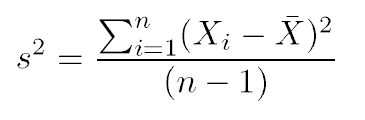
4월14일~4월22일

∙ 저 번주 연구노트에 이어서 논문 주제로 선정한 Principal Components Analysis(PCA)에 대한 mathematical concepts을 공부합니다.

**1.2 Variance**

분산은 데이터 세트에서 데이터의 확산을 측정하는 또 다른 방법이다. 사실 그것은 거의 표준 편차와 동일하다. 공식은 다음과 같다.



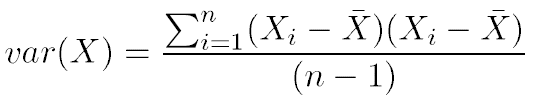
기호(와 수식(분산 수식에는 제곱근이 없다)에서 표준 편차 제곱임을 알 수 있다. 는 표본의 분산에 대한 일반적이 기호이다. 이 측정은 모두 데이터의 확산을 측정한 것이다. 표준 편차가 가장 일반적인 척도 이지만 분산도 사용된다. 표준 편차 외에 분산을 도입을 한 이유는 다음 섹션인 공분산에서 탄탄한 플랫폼을 제공하기 위해서이다.

**1.3 Covariance**

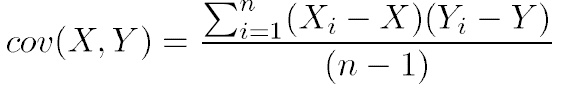
앞에서 살펴본 것은 1차원 관점 이였다. 예를 들어, 데이터 세트안에 있는 모든 사람들의 키, 마지막 COMP101 시험에 대한 점수 등이 될 수 있다. 그러나 많은 데이터 세트에는 둘 이상의 차원이 있으며 이러한 데이터 세트의 통계 분석의 목적은 일반적으로 차원 간에 관계가 있다. 예를 들어, 우리는 데이터로 클래스의 모든 학생의 키와 시험 점수를 둘 수 있다. 그런 다음 통계 분석을 수행하여 학생의 신장이 자신의 점수에 어떠한 영향을 미치는지 확인할 수 있다.

표준 편차 및 분산은 1차원에서만 작동하므로 독립적으로 데이터 집합의 각 차원에 대한 표준 편차를 계산할 수 있다. 그러나 차원이 서로 얼마나 다른지 알아 내려면 비슷한 방법을 사용하는 것이 좋다.

공분산은 항상 2차원 사이에서 측정된다. 하나의 차원과 그 자체의 공분산을 계산하면 분산이 생긴다. X와 y차수, x와 z차수, y와 z차수 사이의 거리이다. X와 x 또는 y와 y 또는 z와 z 사이의 공분산을 측정하면 x, y 및 z차원의 분산이 각각 나타난다. 공분산의 공식은 분산 공식과 매우 유사한다. 분산 수식은 다음과 같이 작성할 수 있다.



여기서 단지 두 부분을 보여주기 위해 제곱근을 확장했다. 따라서 다음과 같은 공식을 실현할 수 있다.



대괄호의 두 번째 세트에서 X가 Y로 대체된다는 것을 제외하면 정확히 동일하다. 이것은 “각 데이터 항목에 대해 x값과 x평균값의 차이에 y값과 y평균값의 차이를 곱한다. 이 모든 것을 더하고, (n-1)로 나눈다.”

몇 가지 궁금증에 대해 물어볼 수 있다. “cov(X,Y)가 cov(Y,X)와 같은가요?” 라고 물을 수 있다. 공식을 살펴 보면 cov(X,Y)와 cov(Y,X) 사이의 차이는 가 로 대체된다는 것이다. 그리고 곱셈은 교환적이기 때문에, 두 개의 숫자를 곱하는 방법은 상관 없다. 항상 같은 수를 얻는다. 이 두 방정식은 같다고 할 수 있다.