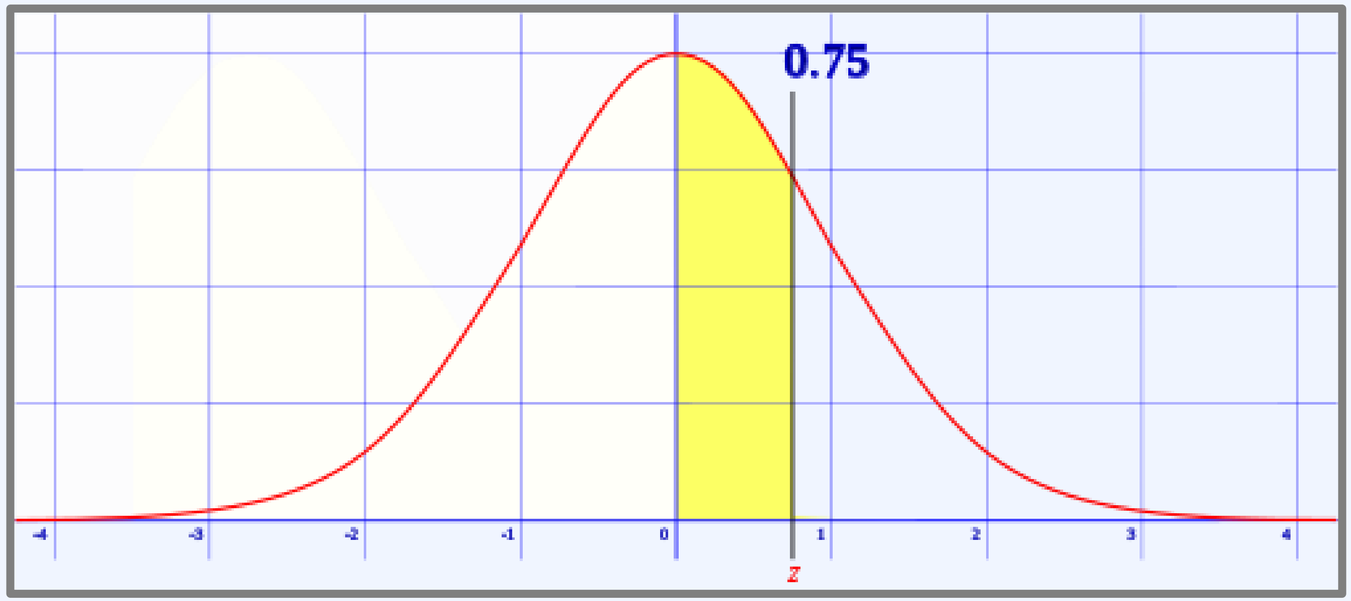
- 정규 분포의 **누적 분포 함수 값**에 대한 표이다.

z	+0.00	+0.05
0.0	0.50000	0.51994
0.2	0.57926	0.59871
0.5	0.69146	0.70884
0.7	0.75804	0.77337
0.8	0.78814	0.80234
1.0	0.84134	0.85314
1.5	0.93319	0.93943
2.0	0.97725	0.97982
2.5	0.99379	0.99461
3.0	0.99865	0.99886

* https://ko.wikipedia.org/wiki/%ED%91%9C%EC%A4%80%EC%A0%95%EA%B7%9C%EB%B6%84%ED%8F%AC%ED%91%9C

표준 정규 분포 표

- 표준 정규 분포에서 $P(0 \leq z \leq 0.75)$는 얼마일까?
- $P(0 \leq z \leq 0.75) = P(z \leq 0.75) - P(z \leq 0) = 0.77337 - 0.5 = 0.27337$



* https://ko.wikipedia.org/wiki/%ED%91%9C%EC%A4%80%EC%A0%95%EA%B7%9C%EB%B6%84%ED%8F%AC%ED%91%9C

표준 정규 분포 예시 ① 확률 계산하기

- 표준 정규 분포에서 z가 0.8 이하일 확률은 얼마일까?
- 구하고자 하는 확률 값은 $P(z <= 0.8)$이다.

[해설] 표준 정규 분포 표에 따르면, 확률은 0.78814이다.

z	+0.00	+0.05
0.0	0.50000	0.51994
0.2	0.57926	0.59871
0.5	0.69146	0.70884
0.7	0.75804	0.77337
0.8	0.78814	0.80234

표준 정규 분포 예시 ② IQ 계산하기

- 흔히 **지능 지수(IQ)**를 판단할 때, 평균 IQ를 100으로 설정한다.
- 한국에서는 기본적으로 표준 편차 σ를 24로 설정한다.
- 그렇다면 IQ가 148이라면, 상위 몇 %에 해당하는가?

