알고리즘 설명 보고서

안녕하세요, 저희는 팀 '대체로 맑음'입니다.

저희는 아래의 글을 통해 제출한 코드와 알고리즘에 대한 이해를 돕고자 합니다. 코드에 대한 구체적인 설명은 깃허브와 제출한 알고리즘 코드에 자세하게 기술해 놓았습니다.

https://github.com/mostly-sunny

본론으로 들어가기 전에, 이번 해커톤에 참가하면서 많은 경험을 할 수 있었고, 이러한 기회와 데이터를 제공해주신 모든 분들께 감사의 말을 드리고 싶습니다. 그러면 지금부터, 본격적으로 저희가 제출한 코드에 대한 설명을 해보도록 하겠습니다.

글의 차례는 다음과 같습니다:

가) 전체적인 알고리즘의 설계

나) 비선형 Cox regression Model을 사용한 모델 구성

다) 교란 변수 문제의 해결

라) 답의 도출 과정

마) 참고문헌

가) 전체적인 알고리즘의 설계(1단계, 2단계, 3단계)

저희가 구성한 알고리즘은 크게 3개의 단계로 구성되어 있습니다.

1단계: Cox Proportional-Hazards Regression을 기반의 기계학습

저희는 데이터를 받고 가장 먼저, 어떤 변수가 종속 변수이고, 어떤 변수가 독립 변수인지 판단하는 과정을 가졌고, 저희가 받은 데이터들 중 '생존시간'만이 종속 변수이고, 나머지 변수들(유전자 변이, 임상 변수, 치료 여부)은 모두 독립 변수라고 판단하였습니다. 여기서 주목해야하는 점은 생존 시간에 사망 여부를 적절히 반영해야 Right censored data의 문제를 해결할 수 있다는 것이었습니다. 저희는 종속 변수가 생존 시간으로 유일하고, 연구 기간 동안의 사망 여부에 따라 우편향 될 수 있는 데이터라는 점에서 생존 분석을 해야한다고 판단하였습니다.

다음 데이터에 적절한 생존분석 방법을 고민하던 중, 주어진 데이터 중 시간에 따라 변하는 값이 없고, 생존기간이 특정한 분포를 따른다는 가정을 할 수 없는 상태에서 기계학습을 통하여 생존과 관련된 여러 인자들의 영향을 동시에 판단하기 위하여 Cox Proportional-Hazards Regression[1][2]을 기반으로 모델을 설계하였습니다. 모든 인자가 Linear한 관계를 가지고 있지 않을 것이라는 가정에 Nonlinear Cox regression을 적용한 Deepsurv 모델을 참고하여 코드를 작성하였습니다.[3]

우선, 주어진 데이터 중, 생존시간이 음수인 값을 가진 환자 데이터는 정보 수집에 오류가 있었다고 판단하여 제거하고 진행하였습니다. 또, 임상변수의 값이 문제에서 0에서 9까지로 설정되었지만 10과 11의 값을 가진 환자 데이터가 존재하여, 이 또한 제거하고 진행하였습니다.

다음으로, 유전자 데이터와 치료 여부의 데이터의 경우, 0 과 1의 이진 데이터로 주어져 있는 반면, 임상 변수 데이터의 경우 0~9까지의 값으로 주어져 있어 최대최소법을 적용하여 모든 데이터가 0과 1 또는 그 사이의 값으로 Scaling을 해주었습니다.

위의 데이터 처리 과정을 거친 후 유전자 데이터 300개, 임상 변수 데이터 10개, 치료 여부 데이터를 모두 합쳐, 311개의 열과 966개의 행을 가진 데이터가 준비 되었습니다. 이 데이터를 [학습 데이터: 테스트 데이터]를 [7:3]으로 나누고, 다시 학습 데이터를 [훈련 데이터: 검증 데이터]를 [8:2]로 설정하여 학습시켜 HR(Hazard Ratio, 위험비)를 얻을 수 있는 모델을 제작하였습니다.

