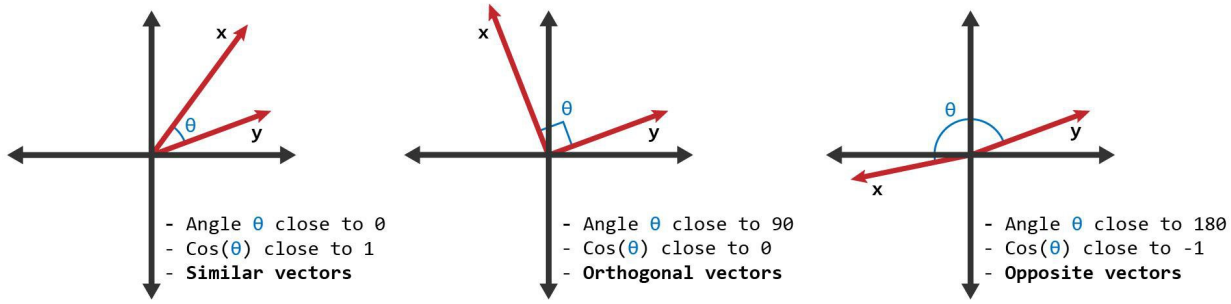
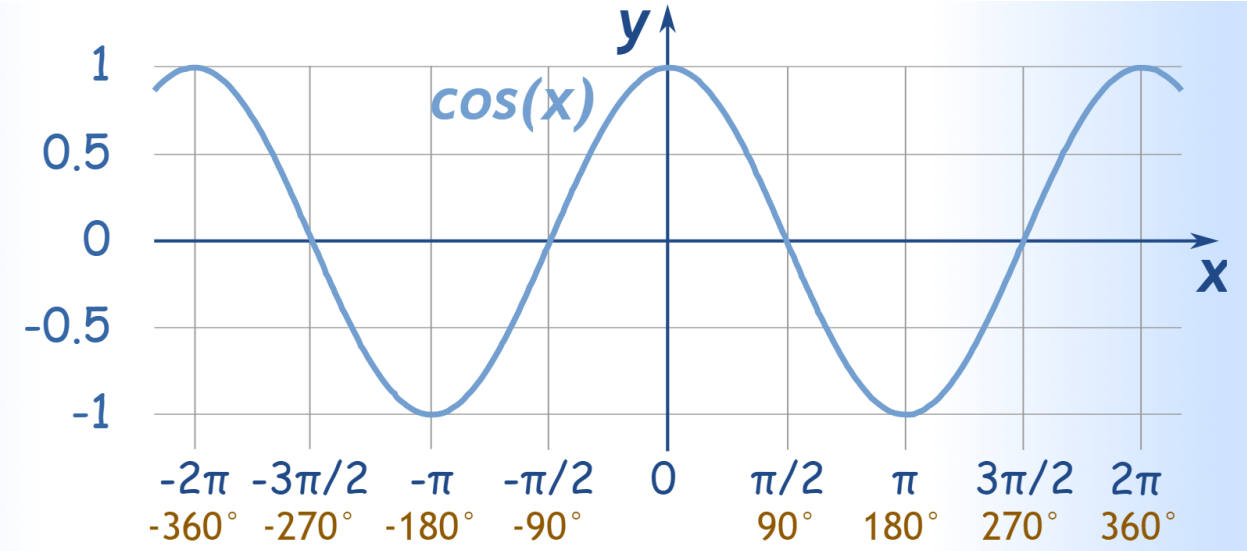
Cosine Similarity and Manifold Learning