표준 정규 분포 예시 ② IQ 계산하기

- 확률 변수 X가 $X \sim N(100, 24^2)$일 때, X가 148 이상일 확률을 구해보자.
- $P(X > 148) = P\left(\frac{X-\mu}{\sigma} > \frac{148-\mu}{\sigma}\right)$

$$= P\left(Z > \frac{148-\mu}{\sigma}\right)$$

따라서 일반적으로 지능 검사에서 **상위 2%**의 성적을 얻었다면, **IQ 148**로 본다.

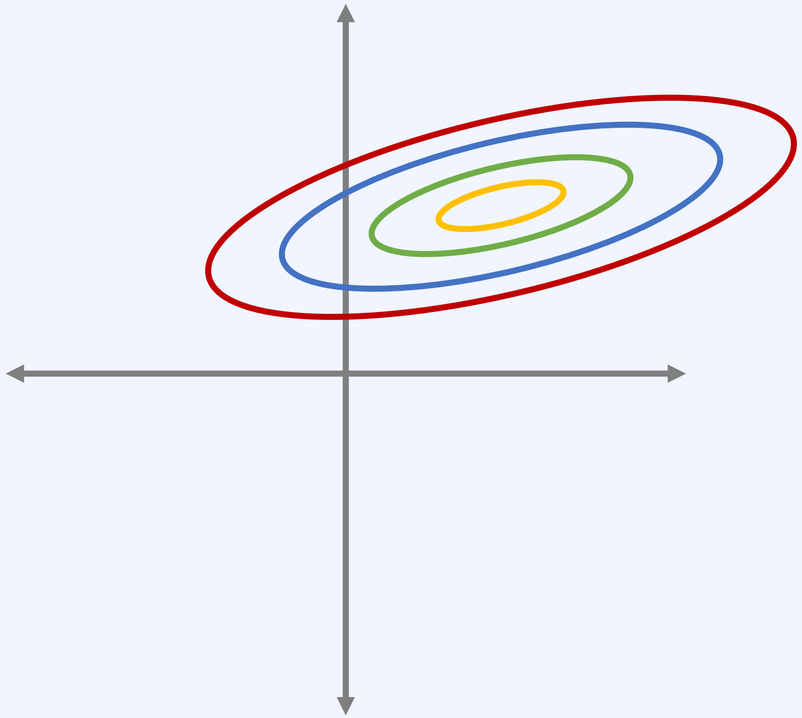
$$= P\left(Z > \frac{148-100}{24}\right)$$

$$= P(Z > 2) = 1 - f_Z(2) = 1 - 0.97725 = 0.02275$$

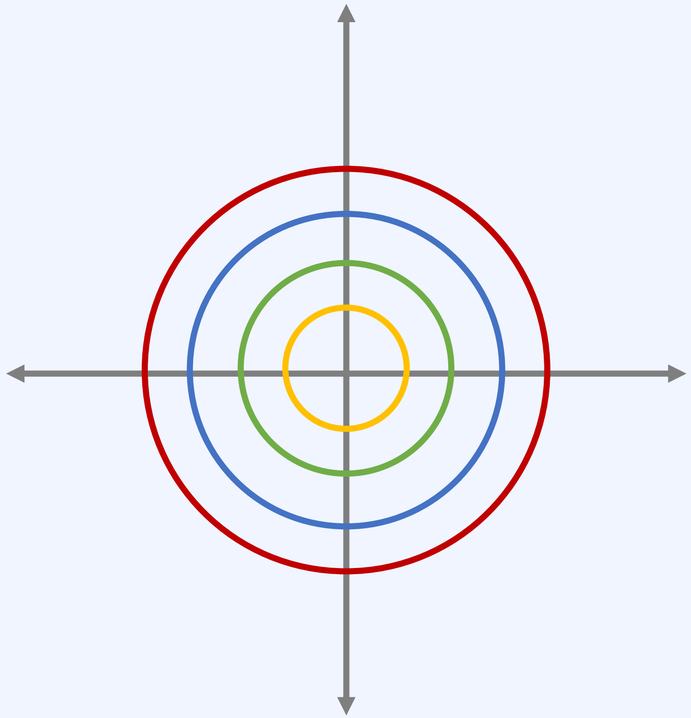
딥러닝 분야의 입력 정규화(Input Normalization)

- 입력 데이터를 정규화하여 **학습 속도(training speed)를 개선**할 수 있다.

[정규화 전]



[정규화 후]



딥러닝 분야의 입력 정규화(Input Normalization)

- 입력 데이터가 $N(0,1)$ 분포를 따르도록 **표준화**하는 예제는 다음과 같다.

