Term project: SVD and PCA

Index

SID: 2017041930
NAME: 김선동

0. Introduction

1. Implement SVD for randomly generated 100-D vectors
2. Find some principal component
3. Represent 100-D vectors with the selected principal component vectors
4 Discuss the errors in representing vectors with partial basis (less than 20~30 basis vectors)

1. Principal Component Analysis

- 우선, A^TA를 구한다. A의 대한 구성 요건은 추후에 2번에서 설명하고 ATA를 구하는 이유에 대해 설명하겠다.

- ATA는 ppt에 나와 있듯이, Correlation matrix라고 한다. 즉, 서로 다른 임의의 두 데이터가 얼마나 상관되어 있는가를 확인하는 것이다.

- 결과적으로, 상관 행렬로 만들어서 차원을 축소하여 데이터의 분포나 이 분포의 주성분을 알 수 있게 된다.

- 이번 과제에선 이런 주성분을 통해서 벡터를 나타내고, 그 주성분의 중요도가 얼마나 크냐에 따라서 average vector distance가 어떻게 정해지는 지를 다룬다.

2. Generate Matrix A

- 1번의 PCA를 하려면, 우선 그를 위한 행렬 A를 생성해야 한다.

- 본 과제에서는, 100차원의 vector 1000개를 생성하게 한다. 이는 곧, 각 데이터의 변수가 100개이며, 1000개의 데이터가 주어진 상황이라고도 볼 수 있다.

- 결과적으로, A^TA를 함으로써 그 변수들이 서로 어떻게 상관 되어있는지를 알게 되는 것이다.

- 그러나 이러한 A를 마구잡이로 생성한다면 domain한 주성분을 못 뽑아낼 수 있기 때문에 일정한 조건을 부여한다. 본 과제에서는 임의의 2개의 element에 relation을 부여하고 그러한 relation 5개를 만드는 방식을 사용하였다. 그 외의 elements 또한 각각의 number range가 있어서 domain이 너무 커지지 않게 방지하였다. 5개의 relation을 어떻게 결정하느냐는 중요하지 않다. PC를 잘 뽑아내기 위해서이기 때문이다.

- 본 과제에서 A는 아래와 같은 조건들을 만족하게 하여 생성하였다.
① A는 [x_1, x_2, x_3, …., x_{100}]과 같은 100-D vector 1000개를 지닌 1000*100행렬이다.
② $x_1 = x_2$, $x_3 = 3*x_9$, $x_5 = -2*x_7$, $x_8 = -x_{10}$, $x_4 = 2*x_6$ 과 같은 5개의 relation을 사용하였다. (relation은 중요치 않다)
③ x_1~x_{10}까진 [-100, 100]을, x_{11}~x_{55}까진 [-20, 20]을, x_{56}~x_{100}까진 [-5, 5]의 range를 만족한다.

3. Singular Value Decomposition and PCA

- 2번에서 생성한 A를 SVD를 통해 100 차원의 basis vector들인 행렬 U를 구한다.

- 이러한 SVD를 통해 우린 eigenvalue의 ½제곱인 시그마들을 내림차순으로 정렬된 형태로 얻을 수 있다.

- 즉, 손쉽게 dominant한 eigenvector들을 구할 수 있게 된다.

이제껏 분석한 것들을 다시 정리해보자.

① 생성한 A 행렬은 그저 100개의 변인을 갖는 1000개의 데이터이다.

② 이 때, 이 변인들이 어떻게 연관되어 있는 지를 조사하고자 한다.

③ 변인들의 연관성을 조사한다는 것은 곧, 변인들이 어떠한 실험식으로 얼마만큼의 variance를 갖느냐를 알아내는 것이다.

Cf) 예를 들어, x1과 x5를 2D좌표에 찍어보았더니 확실한 패턴으로 $x_1+x_5=10$을 중심으로 쫙 퍼져 있다고 해보자. 이 때, 저 식을 하나하나 찍어보지 않고 어떻게 구하는가?라는 물음이 생기는데, 이는 곧 변인간의 선형성들을 systemic하게 구하는 알고리즘이 뭔가 없을까?와 같은 말이다.

3. Singular Value Decomposition and PCA – 정리(Contd.)

④ 먼저, 변인간의 상관관계를 가장 간단히 수치화하는 것은 그냥 내적을 해버리는 것이며, 그게 A^TA이다.

⑤ 내적을 했는데, 각 성분들의 크기가 어떠한 단위들이 합성되어 상관수치를 만드는 지 알고 싶은 것이다. 그 단위의 수치가 클수록 상관력이 더 강하다는 말이기 때문이다.

⑥ 이 때, 이 단위가 본 문제에선 eigenvector이며 크기는 해당 eigenvector의 eigenvalue이다.