2단계: 직접 가공한 데이터를 사용하여 '치료효능률' 계산

다음으로 저희는 직접 가공한 데이터를 사용하여 위험비 사이의 비율을 새롭게 계산하였습니다.

위의 1단계에서 총 311개의 데이터(유전자 데이터 300개, 임상변수 데이터 10개, 치료 유무)를 독립변수, 생존시간을 종속 변수로 하는 비선형 모델이 완성되었고, 다음으로 저희는 각 유전자의 변이가 있을 때 치료를 진행했을 때와 진행하지 않았을 때의 위험비가 얼마나 변하는지를 계산하기로 하였습니다. 저희가 새롭게 정의한 '치료효능률'의 정의는 아래와 같습니다.

$$\text{치료효능률} = \frac{(\text{유전자 } A\text{에만 변이가 있고 치료하였을 때의 위험비})}{(\text{유전자 } A\text{에만 변이가 있고 치료하지 않았을 때의 위험비})}$$

따라서 저희는 311개의 열(유전자 변이 데이터 300개 + 임상변수 데이터 10개 + 치료여부)의 형식을 가지고 각 행마다 하나의 열만 1, 나머지는 0으로 둔 가상 환자 데이터를 새로 제작하여 만들어진 모델에 넣었습니다. 새로 저희가 제작한 데이터는 아래와 같이 구성되어 있습니다:

	유전자 데이터 1~300	임상 변수 데이터 10개	치료여부
1 번	1000000~0000	0,0,0,0,0,0,0,0,0,0	0
2 번	0100000~0000	0,0,0,0,0,0,0,0,0,0	0
3 번	0010000~0000	0,0,0,0,0,0,0,0,0,0	0
~	~	0,0,0,0,0,0,0,0,0,0	0
300 번	0000000~0001	0,0,0,0,0,0,0,0,0,0	0
301 번	1000000~0000	0,0,0,0,0,0,0,0,0,0	1
302 번	0100000~0000	0,0,0,0,0,0,0,0,0,0	1
303 번	0010000~0000	0,0,0,0,0,0,0,0,0,0	1
~	~	~	1
600 번	0000000~0001	0,0,0,0,0,0,0,0,0,0	1

1단계에서 제작한 각 모델에 각 행을 대입한 후 얻은 HR(위험비)를 가지고, i번째 행에 대하여 치료효능률을 다음과 같이 정리할 수 있습니다. (i번째 행은 i번째 유전자의 변이가 있음을 의미합니다.)

$$\text{치료효능률}_i = \frac{(i+300)\text{의 위험비}}{(i)\text{의 위험비}} \ (i = 1,\ 2,\ \ldots\ 300)$$

3단계: 유전자별 평균 치료효능률 계산을 통한 정답 도출

마지막으로 저희는 데이터의 수가 적은 한계를 극복하고, 정답률을 높이기 위해, 위의 과정을 반복하였습니다. Layer, Node 등을 변화시킨 1120개의 모델에 대하여 각 유전자의 치료효능률 값을 저장하였습니다.

여기서 저희는 모든 반복 회차의 치료효능률의 값과 Layer, Node 등의 hyper parameters, 생존 분석의 평가 요소로 쓰이는 C-index를 함께 저장했습니다.

일반적으로 생존분석에서 C-index가 0.7이상인 모델을 good model로 평가하므로, C-index가 0.7 이상인 모델[4]에 대해서만 치료효능률의 평균을 구하였습니다. 다음으로, 치료효능률의 평균을 오름차순으로 정리하여 300개의 유전자 중 앞에서부터 10개의 유전자를 뽑아 정답을 도출하였습니다.

나) 비선형 Cox regression Model을 사용한 모델 구성

주어진 데이터의 수가 적어 훈련 데이터만을 과도하게 훈련해 일반적인 데이터에 대한 정확도가 떨어지는 오버피팅 문제가 발생할 수 있다고 판단하였습니다. 이를 완화하기 위해

네트워크 생성 시 dropout 비율을 0.2로 설정하였고, 다음 은닉층의 노드 감소 비율을 설정했으며, 4개 이상의 은닉층을 가지지 않도록 했습니다.

