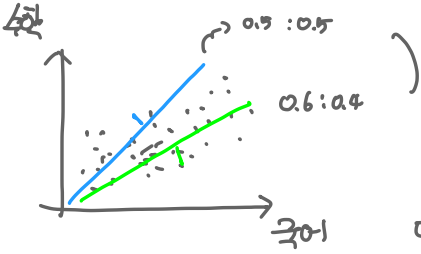
PCA (Principal Component Analysis)

=> 데이터 분포에서 어떤 선분으로 정사영 시켰을때 데이터 분포 잘 표현 되었나? (국어 영어 중요성을 반영 비율 (벡터) 찾기



=> 이렇게 변형 비율 다르

5:5로 변형한거 6:4로 반영한거

데이터 분포따라 무엇이 잘 반영한건지 다르다! 최적의 projection 할 line 찾기

↳ 2차원 영역은 2개의 벡터를 찾아야 한다

principle component (주성분) = data의 분산이 가장 큰 방향 벡터 (잘 퍼짐)

ex) 얼굴인식, 1800개 사진 변화 1800개 주성분 벡터중 분산이 큰 20개 뽑아서 사용 (얼굴 수직 얼굴 전반적 형태, 뒤얼수록 세부적)

K-mean clustering

=> 키와 몸무게 에 따라 옷을 생산 할때, 모든 사이즈 고려할수 없어 S.M.L 로 사이즈 나눌때 이용가능 하다.

* Centroid · Cluster 의 중심

데이터 분포, k 정함. Centroid 설정

while (변함 없을)

모든 데이터 Centroid 와 유클리드 거리로 분류

Centroid 를 Cluster 된 데이터의 중간으로 다시 설정

다시 새로운 Centroid 로 모든 데이터 유클리드 거리로 분류

=> K 최적 값을구 단점

Soft max regression

=> 로지스틱 회귀를 multi classification 이다.

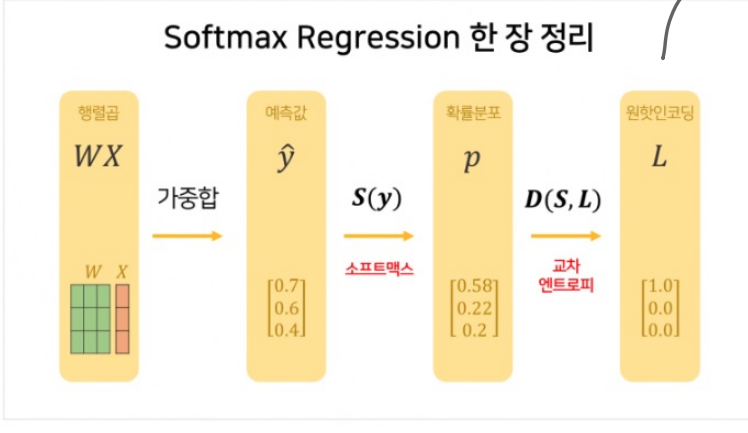
(로지스틱은 성공/실패 2가분류, soft max는 A, B, C 처럼 3개이상 classification)

- 확률 총합인 1의 확률 분포로 바꾸기 위해 이용 $S(y)$로 표현 되고 이를거쳐 확률분포 p로 바뀐다

$$S(y_i) = \frac{e^{y_i}}{\sum_j^k e^{y_i}}$$

전체 그림

1 2 3. 가중치 다르니 0과 1로 가중치 없애고 공평하게 효율적임!



SVM (Support vector machine)

=> margin 을 최대화 (class 간 분류) 하는 선을 긋는것

* Support Vector = 선과 가장 가까운 point
* margin = Support vector 와 선의 distance
* decision boundary = 두 data 구분하는 선
* robust 하다 = outlier의 영향을 덜 받는다.

e.g.

	1	2	3	4	5
	1	2	3	4	100

	mean	median
	3	3
	22	3
	↓	↓
	not robust	robust

(outlier에 취약. 정확)

-> 무조건 margin이 큰순이 분류 잘하지 않음

그래서 원 데이터를 분류하는 범위 원리 찾고

그 범위 안에서 margin 최대화 하는선 찾음

but 목표가 outlier 있는 경우 어느정도 outlier 무시하고 선찾음

-> Gamma 와 decision boundary 휘어짐. 작다. 직선에 가깝다.