⑦ 즉, 상관행렬 A^TA의 eigenvector와 eigenvalue를 구하면 되며, 어떤 행렬의 eigenvalue와 eigenvector를 method들 중 본 과제는 SVD(technical한 부분)를 이용한다.

⑧ SVD가 좋은 도구인 이유는, 행렬 A를 SVD하면 USV^T가 되는데, 여기서 S는 대각성분들이 전부 eigenvalue인 singular행렬이며, U와 V는 각 eigenvector를 담고 있기 때문이다.

⑨ ATA는 100*100 행렬에 noise가 있는 행렬이기에 100개의 단위(Eigenvector)로 합성이 되어있다고 볼 수 있는데, 여기서 100개의 변인들이 5개의 significant한 패턴으로 이항선형상관관계를 가진다고 한다면, 5개의 eigenvalue가 나머지 95개의 eigenvalue보다 훨씬 큰 value를 갖게 될 것이며, 이러한 significant한 상관성을 찾는 것을 PCA라고 한다.

4. Vector Representation

- 앞서 SVD와 PCA를 통해서 eigenvector들이 얼만큼의 영향력을 갖고 있는 지 알게 되었다.

- 본 과제는, vector representation을 통해서 representation에 error가 얼만큼 섞였는가를 알아보게 하였다.

- 즉, PCA를 잘 수행하였는지 검산을 하는 것을 말한다.
우선, A를 생성할 때 사용한 relation들, 5개(본 과제에선 5개를 사용하였으므로)의 가장 영향력이 큰 eigenvector들이 이 A의 significant한 특징이고 나머지는 그저 noise를 말한다. 결과적으로, SVD를 하면 가장 큰 5개의 eigenvalue를 제외한 나머지 95개는 크기가 현저히 작을 것이다.

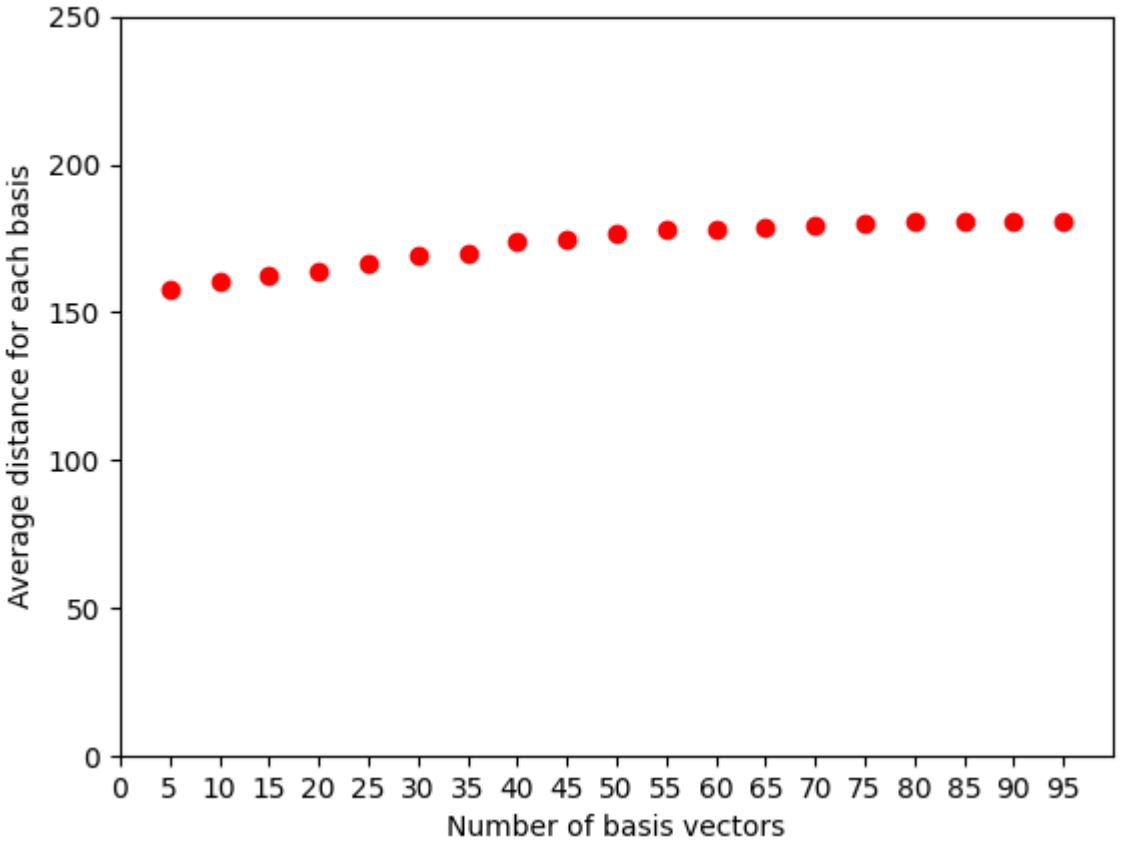
이것을 어떻게 검산하는가?
- 5개의 eigenvector들을 basis로 하는 100개의 벡터를 생성한다. 이로써 eigenvector들로 represent한 벡터를 얻는다.(represented vector)

- 이렇게 생성된 벡터들과 generated한 벡터의 각각의 거리를 구하고 그들의 평균거리를 구한다.

