4. 비지도 학습

- 데이터 전처리
  - 일부 알고리즘은 데이터의 스케일에 매우 민감함.
  - 민감한 알고리즘 예 : 신경망 계열, SVM, ...
  - 스케일에 영향을 받지 않는 계열 : 결정 트리를 기반으로 하는 알고리즘들

- 전처리 종류
  - StandardScaler : 평균 0, 분산 1로 변환
  - MinMaxScaler : 0 ~ 1 사이의 값으로 변환
  - RobustScaelr : 평균대신 중간값, 분산대신 4분위값(q1, q3) 사용. 이상치 데이터 처리에 유리.
  - Normalizer : 특성벡터의 유클리디안 길이를 1로 조정. 방향만 중요한 데이터에 유용.

- 예제
  - 암데이터셋에 MinMaxScaler 를 적용(cell 4 ~ cell 9)
  - 주의) 훈련세트로 조정된 scaler를 테스트세트에도 적용해야 함.
  - 전처리 없이 SVC 알고리즘 적용결과(cell 11) : / 0.63(테스트세트 정확도)
  - MinMaxScaler 적용후 SVC 적용결과(cell 12) : / 0.95(테스트세트 정확도)
  - StandardScaler 적용후 SVC 적용결과(cell 13) : / 0.97(테스트세트 정확도)

4. 비지도 학습 – PCA

- 비지도 변환(차원축소)
  - 데이터를 새롭게 표현하여 분석자나 머신러닝 알고리즘이 원래 데이터보다 쉽게 해석할 수 있도록 만드는 알고리즘.
  - 특성이 많은 고차원 데이터를 특성의 수를 줄이면서 꼭 필요한 특징을 포함한 데이터를 표현하는 방법

- 군집화
  - 데이터를 비슷한 그룹으로 묶음.
  - 예) 사진의 주인공을 모르는 상태에서 여러 사진 중 같은 인물의 사진으로 분류

- 이슈
  - 레이블이 없기때문에 알고리즘의 성능평가 어렵다.

- 용도
  - 데이터를 더 잘 이해하고 싶을 때 탐색적 분석단계에서 사용.
  - 지도학습의 데이터 전처리 단계로 사용
  - 새롭게 표현된 데이터를 사용해 학습하면 지도학습의 정확도 높아지거나 자원 절약 가능.

4. 비지도 학습 - PCA

- 주성분 분석 PCA(Principal Component Analysis)

  - 특성들이 통계적으로 상관관계가 없도록 데이터셋을 회전시키는 기술
  - 데이터셋에서 가장 유용한 성분만을 추출하여 데이터셋의 차원을 축소하는 기법

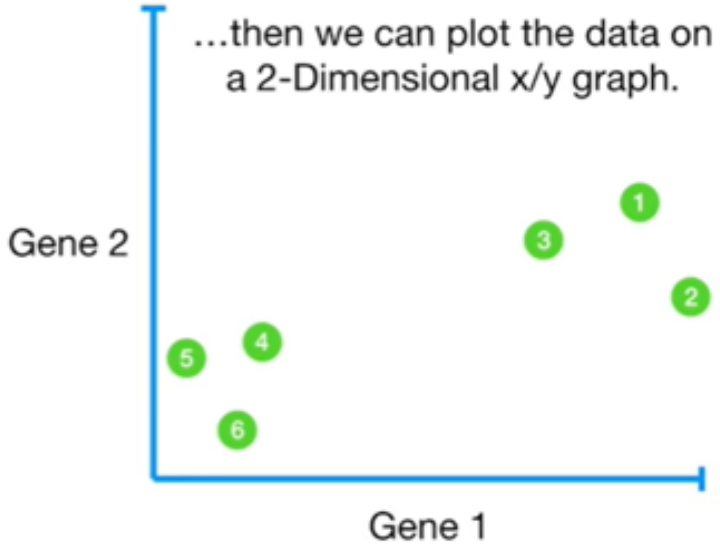
- 용도 및 장단점
  - 특성간에 상관관계가 클 때 이를 제거하기 위한 목적.
  - 데이터를 더 잘 이해할 수 있다.
  - 자원소요를 줄이고 학습속도 향상
  - 최대분산 방향이 학습 성능을 향상한다는 보장이 없다.
  - 비선형 데이터에는 적용 불가.