은닉층 수, Node 수, 노드 감소 비율, 학습률을 다르게 하여 1120개의 서로 다른 모델을 훈련한 뒤 각 유전자별 치료효능률을 csv파일에 기록했습니다.

다) 교란 변수 문제의 해결

교란변수를 데이터에서 찾는 방법은 교란변수의 정의로부터 출발하였습니다. 교란변수는 독립변수와 통계적 연관성을 가져야하며, 동시에 종속변수와 원인적 연관성을 가져야 합니다. 그러므로 교란변수는 두 개의 큰 전제를 만족해야하고 아래의 두 개의 계수를 구하여 독립변수 사이의 통계적 연관성이 있는지 확인하였습니다.

우선, 선형 상관 분석[5]이 가장 기본적인 EDA라 판단하여 임상변수-임상변수, 임상변수-유전자, 유전자-유전자 사이의 상관계수 값을 계산하였지만, 최대 0.153, 최소 -0.115로 의미있는 값을 얻지 못하였습니다.

다음으로 유전자 데이터는 이진 데이터이므로 Jaccard 계수[6]를 사용하여 유전자 데이터 사이의 상관관계를 분석하였습니다. Jaccard 계수를 사용하면 Sparse data(0의 값을 많이 가진)의 특성을 가지는 0과 1로 이루어진 이진 데이터에 대한 상관 계수를 얻을 수 있기 때문입니다. 주어진 유전자 데이터의 경우 한 유전자에 대하여 값이 1인 데이터가 평균적으로 1000개중 30개이므로 Jaccard 계수를 적용하기에 적합하다고 판단하여 유전자-유전자 사이의 Jaccard 계수를 구하여 EDA를 진행하였습니다.

결론적으로, 선형 상관분석과 Jaccard 계수를 모두 Max 값이 0.2를 넘어가는 값이 존재하지 않아 변수 사이에서 충분한 통계적 상관관계가 있는 교란변수가 없다고 판단하였습니다.

라) 답의 도출 과정

여기서 저희가 답을 도출한 과정, 결론적으로 치료와 유의한 관계를 가지는 유전자 후보 10개를 뽑아낸 과정에 대하여 조금 더 자세하게 정리하고자 합니다.

저희는 위에서 설명한 과정을 통하여 총 1120개의 모델로 각 유전자에 대한 치료효능률과 C-index 값을 계산하였습니다. 다음 의미있는 생존분석 결과를 보여준 기계학습 모델을 선별하기 위하여 C-index 값이 0.7 이상인 모델을 선별하였고, 355개의 모델이 C-index의 조건을 만족하였습니다.

355개의 모델 가운데, 유전자별 평균 치료효능률을 구하였고, 오름차순으로 순위를 매겨 1등부터 10등까지의 유전자를, 치료 효과를 증가시키는 인과 관계과 확실한 유전자 후보, 즉 주어진 문제에 대한 답으로 정하였습니다. 오름차순으로 정리한 이유는 치료효능률이

작을수록 치료를 하였을 때 위험비가 더 크게 줄어든다는 것을 의미하기 때문입니다. 저희가 도출한 답은 아래와 같습니다:

G280, G88, G196, G111, G180, G284, G35, G156, G137, G74

마) 참고 문헌

[1] David R. Cox. Regression models and life-tables. In Breakthroughs in Statistics. Springer, 1992.

[2] Ping Wang, Yan Li, and Chandan k. Reddy. Machine Learning for Survival Analysis: A Survey. ACM Comput. Surv. 51, 6, Article 110 (February 2019), 2019

[3] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, Yuval Kluger. DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network. BMC Medical Research Methodology, 18:24, 2018

[4] Hajime Uno,Tianxi Cai, Michael J. Pencina, Ralph B.D'Agostino, L. J. Wei. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. Stat Med, 10: 30(10), 2011

[5] J. H. Oh, S. O. Jung. Multivariate analysis for Clinicans. Journal of the korean Shoulder and elbow society, 63:72, 2013

[6] Jaccard, Paul. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. New Phytologist. 11 (2): 37–50, 1912