모델을 작성하기 위해서는 우선 데이터셋을 찾는 게 선행되어야 한다고 생각해서 데이터셋을 찾아봤습니다.

원하는 데이터셋: 일별 기준으로 기록된 혈압(수축기, 이완기), 수면 시간, 심박수, (키, 체중)->(체질량지수 계산을 위함) + 개인의 정보인 나이, 성별, 흡연 여부, 유산소 운동 여부, 고혈압 여부 등이 기록된 데이터셋

* 개인과 밀접히 관련된 정보이기 때문에 막상 필요할 것 같은 데이터셋을 찾아도 접근하기가 쉽지 않습니다.
* 조사를 진행하다가 EHR(Electronic Health Record, 전자 건강 기록)이라는 것을 알게 되었고, 그럼 무료로 접근할 수 있는 EHR 데이터셋이 있나 찾아보다가 아래의 데이터셋들을 발견하게 되었습니다.
* 일단 원하는 데이터셋에 가장 가까웠던 3개를 서술하도록 하겠습니다.

1. MIMIC

MIMIC은 Medical Information Mart for Intensive Care의 약자로 Beth Israel Deaconess Medical Center 중환자실에서 발생한 데이터를 MIT에서 정리하고 개발한 데이터셋입니다. 혈압 측정에 관한 데이터가 존재할 것 같아 이 데이터셋의 최신 버전인 MIMIC-IV의 샘플 데이터를 받아 조사해보았습니다. 조사 결과, 환자에 대한 혈압 측정 주기가 들쭉날쭉(이틀부터 길게는 4달 주기도 있습니다.)해서 사용할 수 없다고 판단했습니다. 또한 수면 시간, 심박수 등의 데이터도 존재하지 않습니다.

1. KoGES

KoGES는 한국인유전체역학조사사업을 말합니다. 반복 추적 조사를 통해 데이터를 수집했다고 해서 좀 더 조사해 보았지만, 2~4년 주기로 조사를 진행했다고 해서 사용할 수 없다고 판단했습니다.

1. AIhub ‘만성질환 관련 임상 및 생활습관 데이터’

웨어러블 데이터로 수집한 데이터가 있다고 해서 문의를 넣고 결과를 기다렸습니다. 고혈압 수치는 45일 당 1번 검사하였으며, 총 90일간 조사를 수행하였으니 고혈압 수치는 총 2개 밖에 없습니다. 또한 AIhub는 해당 데이터에 대해 ‘안심존’이라는 서비스 내에서 작업을 수행할 수 있도록 하는데, 여기서 학습시킨 모델을 반출시키는 것에 대한 내용이 불투명해서 사용할 수 없다고 판단했습니다.

그 외에도 이런저런 데이터셋들을 많이 찾아보았으나 아예 시계열 데이터가 아니거나, 필요한 요소들이 다수 빠져 있다 같은 문제로 사용할 수 없었습니다.

그래서 다른 방법을 생각해봤습니다.

1. 시계열 데이터를 임의로 구성하기

처음에 제안서에 요구된 AI 모델이 시계열 데이터니까 억지로라도 시계열 데이터를 구성해서 사용한다면 어떻게 해야 하는가에 대해 고민하다가 떠올린 방법입니다. 일단 계속 읽어주시길 바랍니다.

저번에 SVM 모델을 작성하고 학습시키면서 국민건강영양조사라는 공개 데이터셋을 사용하였습니다. 이 데이터를 활용할 것입니다.

우선 하나의 가정이 필요합니다. 고혈압이 없는 사람들은 공통적으로 고혈압과 거리가 먼 생활 패턴을 갖고 있을 것이고, 고혈압이 있는 사람들은 공통적으로 고혈압과 거리가 가까운 생활 패턴을 갖고 있을 것이라고 가정합니다.

그러면 우리가 얻을 수 있는 데이터셋에서 우선 고혈압 여부로 데이터를 나눈 다음(없는 사람 – 0, 있는 사람 – 1), 각각의 경우에 대해 일정 부분을 샘플링하고 셔플해줍니다. 그리고 그 데이터를 시계열 데이터라고 가정하고 AI 모델을 학습시킵니다.

텍스트, 도표, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

같은 영역에 속하는 사람들은 생활 패턴이 비슷하기 때문에 임의로 시계열 데이터로 구성해도 괜찮을 것이라 생각했기 때문입니다.

Q. 신빙성이 있는가?

A. 신빙성이 없다 쪽에 가깝습니다. 하지만 현재 상황으로 미루어 볼 때, 시도해볼 만한 가치는 있다고 생각합니다. 데이터셋을 찾는 걸 포기한 게 아닙니다. 데이터셋은 계속 찾아볼 것이고, 당장 모델을 구성하고 학습을 시키고 백엔드와 연동 같은 작업을 해야 한다고 했을 때, 임시로 고안한 아이디어일 뿐입니다.

Q. 성별, 나이, 흡연 여부, 유산소 운동 여부는 학습시킬 때, 값이 변하지 않는다. 이에 대한 대책은 있는가?

A. 고혈압 여부로 자른 데이터셋에서 더 세부적으로 자른 다음 데이터셋을 구성하면 된다고 생각합니다. 우선 성별, 그 다음으론 나이대(~10, 20, 30, 40, 50, 60~), 그 다음으론 흡연 여부, 그 다음으론 유산소 운동 여부로 자르면 될 것입니다. 이렇게 더 세부적으로 자른 데이터셋에서 표본을 추출해 시계열 데이터를 구성하면 될 것 같습니다.