Cosine Similarity



intuitive graph



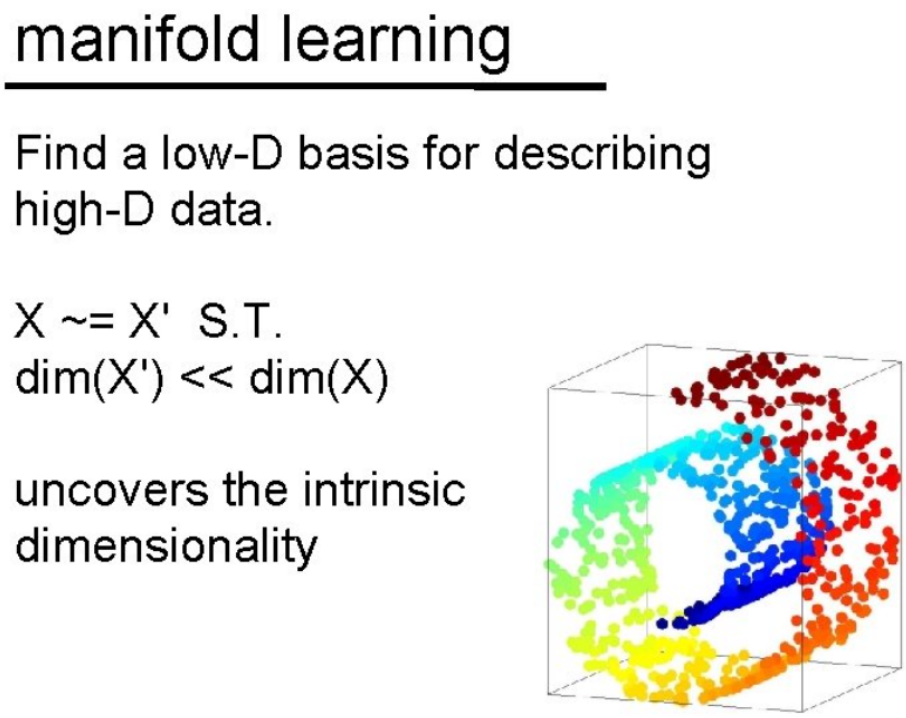
cosine graph

Cosine similarity: a metric used to measure **how similar** the documents are **irrespective of their size. (→Euclidean distance 보다 high dimensional data에서 작**

동 더 잘할 가능성)

The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, higher the cosine similarity.

Similarity = $Cos\theta = \frac{A \cdot B}{||A||||B||} = \frac{\sum_i^N a_i b_i}{\sqrt{\sum_i^N a_i^2}\sqrt{\sum_i^N b_i^2}}$



manifold learning

Manifold Learning이란?

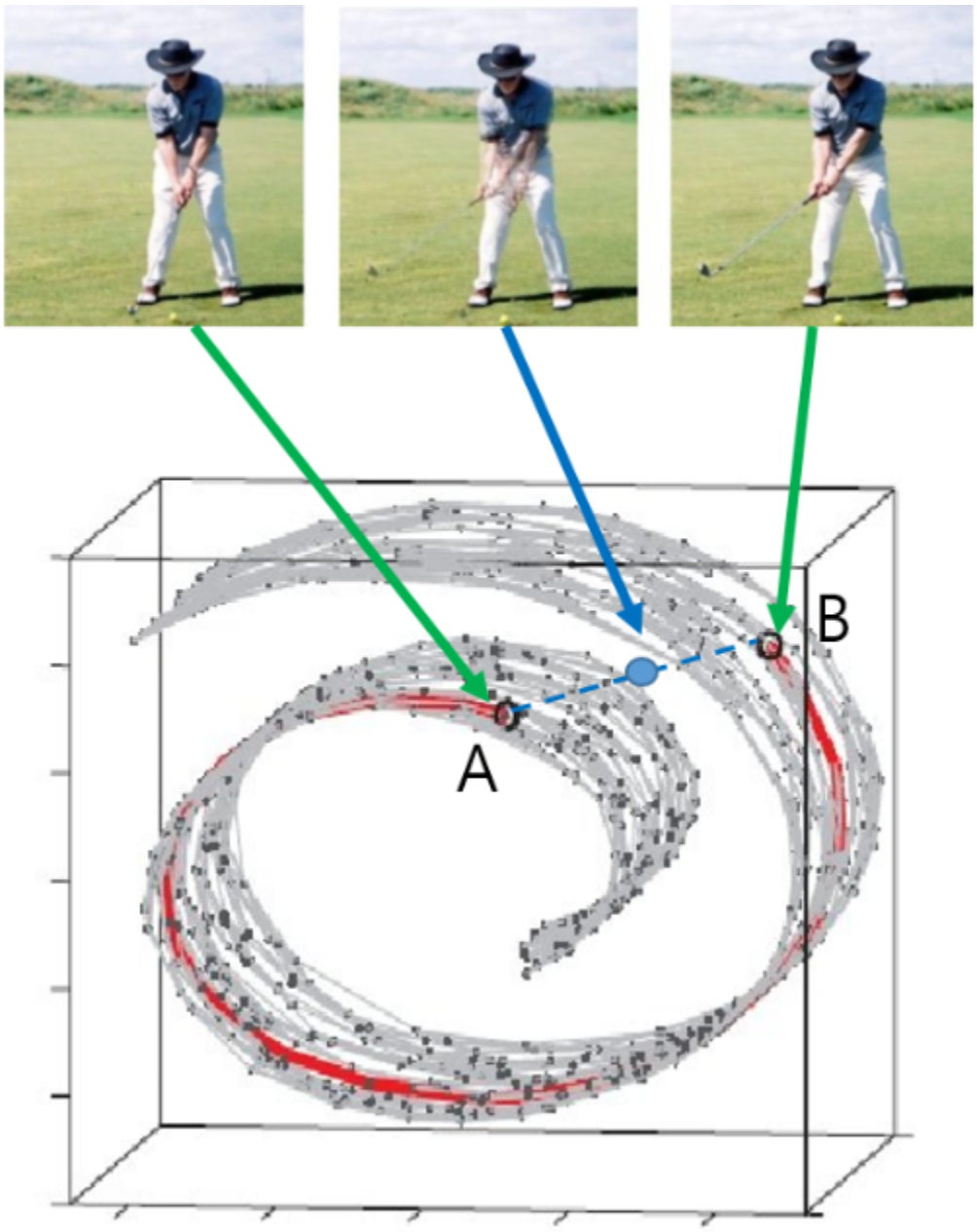
Manifold란 고차원 데이터(e.g Image의 경우 (256, 256, 3) or...)가 있을 때 고차원 데이터를 데이터 공간에 뿌리면 **sample들을 잘 아우르는 subspace가 있을 것**이라는 가정에서 학습을 진행하는 방법 →이 manifold는 데이터의 차원을 축소