  - 간단한 PCA 이용한 차원축소/복원(cell 14)
  - 1) 두개의 주성분 방향 찾기(분산이 가장 큰 방향 = 정보를 가장 많이 담고 있는 방향)
  - 2) 원점을 중심으로 회전
  - 3) 차원축소 : 첫번째 주성분만 나타냄(y = 0)
  - 4) 복원 : 첫번째 주성분만으로 복원

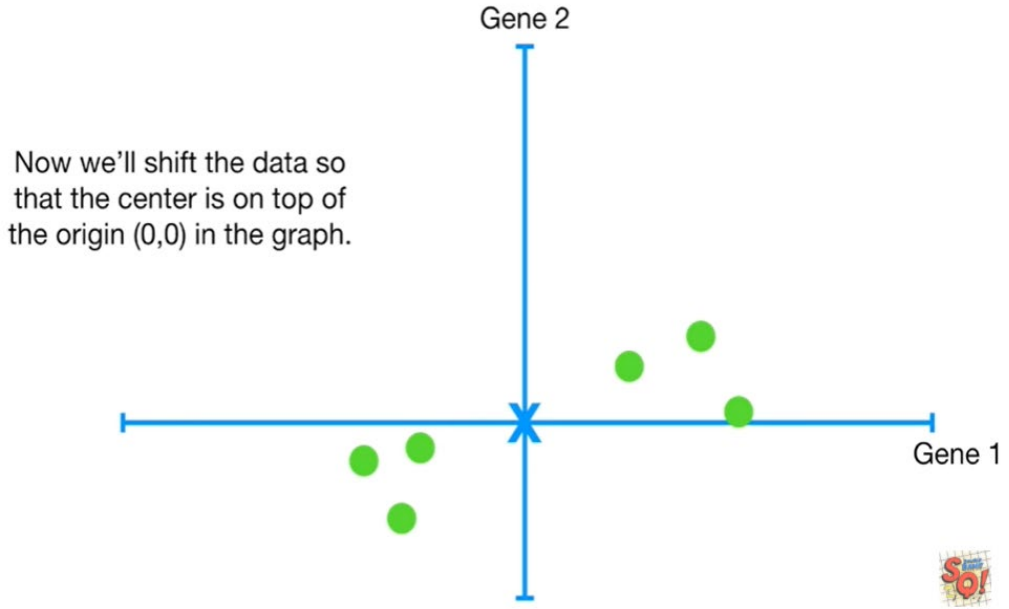
4. PCA

데이타셋 : 두개의 유전자로 표현된 쥐

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



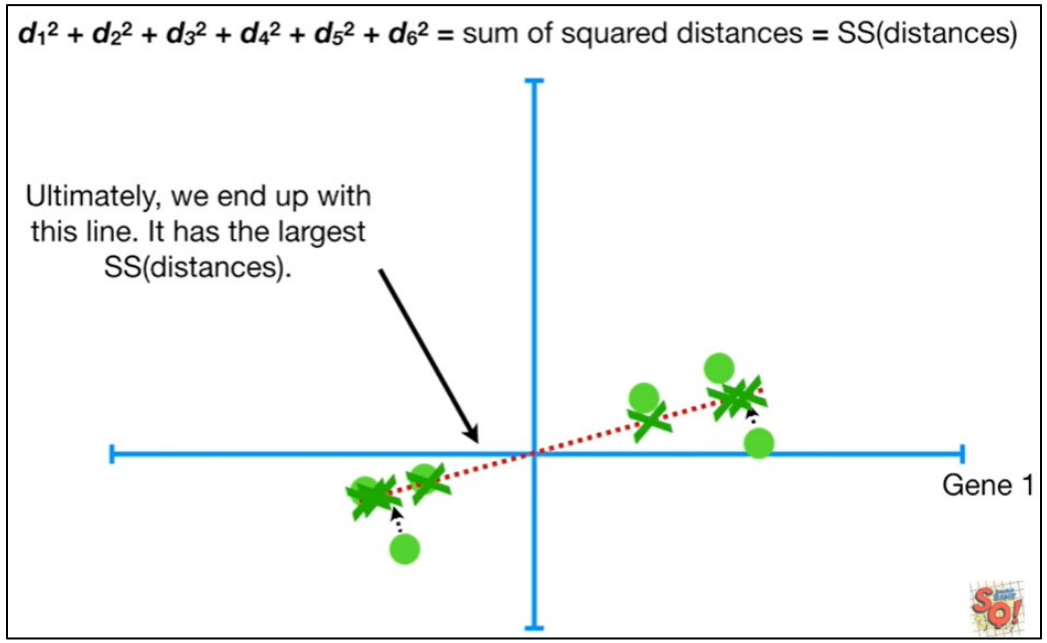
1. 데이터를 원점을 중심으로 이동



