15wk-2: 기말고사

1. SVD

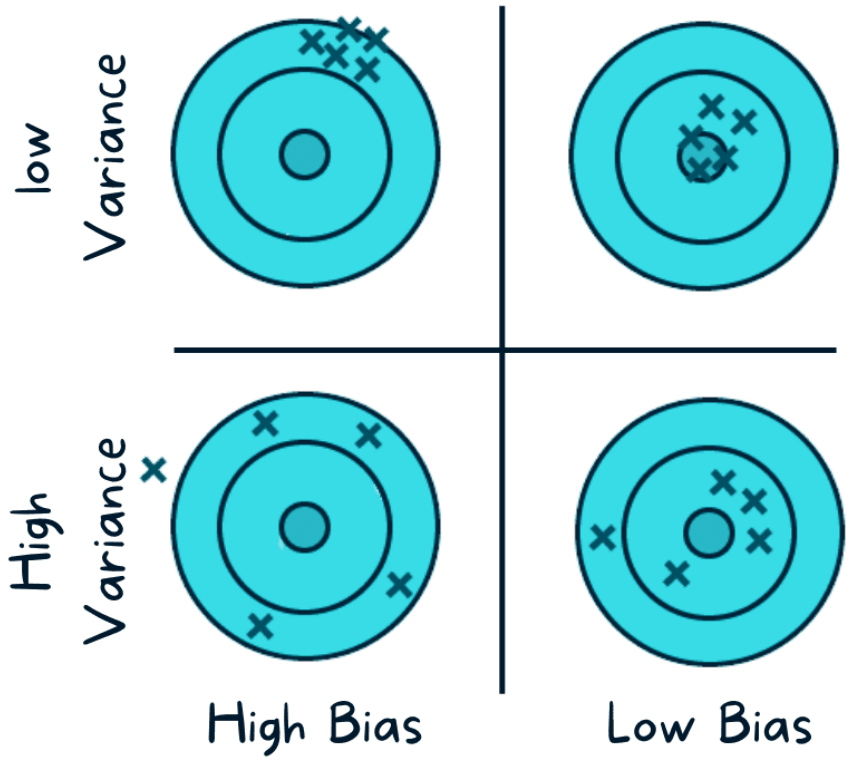
통계학에서 SVD는 어떻게 활용될 수 있는가? 활용분야를 목록화하고 간단히 서술하라.

- 이미지 압축 : 이미지 데이터를 분해해서 중요한 정보만 남긴다.
- 노이즈 제거 : 데이터에서 노이즈를 제거해서 중요한 신호를 복원한다.
- 이미지 처리 : 이미지를 저차원 공간에 매핑해서 패턴을 인식한다.

2. 다중공선성

토익,텝스,학점을 설명변수 $\mathbf{X}_{n\times 3}$ 로 설정하고 이를 바탕으로 연봉 $\mathbf{y}_{n\times 1}$를 추정하고자 한다. (이때 학생들의 토익과 텝스점수는 서로 비슷하다고 가정한다. 즉 토익점수가 높은 학생은 대체로 텝스점수도 높으며, 반대의 경우도 그러하다고 가정한다) 다음을 잘 읽고 물음에 답하라.

(1) 선형회귀를 사용하여 계수(토익,텝스,학점이 연봉에 미치는 영향)를 추정하고자 한다. 이러한 상황은 그림1에서 무엇과 관련이 있는가? 왜 그렇다고 생각하는가?



ref: https://www.kdnuggets.com/2022/08/biasvariance-tradeoff.html

- 분산이 높고 편향이 낮은 오른쪽 하단 그림에 해당한다.

토익 점수와 텝스 점수가 비슷한 관계를 보이고 있어서 추정값들을 모아서 분산을 구하면 분산이 매우 크기 때문이다.

(2) 능형회귀를 이용하여 계수를 추정한다고 하자. 여기에서 λ는 어떠한 역할을 하는가? 그림과 연관시켜 설명하라.