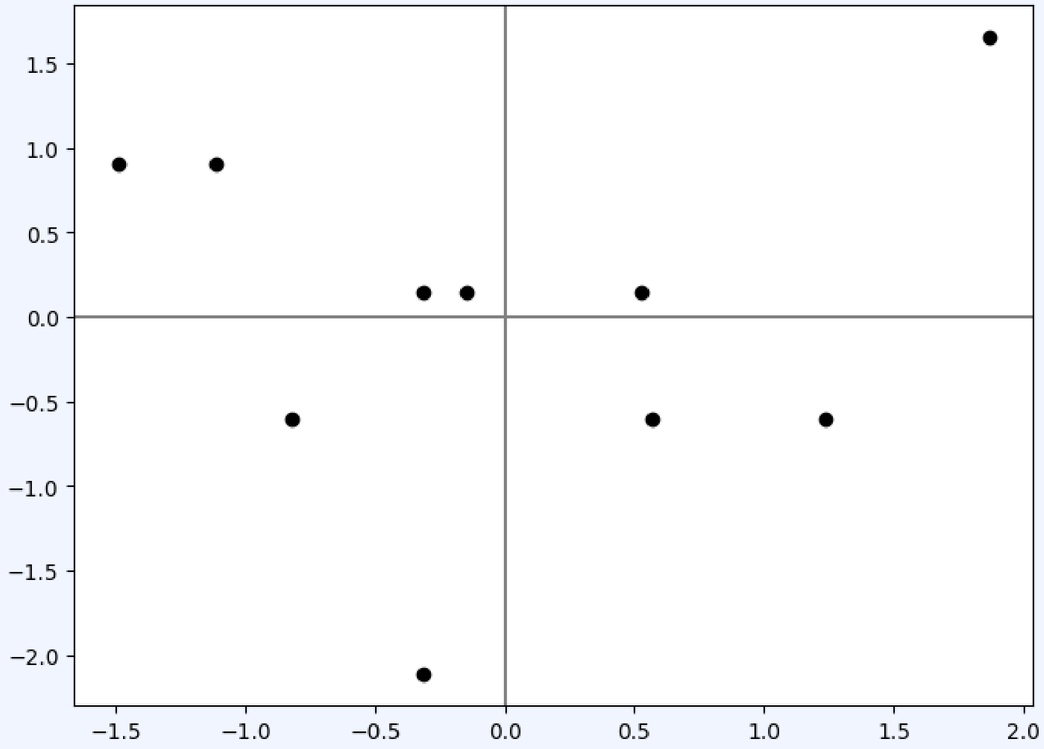
$$\hat{x} = \frac{x - E[x]}{\sqrt{Var[x]}}$$

```
x1 = np.asarray([33, 72, 40, 104, 52, 56, 89, 24, 52, 73])
x2 = np.asarray([9, 8, 7, 10, 5, 8, 7, 9, 8, 7])

normalized_x1 = (x1 - np.mean(x1)) / np.std(x1)
normalized_x2 = (x2 - np.mean(x2)) / np.std(x2)

plt.axvline(x=0, color='gray')
plt.axhline(y=0, color='gray')
plt.scatter(normalized_x1, normalized_x2, color='black')
plt.show()
```

딥러닝 분야의 입력 정규화(Input Normalization)

- 입력 데이터가 $N(0, 1)$ 분포를 따르도록 **표준화**하는 예제는 다음과 같다.

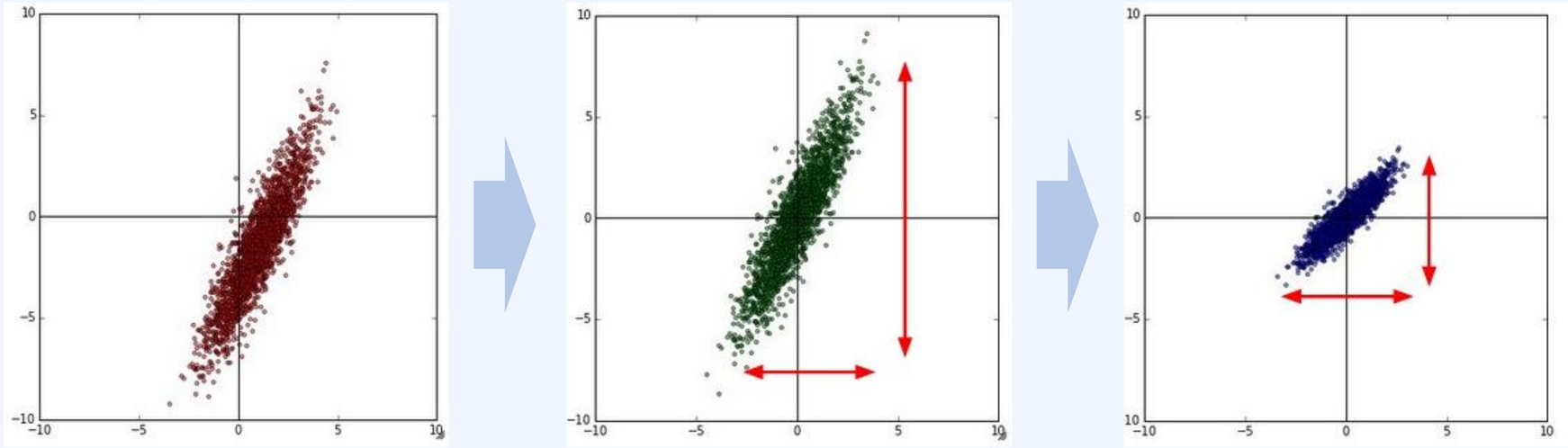
[평균(mean) = 0, 분산(variance) = 1]