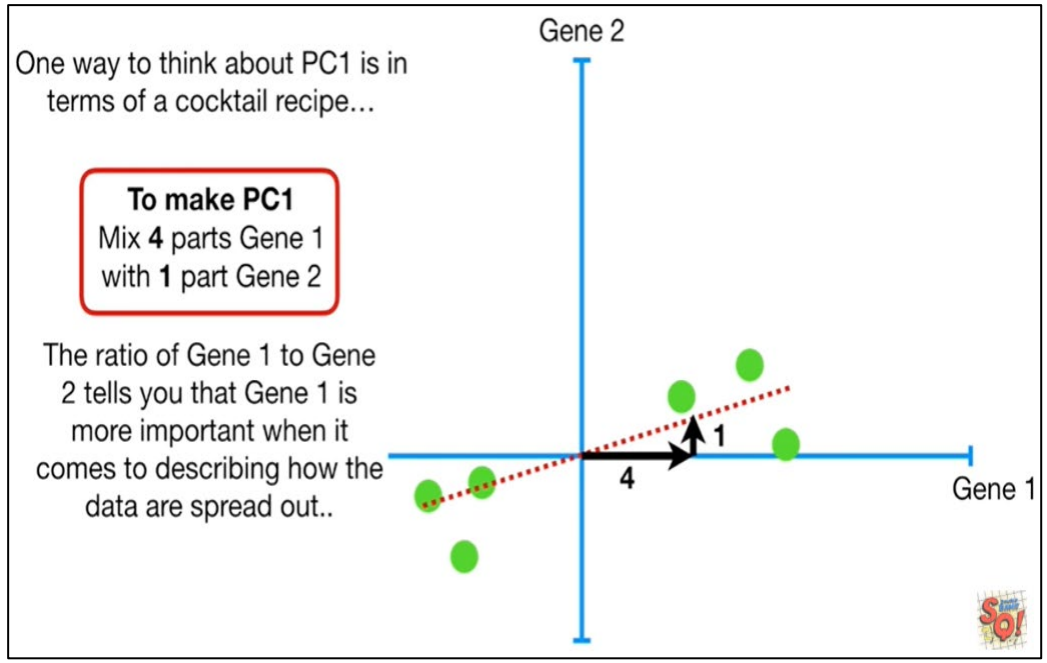
https://www.youtube.com/watch?v=FgakZw6K1QQ&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=28

4. PCA

2. SS(sum of squared distances) 최대가 되는 직선 선택
   ✓ SS : 원점과 데이타포인트 사이의 거리



3. PC1 생성 –
   ✓ Gene1이 Gene2 에 비해 4배 중요도가 있음.



4. PCA

4. PC2 생성 –

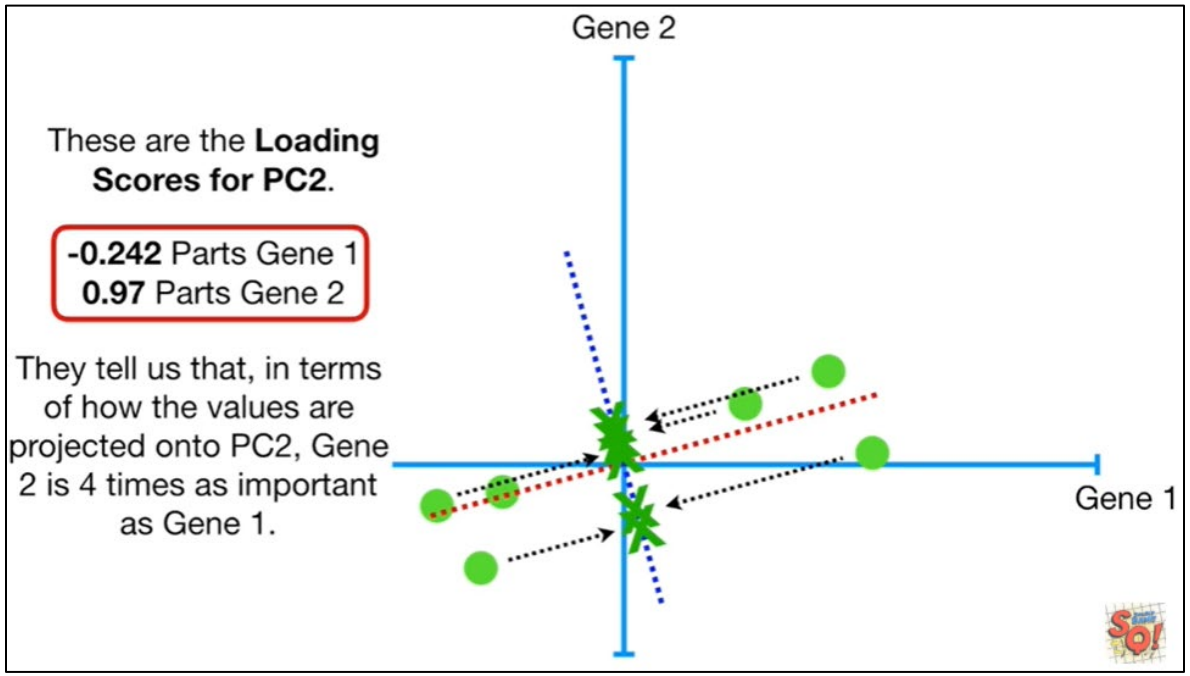
✓ PC1과 직교하는 축. PC1에 누락된 정보 포함.