- λ는 계수 추정값들의 절댓값이 크게 나오는 경우 손실함수를 더 크게 만들어 주는 역할을 한다.

손실을 적게 하는 것이 중요하니까 람다를 크게 하면 해가 (0,0)으로 가게 되고, 너무 작게 하면 원래의 해와 똑같이 나오므로 적절한 λ값을 정해야 한다.

- λ를 키울수록 추정량의 분산은 줄어들고 편향은 커져서 왼쪽 위 그림에 가까워진다.
- λ를 키울수록 추정량이 작은 값을 가지고 분산은 높아지고 편향이 줄어든다. 오른쪽 하단 그림과 가까워진다.

(3) 주성분 회귀 (Principal component regression, PCR)을 이용하여 계수를 추정하고자 한다. 이때 principle componet 수를 작게 설정할때와 크게 설정할때 어떠한 일이 생기는지 설명하라.

- 주성분 개수를 적게 설정하면 차원 축소가 많이 일어나서 복잡도가 줄어든다. 데이터의 손실을 일으켜서 편향이 커지게 되고, 작아진 복잡도 때문에 분산이 줄어든다.
- 주성분 개수를 많게 설정하면 차원 축소가 적게 일어나서 거의 모든 데이터를 사용해 복잡해진다. 많은 데이터가 있어서 편향은 감소하게 되고, 복잡도가 커져서 분산이 커진다.

(4) 능형회귀에서 $\lambda = 0$ 으로 설정하거나 $\lambda = \infty$로 설정하는 것이 어떠한 의미를 가지는 주성분 회귀와 연결시켜 설명하라.

- λ가 0이면 선형회귀와 같아진다. 주성분 분석에서 모든 주성분을 사용하는 것과 비슷하다.
- λ가 커질수록 회귀선의 기울기가 점점 낮아지면서 0까지 내려간다. 계속 커지면 회귀계가 0이 된다. 주성분을 매우 적게 설정하는 것과 같다.

3. 면접 질문?

(1) 능형회귀에 대하여 간단히 설명하라.

- 선형회귀에 정규화를 추가해서 다중공선성 문제를 해결하고 일반화 능력을 향상시킨다. 모델의 예측 성능을 높이고 과적합을 방지한다.

(2) 다중공선성이란 무엇이며 어떤 문제를 일으키는 간단히 서술하라.

- 다중공선성 : 회귀 분석에서 사용된 모형의 일부 설명 변수가 다른 설명 변수와 상관관계가 커서 데이터 분석 시 부정적인 영향을 미치는 현상

다중공선성이 높으면 반응변수에 대한 설명변수의 설명력이 낮게 해석되어 분석에 오류가 생길 수 있다.

(3) $\mathbf{X}_{n\times p}$, $p > 2$ 일 경우 $\mathbf{X}$를 시각화하는 방법에 대하여 간단히 서술하라.

- 주성분 분석을 사용한다.

분산을 최대화하는 새로운 축을 찾아내고 이를 이용해 새로운 좌표로 변환한다.

데이터를 표준화한다 -> 표준화된 데이터의 공분산 행렬을 계산한다 -> 공분산 행렬의 고유값과 고유벡터를 계산한다 -> 고유값이 큰 순서대로 주성분을 선택한다 -> 주성분으로 데이터를 변환한다

(4) 직교변환이 가지는 의미를 간단히 서술하라.

- 어떤 벡터에 직교행렬이 변환으로 적용되면, 그 벡터와 크기는 각도가 보존된다.

열 벡터에 적용되면 $||X_1||^2 = ||AX_1||^2$ 이므로 크기가 보존되고, $X_1X_2 / ||X_1||||X_2|| = (AX_1)(AX_2) / ||AX_1||||AX_2||$ 이므로 각도가 보존된다.