딥러닝 분야의 입력 정규화(Input Normalization)

- **입력 정규화**를 이용해 각 차원의 데이터가 동일한 범위 내의 값을 갖도록 만들 수 있다.
- 모든 특성(feature)에 대하여 각각 평균만큼 빼고 특정 범위의 값을 갖도록 조절할 수 있다.

[2개의 특성(feature)으로 구성된 데이터셋의 정규화 예시]



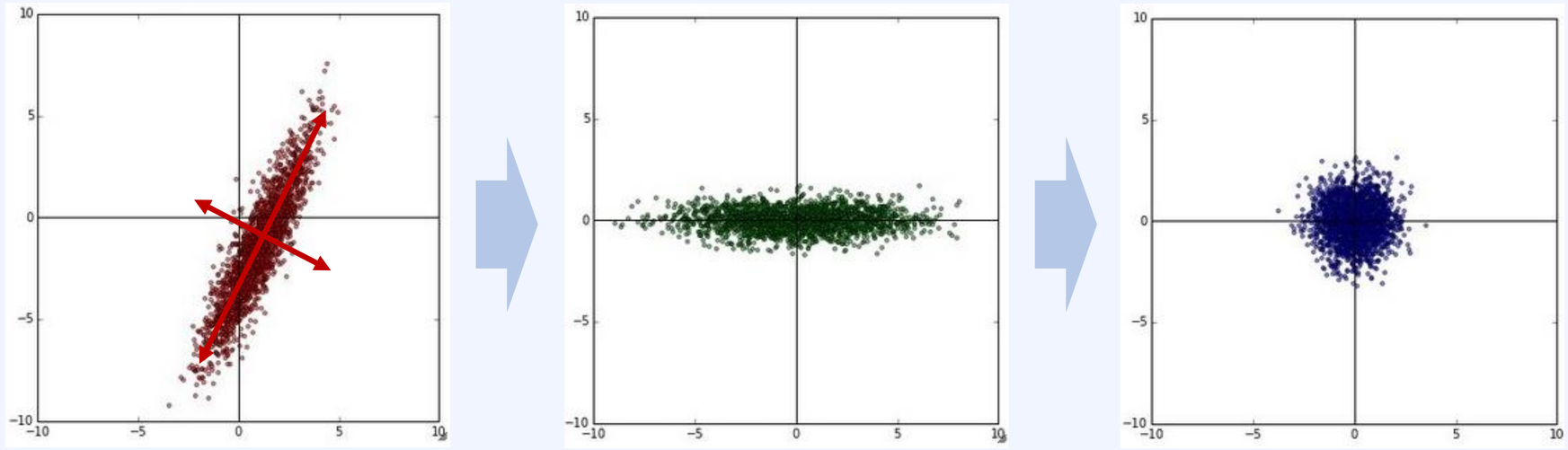
원본 데이터셋 평균이 0이 된 데이터셋 정규화된 데이터셋

* https://cs231n.github.io/neural-networks-2/

딥러닝 분야의 입력 정규화(Input Normalization)

- 화이트닝은 평균이 0이며 공분산이 단위행렬인 정규분포 형태의 데이터로 변환한다.
- 일반적으로 딥러닝 분야에서는 PCA나 화이트닝보다는 정규화가 더 많이 사용된다.

[2개의 특성(feature)으로 구성된 데이터셋의 화이트닝 예시]



원본 데이터셋 Decorrelated 데이터셋 화이트닝된 데이터셋

* https://cs231n.github.io/neural-networks-2/