투이워너비

Bootstrap 방법을 이용한 Machine Learning

2018. 02. 02 (金)

투이워너비 4기 교육생 허성욱

actto8290@gmail.com

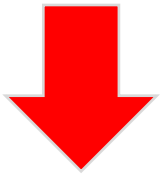
https://github.com/AnalystH

분산(σ^2) & 중앙값

자료 및 코드 : https://github.com/AnalystH

$$X = \{5, 10, 2, 3, 6, 9, 1, 20, 8\}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$



$$X' = \{1, 2, 3, 5, 6, 8, 9, 10, 20\}$$

중앙값 = 6

중앙값의 분산

자료 및 코드 : https://github.com/AnalystH

표본에 중앙값이 하나 밖에 존재하지 않은데 분산을 어떻게 구해야 될까요?

분산의 중요성

자료 및 코드 : https://github.com/AnalystH

가설검정

Test For	Null Hypothesis (H_0)	Test Statistic	Distribution
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x} - \mu_0)}{\sigma/\sqrt{n}}$	Z
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x} - \mu_0)}{s/\sqrt{n}}$	t_{n-1}
Population proportion (p)	$p = p_0$	$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	Z
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Z
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	t distribution with df = the smaller of $n_1 - 1$ and $n_2 - 1$
Mean difference μ_d (paired data)	$\mu_d = 0$	$\frac{(\bar{d} - \mu_d)}{s_d/\sqrt{n}}$	t_{n-1}
Difference of two proportions ($p_1 - p_2$)	$p_1 - p_2 = 0$	$\frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	Z

The Black-Scholes Formulas

$$c = S_0\, N(d_1) - K\, e^{-rT} N(d_2)$$

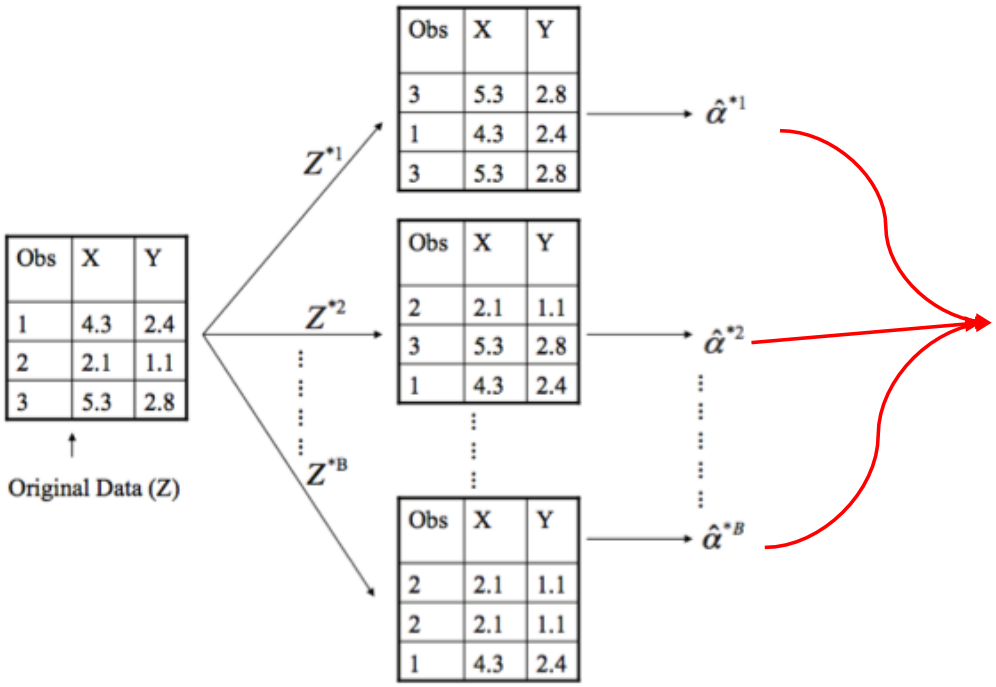
$$p = K\, e^{-rT}\, N(-d_2) - S_0\, N(-d_1)$$

$$\text{where} \quad d_1 = \frac{\ln(S_0/K) + (r + \sigma^2/2)T}{\sigma\sqrt{T}}$$

$$d_2 = \frac{\ln(S_0/K) + (r - \sigma^2/2)T}{\sigma\sqrt{T}} = d_1 - \sigma\sqrt{T}$$