예를 들어 원점 대칭인 경우 -> 회전 시켜도 점들이 원점에 대한 거리가 똑같다

행 벡터에 적용되면 $||x_1^T||^2 = ||x_1^TA||^2$이므로 크기가 보존되고, $x_1^Tx_2^T / ||x_1^T||||x_2^T|| = (x_1^TA)(x_2^TA) / ||x_1^TA||||x_2^TA||$이므로 각도가 보존된다.

예를 들어 원점 대칭인 경우 -> 우산의 양 끝 점이 원점과 이루는 각도가 대칭을 시켜도 항상 같다. 임의의 두 점에서 성립함.

직교변환이 데이터를 의미하는 행렬 X 뒤에 곱해질 경우 각 관측치의 크기 및 각도가 모두 보존된다. 따라서 데이터 X에 적당한 직교변환을 해서 Z를 얻었다면 $Z = XA$, $A^TA = I$가 성립한다.

(5) $\mathbf{X}$가 이변량 정규분포를 따른다고 가정하자. $\mathbb{V}(\mathbf{X})$의 고유벡터행렬을 활용하는 통계적 처리기법을 있는가? 있다면 서술하라. (하나만 서술해도 무방)

- 주성분분석 : 차원 축소에 이용하는 방법이다. 데이터의 분산을 최대화하는 새로운 축을 찾는 방법이다. 공분산 행렬의 고유벡터는 데이터의 주요 방향을 나타내고, 고유값은 그 방향의 분산을 나타낸다.

원래의 자료 X를 그대로 분석하는 것이 아니라, X를 Z로 바꾼뒤에 분석하는 일련의 기법(즉 X의 주성분을 분석하는 기법)을 통틀어 주성분 분석이라고 한다.

(6) SVD를 이용하여 이미지를 압축하는 방법을 간단히 서술하라.

이미지를 행렬로 변환한다

-> 행렬을 SVD를 통해 분해한다. $A = UDV^T$

-> 가장 큰 k개의 특이값만 남긴다. 나머지는 무시한다.

-> 압축된 이미지를 얻는다.

-> 축소된 U, D, V^T로 이미지를 복원한다.

(7) 주성분분석을 하게 되면 얻게되는 이점을 간단히 서술하라.

- 차원 축소를 통해 데이터에 대한 이해가 쉬워지고, 분산값을 유지하면서 정보 크기를 줄이기 때문에 데이터 특성 훼손 없이도 연산 속도가 개선된다.

(8) 선형변환을 SVD를 이용하여 해석하라.

- A를 선형변환한다는 것은 V^T로 회전변환 한다는 것이고 D로 스케일 변환을 하고 U로 회전변환을 하는 것이다. 결국 V는 선형변환 전 행렬이고 U는 선형변환 후 행렬, D는 특이값을 가진 행렬이다.

(9) 변환을 의미하는 행렬 $\mathbf{A}$가 데이터를 의미하는 행렬 $\mathbf{X}$의 앞에 곱해지는 경우와 뒤에 곱해지는 경우 각각 어떠한 의미를 가지는지 설명하라.

- 변환을 의미하는 행렬A가 데이터를 의미하는 행렬X 앞에 곱해지는 경우, A는 X의 열별로 적용되는 어떠한 선형변환을 의미한다.

-> $AX = [AX_1\ AX_2\ \dots\ AX_p]$

- 변환을 의미하는 행렬A가 데이터를 의미하는 행렬X 뒤에 곱해지는 경우, A는 X의 행별로 적용되는 어떠한 선형변환을 의미한다.

-> $XA = [X_1^TA\ X_2^TA\ \dots\ X_n^T]^T$

(10) R(*lm()*)과 Python(*sklearn.linear_model*)에서 더미변수가 포함된 회귀분석을 수행하는 로직이 다르다. 차이점에 대하여 서술하라.

- R은 범주형 변수를 더미변수로 변환하지 않더라도 자동으로 인식해서 가변수 처리를 한다. (factor 또는 character형...)
- 파이썬은 get_dummies() 함수를 사용해서 범주형 변수를 명시적으로 더미변수로 변환해주어야 한다.