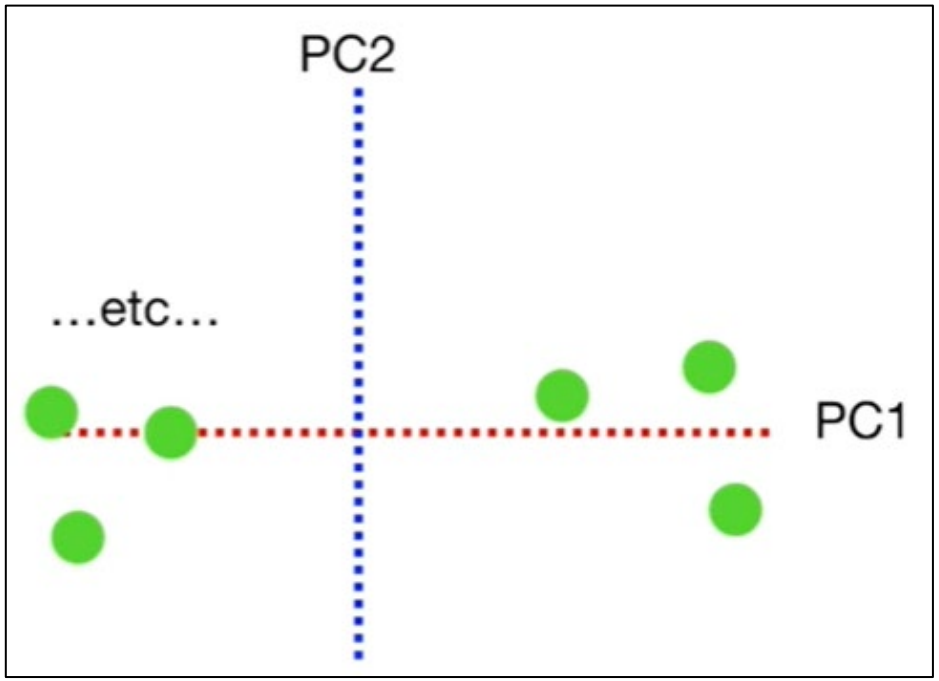
✓ Gene2는 Gene1 에 비해 4배 중요도가 있음.

✓ Gene1의 분포는 ¼로 반영됨.

5. PC1. PC2 축을 기준으로 회전 –

✓ 같은형태이면서 PC1 방향의 분산이 최대가 되도록 변환됨.





