ToBig's 20기 정규세션 4주차 과제

UMAP과 PaCMAP 차원 축소 기법 요약

20기 황태연

1. UMAP(Uniform Manifold Approximation and Projection)

UMAP은 차원 축소 기법 중 하나로, 리만 기하학과 대수 위상에 기반을 두는 학습 기법입니다. 차원 축소 기법은 global distance를 보존하는 PCA와 local distance를 보존하는 t-SNE가 있습니다. UMAP은 local distance를 보존하는 차원 축소 방법이지만 t-SNE보다 global distance를 더 많이 보존하면서 더 큰 데이터에 적용 가능하고, 시각화 성능에서 더 우수하다고 합니다. 또한 임베딩 차원에 계산적 제한이 없어서 머신러닝에서의 차원 축소 기법으로 활용하기에 적합합니다.

UMAP의 차원 축소 과정은 t-SNE와 상당히 유사합니다. t-SNE는 처음에 고차원의 데이터를 저차원의 공간에 랜덤한 순서로 나열한 후에 군집끼리 미는 힘, 당기는 힘이 작용하여 균형에 맞게끔 이동하는 방식으로 차원 축소를 합니다. 각 데이터간의 거리는 t-분포에 의해 정규화되고, 그 거리를 사용하여 여러 번 균형을 맞추는 작업을 반복합니다.

UMAP도 t-분포를 이용하여 균형을 맞춰가는 형태로 학습합니다. 먼저 데이터 간의 거리를 이용하여 한 데이터에 대하여 다른 데이터와의 similarity score를 계산합니다. 그러면 각 데이터 간의 similarity score가 symmetric하지 않게 되는데, 이때 평균을 적절히 활용하여 score가 symmetric하도록 만들어줍니다. 그 다음 Spectral Embedding을 이용하여 각 데이터를 저차원에 배치하고, 계산된 score와 t-분포를 적절히 활용하여 한번에 하나의 데이터가 이동하는 과정을 거칩니다.

데이터가 저차원에 임베딩될 때 t-SNE는 랜덤하게 배치되지만 UMAP은 Spectral Embedding을 사용하여 항상 똑같이 임베딩되기 때문에 차원 축소가 안정적이라는 장점이 있습니다. 그리고 t-SNE는 한번 차원 축소 과정을 반복할 때 모든 데이터가 이동하지만 UMAP은 한번에 하나씩 이동하게 됩니다. 이 점은 큰 데이터 셋에서의 차원 축소에 도움이 된다고 합니다.

2. PaCMAP(Pairwise Controlled Manifold Approximation Projection)

UMAP은 기존의 t-SNE보다 많은 성능 향상이 있었으나 여전히 global distance를 유지하는 데에 어려움이 있습니다. 그에 반해 PaCMAP은 local distance와 global distance를 모두 잘 유지할 수 있는 차원 축소 기법이라고 합니다. 그래서 PaCMAP은 시각화와 여러 실험 환경에서 높은 성능을 보여주고 있습니다.

PaCMAP에서는 local distance와 global distance 모두 잘 유지하기 위해서 새로운 DR 손실 함수를 설계했습니다. 또한 손실 함수와 관련된 그래프 구성 요소(즉, 저차원에서 어떤 데이터 쌍이 서로 밀고 끌어당겨야 하는지에 대한 요소)를 적절히 선택하였습니다.

3. 참고 자료

[1] UMAP paper: https://arxiv.org/pdf/1802.03426.pdf
[2] PaCMAP paper: https://arxiv.org/pdf/2012.04456.pdf
[3] UMAP Youtube: https://www.youtube.com/watch?v=eN0wFzBA4Sc
[4] ToBig's 20기 4주차 정규세션 강의자료
[5] PaCMAP tistory 블로그: https://slowsteadystat.tistory.com/30