- 이를 추출한 eigenvector를 10개, 15개, 20개, ..., rank(A)개 까지 구하는 것이다.

5. Result and Analysis

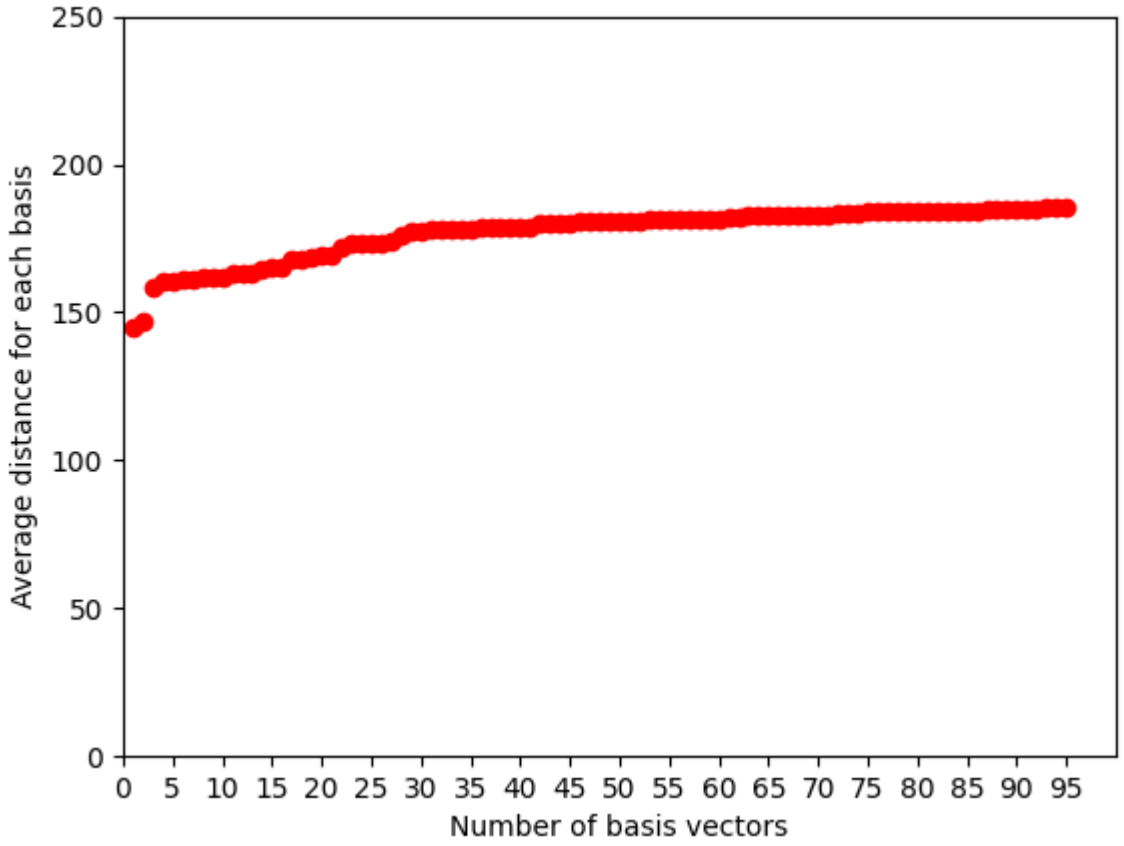
- 위의 ppt의 과정을 토대로 그래프를 만들어보면 아래와 같다.



- 과제 요구사항대로 5개부터 5개씩 rank(A)까지 찍어보았는데, 우리가 예상한 대로 결과값이 출력되었다.

5. Result and Analysis – 정리(Contd.)

- 하지만, 왜 생성한 relation이 significant하며 그것이 그래프에 어떻게 표현되는 지를 알려면 basis vector 가 1개부터 rank(A)까지 represent할 때의 그래프를 그려보는 것이 중요할 것 같아서 그려보았다.



- 이렇게 보면, 우리가 분석한대로 5개까지는 dramatic하고 그 이후부터는 그리 큰 변화 없는 것을 볼 수 있다.

- 만약 relation을 7개, 10개, 15개를 설정했다면 어땠을까? 아마도 7개, 10개, 15개까지는 값이 dramatic하다가 해당 값 이후부터는 격차가 그리 크지 않을 것이다.

감사합니다!