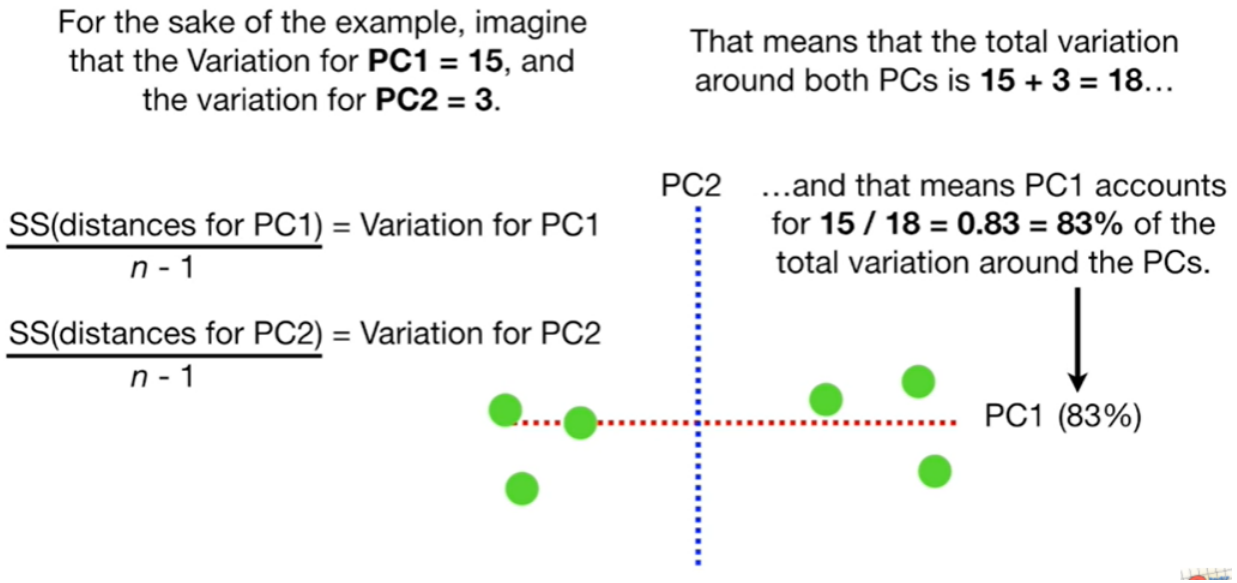
4. PCA

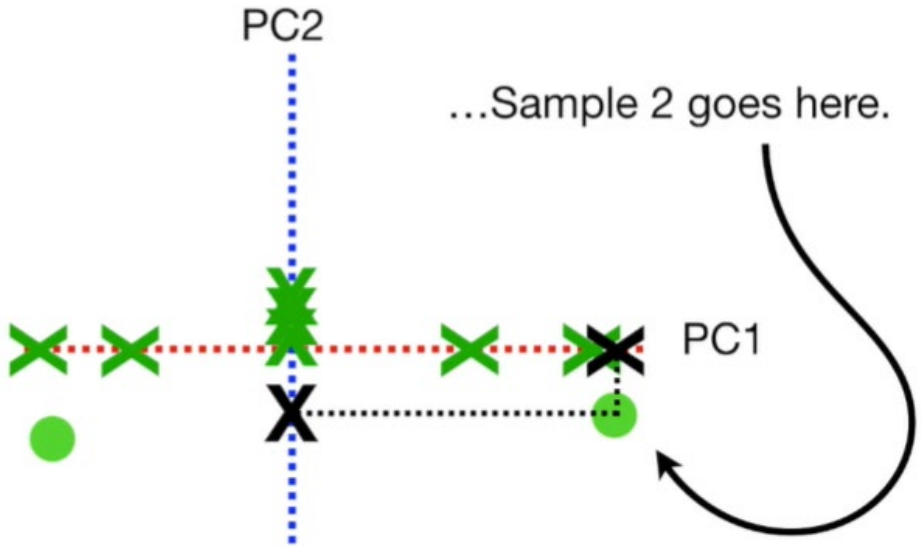
6. 해석

✓ PC1 분산값 : 15, PC2 분산값= 3 이라면,

✓ PC1은 데이터 정보의 83%를 보유함.

즉, PC1 으로 데이터를 축소하면 데이터 정보의 83%를 사용.

