Power Transformation

2016314726 정영준

Power Transformation에 대해 설명하기 전에 먼저 Transformation이 필요한 이유에 대해 설명하겠습니다. 수집된 가공하지 않는 데이터를 raw data라고 하는데 이 raw data는 바로 시각화, 또는 머신러닝 모델에 사용하기에 부적합합니다. 먼저 데이터가 수집될 때 그 단위가 맞지 않기 때문입니다. 예를 들어 소득 데이터를 수집하는데 1,000원 단위, 10,000 단위로 수집된 데이터가 섞여있을 경우 이를 토대로 시각화나 머신러닝을 할 경우 올바른 결과를 얻을 수 없습니다. 그리고 다음으로는 이상치가 존재하기 때문입니다. 하나의 매우 크거나 작은 이상치로 인해 데이터 시각화의 경우 제대로 데이터의 분포를 설명해주지 못하거나 머신러닝 모델의 경우 정확도가 매우 떨어지는 현상이 발생할 수 있습니다. 또 다른 이유로는 feature들 사이의 다중공선성이나 비대칭성이 강한 정규분포와 다른 분포 등이 문제가 될 수 있습니다. Power Transformation은 가우스 분포와 다른 입력 또는 출력 변수를 정규 분포에 가깝게 변환하는 과정입니다. 많은 머신러닝 알고리즘은 입력 데이터가 정규분포라는 가정하에 이루어집니다. 3가지 케이스로 변환이 가능한데 하나씩 알아보겠습니다. 먼저 기초 정의에 기반한 Power Transformation입니다. P = 0 일 때 ax^p + b 로 변환을 하고 p가 0이 아닐 때는 c\*log(x)+d로 변환이 이루어집니다. 조금 더 간단한 적용법 으로는 p>0 일 때 x^p, p가 0일 때 log x, p < 0 일때 -x^p로 변환이 가능합니다. 마지막으로 case 3는 p가 0이 아닐때는 (x^p-1)/p, p가 0일 때는 ln x로 변환이 가능합니다. 이러한 변환의 장점은 하나의 데이터의 배치에서 대칭성을 향상시켜줍니다. 또 배치간에 일정한 분산이 존재하도록 할 수 있습니다. 마지막 case 3의 변환은 Box-Cox 변환이라고도 합니다. 이는 log transformation과 Power transformation이 둘다 사용되는 변환으로 시계열 데이터에서 계절성을 명확히 나타내는데 사용됩니다. 데이터가 모두 양수여야 Box-Cox 변환을 할 수 있기 때문에 데이터 전체의 평균을 이동시키고 변환을 적용하여야 합니다. p값이 1일경우 x는 x-1로 변환이 되는데 이는 거의 변화가 없는 항등변화에 가깝습니다. 신뢰구간에 p가 1일 경우가 들어간다면 이는 분산에 의해 정상성을 충족하지 못하는 경우가 아님을 뜻합니다. 이를 이용하여 시계열 변환에서 Box-Cox 변환을 통해 정상성을 검증합니다. 평균에 의한 정상성의 검증은 다른 방식을 추가로 이용하여야 합니다. 변환이 이루어지는 과정을 쉽게 풀이하여 쓰면 가우시안 분포와 다른 분포가 있을 경우 이를 자료가 조밀하게 밀집되어 있는 구간에서는 척도를 늘려 데이터를 조금 더 퍼지도록 하는 것입니다. 마치 고무줄에 선을 표시하고 이를 늘리면 두 점사이의 거리가 늘어나는것과 같습니다. 얼마나 고무줄을 늘릴 것인가 이것을 결정하는 변수가 p입니다. p값의 조정을 통해 전체적인 원래 데이터의 분포의 밀집된 부분의 척도를 얼마나 늘릴지 결정이 가능한 것입니다. Box-Cox 변환을 통해 정규분포에 가까워진 데이터는 선형 회귀, 신경망 등의 알고리즘에서 전보다 높은 성능을 보입니다.