과제명 : 지식 그래프 기반 음식점 광고 콘텐츠 추천 알고리즘 개발
(Title)

과 제 번 호 :
(Project No.)

5. 벡터의 유사도 (Vector Similarity)

1) 코사인 유사도 (Cosine Similarity)

$$\text{Similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

DTM을 구한 후 Numpy를 사용해서 코사인 유사도를 계산하는 함수로 유사도를 구한다.

* 코사인 유사도는 벡터의 방향에 초점을 두므로 코사인 유사도는 문서의 길이가 다른 상황에서 비교적 공정한 비교를 할 수 있도록 도와준다.

2) 여러가지 유사도 기법

-1. 유클리드 거리 (Euclidean distance)

$$\sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}$$

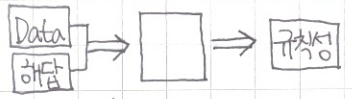
-2. 자카드 유사도 (Jaccard similarity)

자카드 유사도 함수 : J

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

6. 머신러닝 (machine learning)

1) 머신러닝 방식



2) 머신러닝 훑어보기

-1. 머신러닝 모델의 평가

Original data = Training + Validation + Testing

검증용 데이터로 과적합(overshooting)을 판단하며 모델의 정확도를 높이기 위해 하이퍼파라미터를 tuning.

-2. 분류 (Classification) 와 회귀 (Regression)

(1) 이진 분류 ex) 합격과 불합격

(2) 다중 클래스 분류 : 3개 이상의 분류

(3) 회귀 : 연속적인 값의 내에서 예측 값이 나오는 경우.

-3. 지도학습과 비지도 학습

(1) 지도학습 (Supervised Learning)

: 레이블(Label) 이라는 정답과 함께 학습하는 것

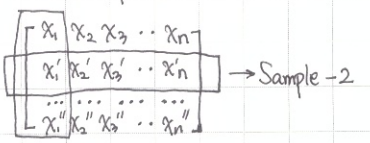
(2) 비지도 학습 (Unsupervised Learning)

ex) LSA , LDA

(3) 자기지도 학습 (Self-Supervised Learning, SSL)

레이블이 없는 데이터가 주어지면, 모델이 학습을 위해서 스스로 데이터로부터 레이블을 만들어서 학습하는 경우

-4. 샘플 (Sample) 과 특성 (Feature)



-5. 혼동행렬 (Confusion Matrix)

	예측 참	예측 거짓
실제 참	TP	FN
실제 거짓	FP	TN

(1) 정밀도 : True라고 예측한 것중 True 인 것.

$$= \frac{TP}{TP+FP}$$

(2) 재현율 (Recall) : 실제 True 중 True로 예측한 것.

$$= \frac{TP}{TP+FN}$$

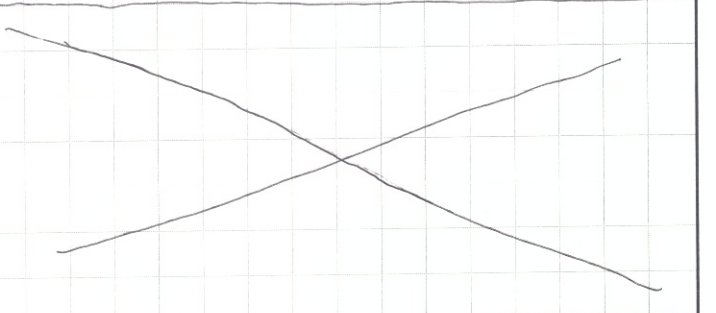
(3) 정확도

$$= \frac{TP+TN}{TP+FN+FP+TN}$$

-6. 과적합 (Overfitting) 과 과소적합 (Underfitting)

Training Data로 너무 많이 학습하면 정확도가 떨어지게 된다. → 과적합

정확도를 더 높일 수 있음에도 적게 학습하는 경우. → 과소적합



기록자 (Recorded by)	서명 (Signature)	일자 (Date)	
이상준	이상준	2022.4.2	이 노트의 지적재산권은 숭실대학교 산학협력단에 있음 (Intellectual property right for this notebook belongs to Foundation of Soongsil)
점검자 (Reviewed by) 정선태	서명 (Signature)	일자 (Date) 2022.4.5	