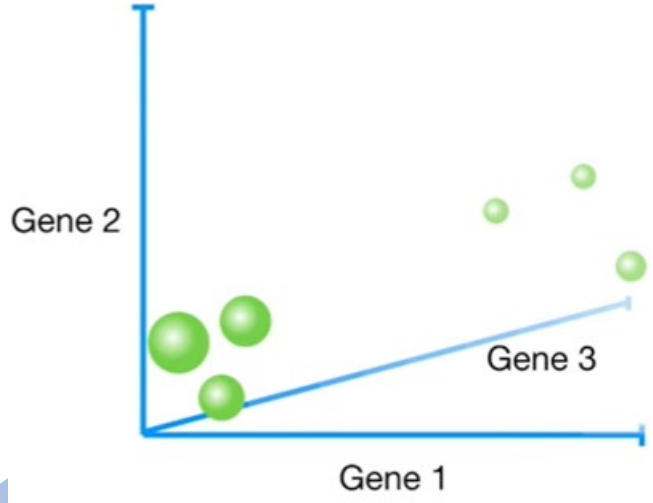




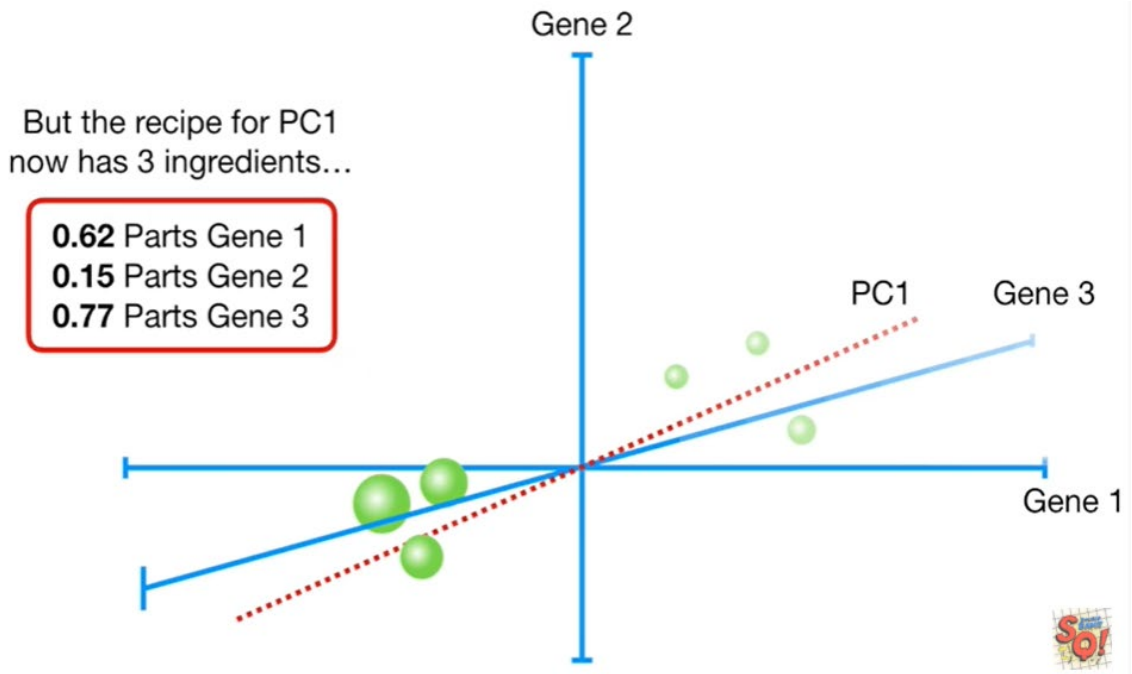
4. PCA

특성이 3개인 데이타

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2

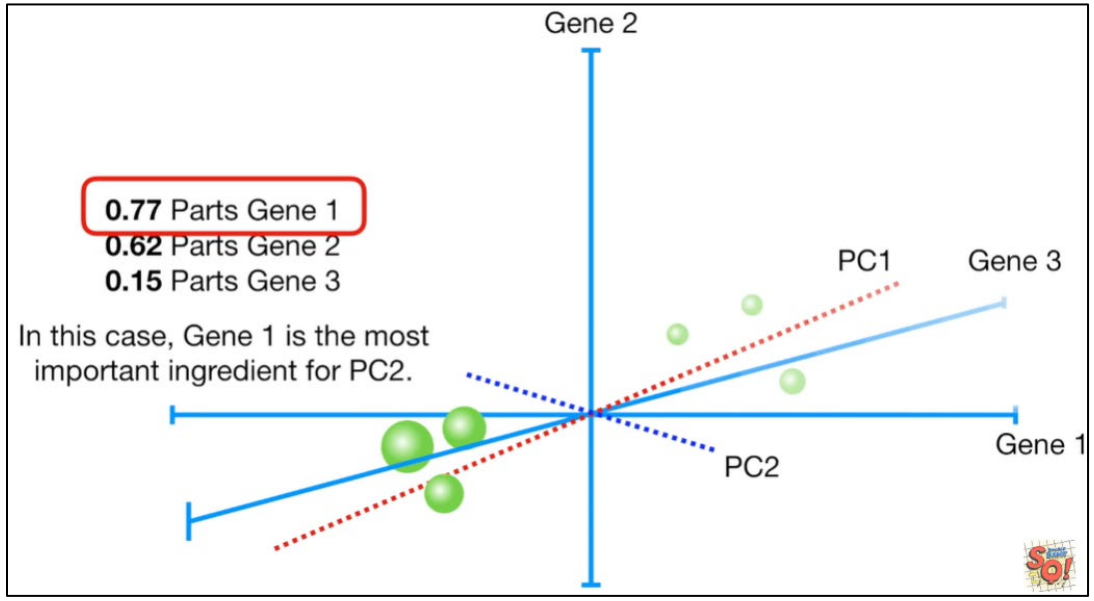


1. 첫번째 주성분 찾음.
