4월7일~4월13일

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

1. **Statistics**

큰 data set을 가지고 있다는 생각을 기반으로 하여, 데이터 세트의 각각의 점 사이의 관계로 그 set을 분석합니다.

* 1. **Standard Deviation**

표준 편차를 이해하려면 데이터 세트가 필요합니다. 통계학자는 대부분 모집단 샘플을 받는 것에 관심이 있습니다. 예를 들어 선거 투표를 사용하려면 모집단은 국가의 모든 사람들이고 표본은 통계학자가 측정하는 모집단의 하위 집합입니다. 통계에 대한 가장 좋은 점은 인구 통계를 측정하여(이 경우 전화 설문 조사 또는 이와 유사한 방법으로) 전체 인구 집단을 사용 경우 측정 값이 될 확률이 가장 높은 항목을 찾아 낼 수 있다는 것입니다. 이 통계 부분에서는 데이터 집합이 더 큰 인구수의 샘플이라고 가정합니다. 다음은 예제 입니다.

X={1 2 4 6 12 15 25 45 68 67 65 98}

간단히 기호 X를 사용하여 전체 숫자 집합을 나타낼 수 있습니다. 이 데이터 세트의 개별 번호를 참조하려면 기호 X에 아래 첨자를 사용하여 특정 번호를 나타냅니다.은 X의 세 번째 숫자, 즉 4를 나타냅니다. 은 일부 교과서에서 볼 수 있는 것처럼 가 아니라 시퀀스의 첫 번째 숫자입니다. 또한 기호 n은 집합 x의 요소 수를 나타내는 데 사용됩니다.

데이터 세트에 대해 계산할 수 있는 여러가지가 있습니다. 예를 들어 샘플의 평균을 계산할 수 있습니다. 이 부분에서는 표본의 평균이 무엇인지를 이해하고 수식만을 제공합니다.



집합 X의 평균을 나타내기 위해 기호 (“X bar”)를 확인 할 수 있습니다. 이 수식은 “모든 수를 더한 다음, 그 수를 나누십시오” 입니다. 불행히도 평균은 중간 점을 제외하고는 데이터에 대해 많은 것을 알려주지 않습니다. 예를 들어, 이 두 데이터 세트는 정확히 동일한 평균(mean)을 갖지만 분명히 상당히 다릅니다.



이 두 set에 대해 다른 점은 다른 데이터의 보급입니다. 데이터 세트의 표준 편차 (SD)는 데이터가 어떻게 퍼져 나가는지를 측정합니다.

SD의 정의는 “데이터 집합의 평균점에서 점까지의 평균 거리” 입니다. 그것을 계산하는 방법은 각 데이터 포인트에서 세트의 평균까지 거리의 제곱을 계산하고, 모든 것을 더하고 그 다음 n-1로 나누고 양의 제곱근을 취하는 것입니다.



여기서 s는 표본의 표준 편차에 대한 일반적인 기호입니다. “왜 (n-1)을 사용하는가? 왜 n이 아닌가? 답변은 약간 복잡하지만 일반적으로 데이터 세트가 샘플 데이터 세트인 경우(예: 당신은 실제 세계의 일부를 택했습니다.-선거에 대해 500명을 조사하는 것과 같음). 그러면 (n-1)을 사용해야합니다. 왜냐하면 이것이 표준 편차에 더 가까운 답을 얻었기 때문입니다. 만약에 n을 사용했다면, 전체 인구를 사용한 것 입니다. 그러나 표본의 표준 편차를 계산하지 않고 전체 모집단에 대해 계산할 경우 (n-1) 대신에 n으로 나눠야합니다.