Bootstrap

자료 및 코드 : https://github.com/AnalystH



$$\frac{1}{B}\sum_{i=1}^{B}\left(\hat{\alpha}^{*i} - \hat{\alpha}\right)^2 \approx Var(\text{중앙값})$$

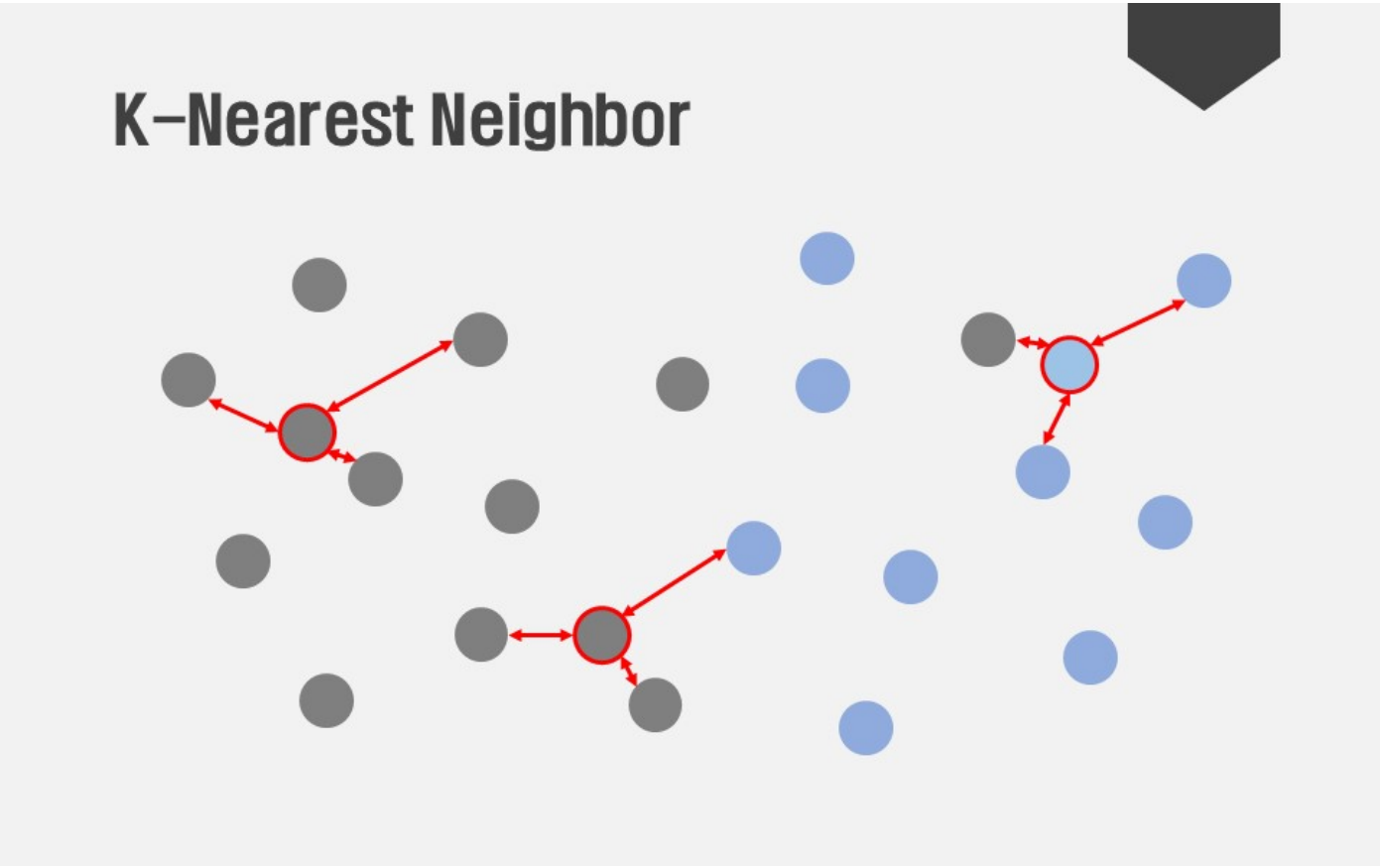
✓ 평균이 5, 분산이 10인 정규분포에서 랜덤으로 100개 추출한 데이터로 테스트 결과

Bootstrap : 0.3696774

실제값 : 0.376

K-Nearest Neighbor(K-NN)