   - ✓ SS 가장 큰 축 계산 → PC1
   - ✓ 분산기여정도를 Eigenvector로 계산하면, Gene3이 가장 큰 기여, Gene2가 다음.

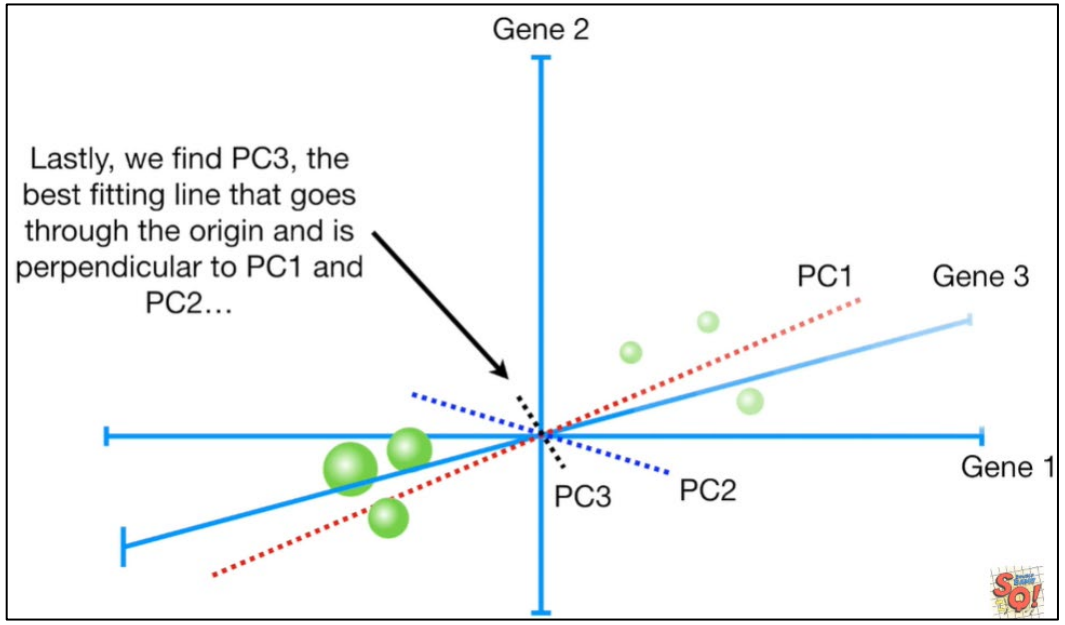


4. PCA

2. 두번째 주성분 찾음.
   - ✓ PC1과 직교하면서 SS 값이 큰 축
   - ✓ Gene1 이 가장 큰 기여, Gene2가 다음.

