FEED FORWARD NEURAL NETWORKS FOR THE ANALYSIS OF CENSORED SURVIVAL DATA: A PARTIAL LOGISTIC REGRESSION APPROACH

**Introduction**

ANN에 대한 소개(multi-layer) due to the back-propagation

prior information의 부족으로 적용에 대한 어려움을 겪었다

* consider flexible modelling approaches

ANN이 non-linear multivariate problem에 적합한 flexible models이라고 여겨짐

-> 의료 분야에서 적용되게 됨

-> 하지만 관측 중단 데이터에 대한 부분을 지적하는 논문이 있다

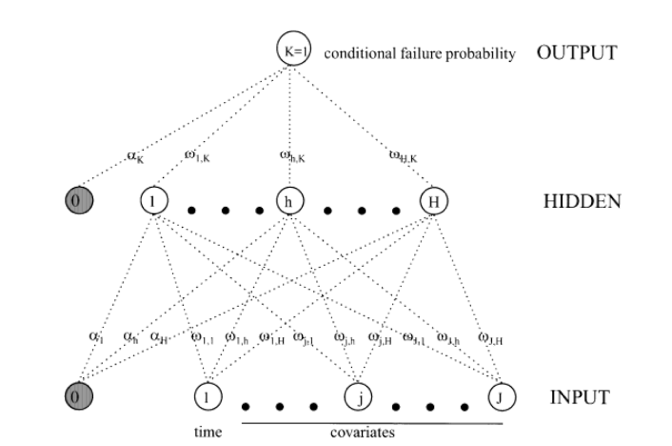
Continuous time data에 대해 Liestol이라는 사람이 constant hazard approach를 제안했고, Faraggi and simon이 ANN predictor로 linear proportional hazard Cox model까지 만들었다. -> 우리가 그냥 당연하게 썼지만 이게 다 censored data를 처리할 수 있도록 만들어져 있던 것

거기서 중요한 issue는 조건부 확률(예후 인자; 결과값)을 구하는 것이다

The aim of the present paper is the proposal of a flexible ANN approach in a discrete survival time context, which provides smoothed hazard function estimation and allows for non-linear covariate effects.

ANN모델(non-linear generalization of logistic regression failure time으로 그룹화된

제안한다! 접근법이 partial likelihood와 관련 됐으므로 이걸 PLANN



안에 들여다 보면 이렇게 생겼어~

Input layer, hidden layer -> fully connected

Hidden layer -> 1 output

Sec 2 -> 통계학적인 framework를 ANN 모델에 적용해보고

Sec 3 -> survival 분석을 위한 discrete model에 대해 설명해줄거야

Sec 4 -> logistic regression의 non-linear generalization이 모델에 어떻게 적용될 수 있는지에 대해 설명해줄거고

Sec 5. > 우리가 만든 ANN model을 previous proposals과 비교할 거야

Sec 6 -> 모델들을 두가지 example을 갖고 비교해볼거야

+ example data에 대한 설명

**SEC 2**

우선 feed forward ANN은 non-linear multivariate regression methods랑 똑같은 의미

Input layer는 그냥 hidden layer에다가 값 분배해주고 hidden layer의 각 노드는 weight과 bias를 구해준다 -> activation 적용해서 결과 값을 얻는다.

(Activation function은 이제 logistic one을 얻게 해줌)

그렇게 logistic regression에서 나타나는 수식이 partial likelihood를 최소화하는 것과 에러를 최소화 하는 기준이 됨

Feed forward ANNs with logistic outputs can be regarded as flexible non-linear regression models for conditional probability estimation -> 뭘로 이렇게 말할 수 있는지는 모르겠음

**SEC 3**

생존함수에 대한 설명

+

i번째 subject는 주어진 시간 간격에 대한 조건부 확률의 곱으로 나타난다… 베르누이 분포로 나타낼 수 있는 이유는? -> 이산적인 결과값

**SEC 4**

에러 구하는 부분 logistic regression에서처럼 구하고 이게 cross entropy

모든 에러값(loss function)은 n명을 l만큼의 시간 간격만큼 보았을 때

하는 말이 결국 sec1에서 말했듯이,

“ ANN에 logistic regression 가져다 이용한건데 어떻게 이용했냐면

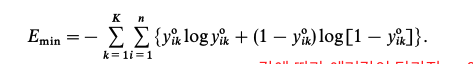
Logistic function을 hidden layer와 output layer의 activation function으로 사용해서 결과값을 뽑았고, 시간이랑 다른 covariate 들이 input으로 들어가기 때문에 시간에 대한 가정은 따로 필요하지 않다.

그리고 그에 대한 결과 값으로 시간에 의존하는 위험 함수와 covariates를 분석하는데 쓰일 수 있다. 이러한 장점으로 time dependent covariate를 쉽게 사용할 수 있다.





위에서 아래 빼면 아래와 같이 나온다



여기에다가 Kullback—Leibler distance 이거 이용해서 grouped and non-grouped cell dataa의 deviance를 계산한다

(쿨백-라이블러 발산은 두 확률분포의 차이를 계산하는 데에 사용하는 함수로, 어떤 이상적인 분포에 대해, 그 분포를 근사하는 다른 분포를 사용해 샘플링을 한다면 발생할 수 있는 정보 엔트로피 차이를 계산한다)

+

Weight decay

NIC

**SEC 5**

다른 ANN 모델들은 multiple output을 갖는다

but smoothed estimates of the hazard function are not directly obtained from the network. Instead PLANN is proposed for the flexible modelling of the hazard function.

Censored data + event indicator에 내용

PLANN is directed to model conditional event probabilities; when data are organized as we proposed in this paper, the subsequent estimation of survival probability is straightforward from relationship

결과는 ppt에서

**Discussion**

Feed forward artificial neural networks represent a particular class of non-linear regression models so, in the statistical framework, they are mainly used for conditional probability estimation in pattern recognition

근데 이게 GLM의 non-linear 버전이랑 거의 동일하게 볼 수 있다. 앞에서 본 것과 같은 이유로! (link function과 activation function and Loss function)

Censored survival time은 몇명 죽은 시점으로 정할 수 있다.

ANN은 multivariate modeling에서 더 좋고, discrete & continuous에 대한 cox모델을 일반화함으로써, censored survival data를 모델링하는 것에 proposed된다.

그래서 PLANN이 straight-forward modeling 할 때도 그렇고 flexible model로도그렇고 여러모로 좋다.

ANN이 coeff. 해석하기는 힘들어도 모델의 성능 높이는 것에는 효과를 보인다.

근데 단점도 있는데; overfitting, long computational times, presence of sub-optimal minima in the error function

+ 두 예시에 대한 내용…. + 선형모델로 결과 좋으면 선형모델 써라