자료 및 코드 : https://github.com/AnalystH



Iris Data

자료 및 코드 : https://github.com/AnalystH

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
42	4.5	2.3	1.3	0.3	setosa
43	4.4	3.2	1.3	0.2	setosa
44	5.0	3.5	1.6	0.6	setosa
45	5.1	3.8	1.9	0.4	setosa
46	4.8	3.0	1.4	0.3	setosa
47	5.1	3.8	1.6	0.2	setosa
48	4.6	3.2	1.4	0.2	setosa
49	5.3	3.7	1.5	0.2	setosa
50	5.0	3.3	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
54	5.5	2.3	4.0	1.3	versicolor
55	6.5	2.8	4.6	1.5	versicolor
56	5.7	2.8	4.5	1.3	versicolor
57	6.3	3.3	4.7	1.6	versicolor
58	4.9	2.4	3.3	1.0	versicolor
59	6.6	2.9	4.6	1.3	versicolor

Sepal.Width : 꽃받침의 너비

Sepal.Length : 꽃받침의 길이

Petal.Width : 꽃잎의 너비

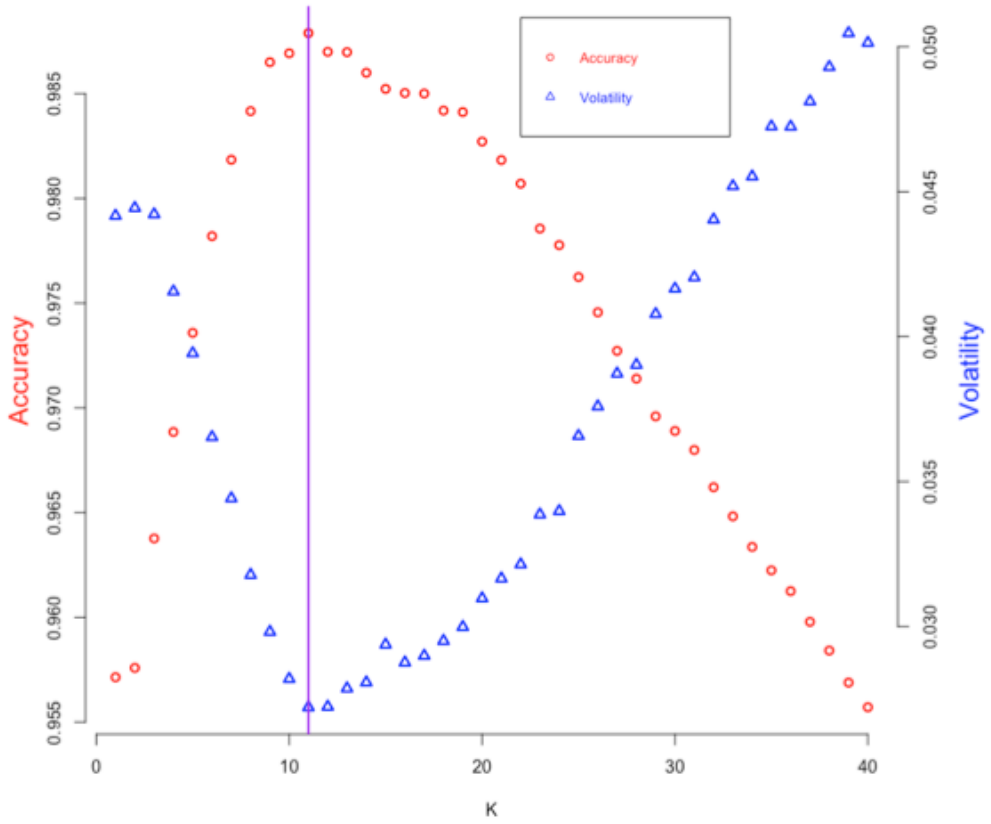
Petal.Length : 꽃잎의 길이

Species : *setosa, versiclolor, virginica*

1. 자료를 복원추출로 랜덤하게 150개를 뽑는다.
2. 1 번 행위를 1000번 반복한다.
3. 각 데이터 셋에서 Train : Test = 2 : 1로 나눈다.
4. Train으로 모델을 만들고 Test로 검증한다.
5. 1000번의 결과에서 평균 및 분산을 구한다.
6. 위 1 ~ 5번 행위를 K = 1 부터 K = 40까지 반복한다.

K-Nearest Neighbor with Bootstrap

자료 및 코드 : https://github.com/AnalystH



Thank you!

허성욱 | actto8290@gmail.com

투이컨설팅