1. 데이터를 불러 들이고
2. 긍정/부정 분류기를 만들 것 -> 그중 10% 정도만 테스트 셋으로 돌릴 것
3. 파싱 – 의미 있는 시퀀스 만들기

단순히 워드 카운트\_마케팅 분야

엔엘피는 피할 수 없음

그 문장에 있는 단어의 발생수를 라이클리 후드로

곱이 안될 때는 로그를 씌워주면 됨. (그럼 곱이 합으로 바뀜!)

각 피쳐가 가지는 중요성\_ 전체 데이터 셋에서 스케

검색에서만 쓰는 게 아니라 주제 찾을때도 쓰임!- TF\_IDF

TF\_IDF+나이브 베이시안

단순히 텀 프리퀸시만 보는 것이 아니라 단어의 중요도는 전체 문장에서 그 단어가 얼마나 중요도가 있는지 분모를 통해서 노말리제이션 하는 것임!

단어마다 웨이트를 줘야함 – 정관사 더 가 문장내에 중요도

그게 아이디 에프 -> 전체 데이터 셋에서 해당 단어가 나오는 수 ( 전체 데이터 셋/ 해당 단어가 나오는 횟수)

1. 티에프를 구한 다음에
2. 아이디에프를 구해서 곱한다!-> 아이디에프에 로그를 취하는 이유는 작게 스케일링 하기 위해서 임!
3. 티에프 아이디에프를 라이클리 후드로 쓰는 것
4. 전체 데이터에서가 아니라 각 문장 별 티에프 아이디에프를 구하는 것이기 때문에 휠씬 강력함!

학습 시키는 것은 데이터를 불러와서 핏팅 하고 트랜스폼 하는 것( 사이클론의 철학이기 때문에 학습의 방법은 같음!)

결론적으로 카운트 보다 더 정확하 이유는 각 단어의 웨이트도 곱하기 때문임!

핏팅을 시키고

슈퍼 바이즈드 러닝이기 때문에 인풋과 라벨이 필요함!

모델.프리딕트 값을 넣으면 됨!

학습 자체는 핏이고 트랜스 폼은

뉴머릭한 데이터를 나이브 베이시안을 하고 싶을 때는 가우시안을 쓰면 됨!!

트랜스폼: 핏은 트레인 데이터 데이터를 핏팅 시켜서 객체를 반환

핏된 데이터는 다 가지고 있음

-선형회귀 모형

오버 피팅/언더 피팅 굉장히 중요한 개념!

다 중요함! 라소랑 릿지, 제너랄리 제이션 정말 중요!

머신 러닝에서는 함수/가설을 찾는 것임( 데이터를 가장 찾을 수 있는 가설을 찾는 것!)

설명변수는 통계학에서 많이 쓰임

데이터에서 직선을 구하는 방법 -> 로스 펑션! (최소 제곱 법\_선형회귀의 로스 펑션)

제가 찾고 싶어하는 직선이 있고 직선만 나오면 어떤 엑스를 넣어도 그 엑스에 대응하는 와이를 찾을 수 있음-> 버티칼 디스턴스를 구함( 그 선과 데이터 간에)-> 그 디스턴스가 작을수록 좋음

* 정당성! 최소 제곱법을 로스 펑션으로 쓴다는 정도만 알아 둘 것! (로스 펑션을 쓴다. 최소 제곱법을 이용해서\_점과 선의 사이로 구하지 않음!)
* 엠에스이나 알엠에스이를 쓰지 않는 이유는?
* 거리의 차이를 최소화 하는 것이 핵심임!

-노말 이큐에이션 : 미분한 다음에 0이 되는 값을 찾는 것.(세터 제로랑 세터 1에 미분을 해야함. 그 나머지는 상수임 따라서 세터 제로와 세터 1의 방정식으로 나옴)

제곱해 준 것(미분한 것)을 다 더하면 됨.

편미분

(최대 최소 구할 때) 미분하고 0이 나온 지점을 구하는 것!

노말 이큐에이션이 어떻게 구해지는 지 알아 둘 것/ 결론이 중요함!

* 노말 에큐에이션은 선형회귀를 통계학적으로 푸는 것!( 머신러닝에서 선형회귀를 구할때는 그래디언트 디센트 를 씀)
* 엑스의 값이 크면 클수록 값을 구하기 어렵다!
* 그래디언트 디센트

코스트 펑션: 거리 최소!

편미분한 코스트 펑션 \* 러닝 레이트를 준만큼 업데이트 함!

초기 값을 설정하고 이것을 계속 업데이트 하는 것!

-배치 그래디언트 디센트

세터를 업데이트 할 때 -> 몇번 업데이트 하는 지가 중요함

데이터 셋이 총 엔 개 있다고 하면 전체 엔 개에 대해서 한번 그래디언트 디센트를 구하면 됨!(한번 트레이닝 셋을 전체 흟고 업데이트 하고\_모든 데이터를 훓음-> 그래서 백프로 글로벌 미니멈(옵티멈)에 수렴함 (왜 전체 데이터에 업데이트를 치기 때문에\_그래서 정규 방정식의 해와 동일하게 됨)

코스트가 많이 듬

-스토카스틱 그래디언트

정확도는 떨어지나 속도는 굉장히 빠름/왔다갔다 하면서 글로벌 미니멈으로 수렴함!

좋은 세터를 구할 수 있겠지만 옵티멀 값은 아니다

로컬 미니멈에 들어가도 다시 빠져나올 수 있는 장점은 있음 (랜덤성이 있으니까, 왔다갔다, 적당히 좋은 해로 수렴할 수 있다)

-미니 배치 그래디언트(배치+스토카스틱)

그만큼만 그래디언트를 업데이트 시켜줌.

두 그래디언트의 장단점을 다 가지고 있음. wlsWk aldksgo bb

학습 속도 빠르고/랜덤성이 줄어들고/ 조금 더 안정적임!/글로벌 옵티멈이 아닐 수 있다

\*그래디언트는 선형회귀 뿐만 아니라 로스 펑션에서도 사용함!

\*글로벌 옵티멈이 나오는 것은 배치 그래디언트랑 노말 이큐에이션임!

- 선형회귀 모형의 확률 인터프리테이션( 단골 인터뷰 문제임)

이거다!!!!! Aldksgo bbb