필요한 이유:

- data compression
- curse of dimensionality
- de-noising
- visualization
- reasonable distance metrics

<linear interpolation: unnatural>

interpolation: image in the middle

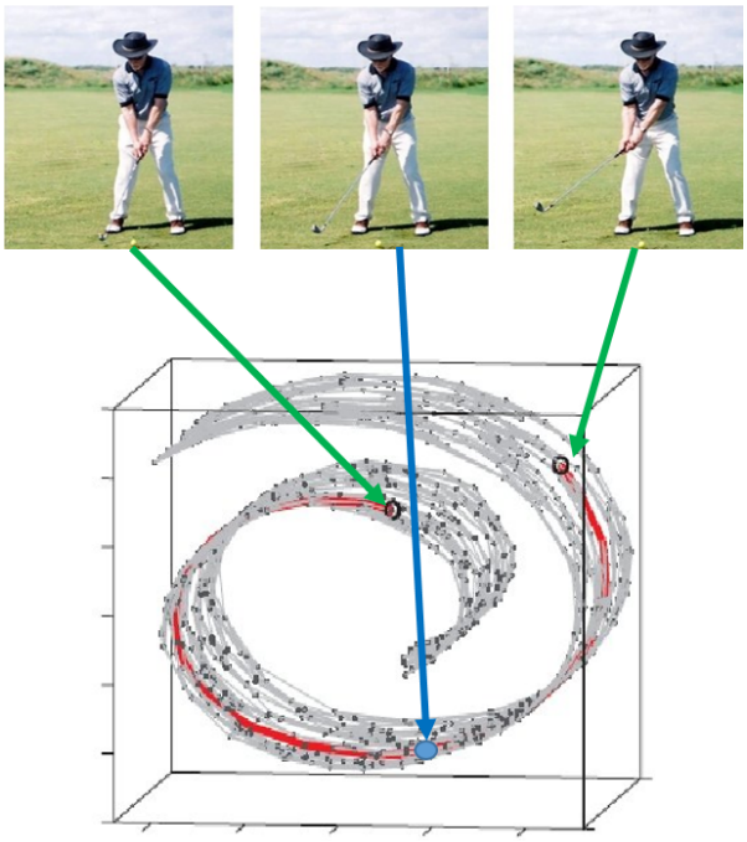


interpolation using euclidean distance (linear)

대체로 Euclidean distance는 high dimension data에선 유사함을 측정하는 데에 있어 잘 작동하지 않는다. (수가 기하 급수적으로 커짐 + noise가 포함되었을 경향이 높음)

<manifold interpolation: natural>

interpolation: image in the middle



manifold interpolation

참고:

https://www.machinelearningplus.com/nlp/cosine-similarity/

https://slidetodoc.com/image-manifolds-a-a-efros-16-721-learningbased/

https://deepinsight.tistory.com/124