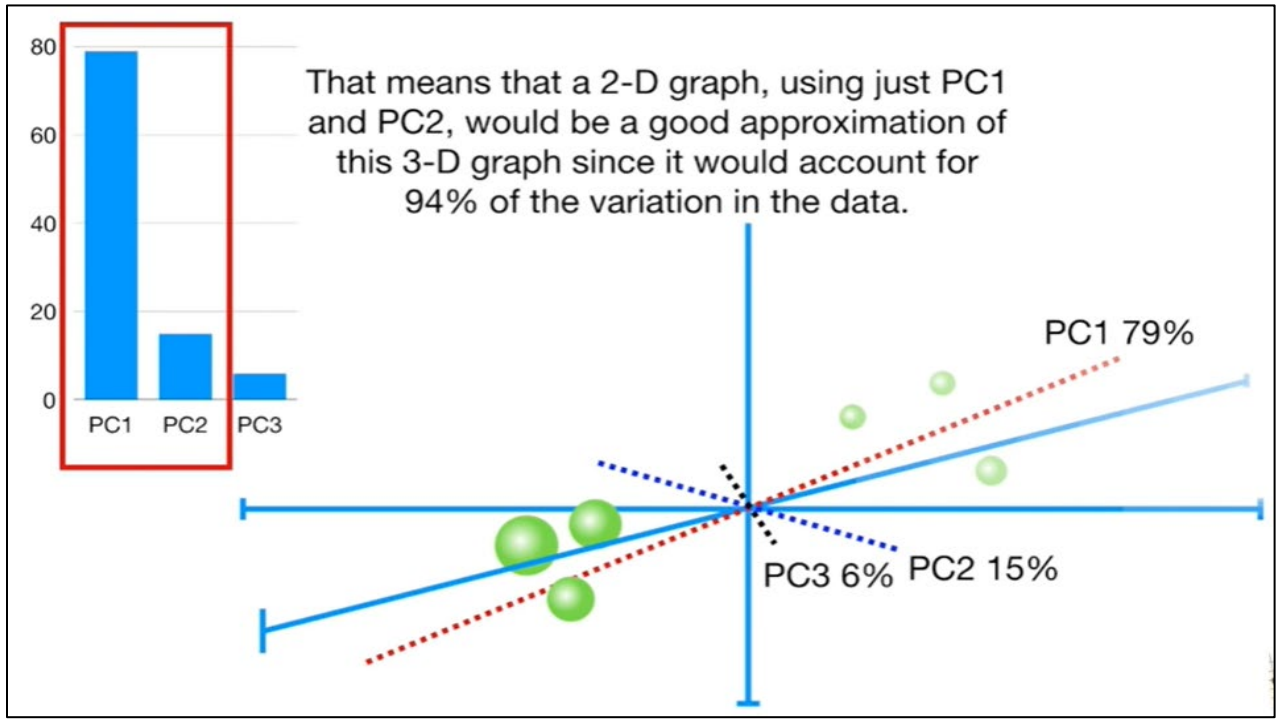


3. 세번째 주성분
   - ✓ 남은 한 축.

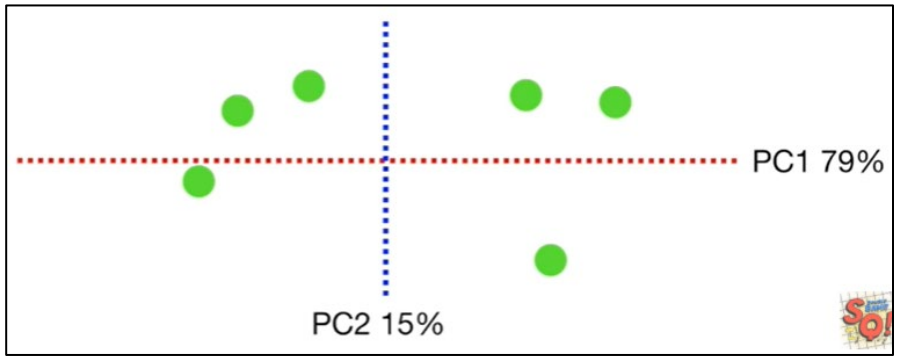


4. PCA

4. 분산값 분석에 의한 PC별 중요도 확인.
   ✓ PC1, PC2가 데이터 분산의 94%를 보유하고 있음.



5. PC1, PC2 만을 사용하여 데이터 변환.



4. 비지도 학습 – PCA

- 데이타셋 시각화

  - 시각화방법 1 : 두개의 특성씩 짝짓는 산점도 행렬(예 : 붓꽃 산점도 행렬)
  - 암데이타셋은 특성이 30개이므로 산점도로 파악하기 쉽지 않음.
  - 대안으로 양성/악성 클래스에 대해 각 특성의 히스토그램 그려봄(cell 15).
  - 특징별 보유 값을 히스토그램 bin으로 정의하고 양성/악성별로 출현횟수를 나타냄.
  - mean radius, mean area, mean perimeter, worst perimeter 등 분류와 관련있어보이는 특성 확인 가능.
  - 그러나, 특성간의 상호작용이 클래스에 미치는 영향은 표시하지 못함(개별특성은 무관해보이지만 혼합특성은 관련있는 경우 등).

- PCA 이용한 데이타셋 시각화
  - 처음 두개의 주성분을 시각화(cell18)
  - pca() 호출시 유지할 성분갯수 지정(n_components=2)
  - (선형 분류모델을 적용할 수 있을정도로 구분됨)
  - 문제는 생성된 축의 해석이 쉽지 않다는 점.
  - pca.components_[] : 생성된 축에 30개의 특성이 기여한 정도.