만약 그렇게 나눈 구간 안의 데이터 양이 너무 적다면 나이대의 범위를 더 포괄적으로 잡거나, 흡연 여부나 유산소 운동 여부를 제거한 후에 고혈압 관리 요망이 떴을 때 흡연 여부를 확인해 담배를 끊으라 조언하거나, 유산소 운동 여부를 확인해 운동을 하자 같은 조언을 해줄 수 있습니다.

Q. 그렇게 분류한다면 성별, 나이, 흡연 여부, 유산소 운동 여부는 결과에 영향을 못 미치게 되는 것 아닌가?

A. 우선 요소들을 고정 인자와 변수 인자로 구분해야 합니다. 고정 인자는 학습을 시키면서 값이 변하지 않는 데이터 요소들을 말하며 이는 성별, 나이, 흡연 여부, 유산소 운동 여부를 말합니다. (고혈압 여부는 결과와 관련된 값이므로 제외했습니다.) 그리고 변수 인자는 학습을 시키면서 값이 변하는 데이터 요소들을 말하며 이는 15초간 맥박수, 수축기 혈압, 이완기 혈압, 체질량 지수, 수면 시간을 포함합니다.

고정 인자들은 1D-CNN, 변수 인자들은 LSTM으로 학습시킵니다. 그리고 나온 결과값을 concatenate해서 FC layer로 변환해 결과값을 낼 수 있겠다고 생각했습니다.

Q. 그럼 위험도는 어떻게 정하는가?

A. 3월 24일에 회의를 하면서 나왔던 내용입니다. 단순히 고혈압 여부(1/0)를 가지고 위험도를 어떻게 나타낼 것인가에 대한 논의였습니다. 그때는 출력 함수로 sigmoid 함수를 사용해 학습을 시킨 후, sigmoid 값을 확인해보자고 결론을 내렸습니다.

위의 방법이 통하지 않았을 때를 대비해 새로운 방법을 생각해봤습니다. 고혈압 여부가 0인 데이터와 1인 데이터를 적절히 혼합하는 것입니다.

그래서 0인 데이터만 있는 건 결과가 0, (안전)

0인 데이터와 1인 데이터 비율이 8:2면 결과가 1, (의심)

0인 데이터와 1인 데이터 비율이 6:4면 결과가 2, (경고)

0인 데이터와 1인 데이터 비율이 4:6이면 결과가 3, (위험)

0인 데이터와 1인 데이터 비율이 2:8이면 결과가 4, (강한 관리 요망)

1인 데이터만 있으면 결과가 5로 mapping 되도록 만들면 될 듯합니다. (병원 추천)

이렇게 하려면 원핫인코딩이 필요하므로, 소프트맥스 함수를 사용해야 합니다. 위험도 %에 대한 내용은 사라지겠지만 위험도에 대한 예측이 가능해집니다.

Q. 그래서 사용할 모델은 무엇인가?

A. 일단은 총정리\_3 파일에 작성한 것처럼 시계열 데이터를 처리할 때는 기본적으로 LSTM을 사용하려고 합니다. 성별, 나이 등의 고정 인자들은 어떻게 다루어야 할 지에 대한 고찰이 좀 더 필요할 듯합니다.

Q. 이대로 진행한다면 무엇부터 해야 하는가?

A. 국민건강영양조사 파일들을 다운로드 받아서 기초적인 데이터를 구축합니다. 그리고 고혈압 여부, 성별, 나이 등에 따른 기준으로 분류해야 합니다. 그렇게 분류한 데이터를 바탕으로 앞서 말했던 일정 비율에 맞춘 데이터셋을 구축하는 것이 일차적인 목표입니다. 또한 학습시킬 때에도 각각을 특정 비율로 나누고 순차적으로 합쳐서 학습 데이터를 만들어야 하겠습니다.

학습 데이터 중에 결측치인 데이터에 대해 어떻게 처리할 것인지에 대해서도 생각해야 합니다. 결측치가 존재한다면 그 값을 채우지 않고 해당 row를 아예 제거한다 같은 결론을 낼 수도 있습니다. (데이터가 충분히 많다면) 아니면 논문 조사할 때 찾았던 방법인 협력적 필터링 기법을 사용할 수도 있습니다.

그리고 앞서 말했던 것들을 바탕으로 모델(코드)을 작성해야 합니다. 그에 대한 조사 및 공부가 필요합니다.

초매개변수(hyperparameter)를 어떻게 정의할 것인가도 성능과 직결되므로 중요한 문제입니다. 우선은 임의로 초매개변수를 정한 후에 성능 개선을 수행할 때, 수치를 바꿔보면서 수행하면 될 것 같습니다. LSTM의 초매개변수로는 학습률(learning rate), 윈도우 길이(window size), 최적화(optimization) 방법, 히든 유닛(hidden unit) 개수, 드롭아웃(dropout) 비율, 에폭(epoch) 등이 있습니다.

1. 시계열 데이터를 사용하지 않기

다음은 3월 24일자 회의에서 나왔던 내용입니다. 데이터를 수집하는 과정에서 사용자 개인의 문제로 측정을 일주일 마다 한 번씩 하다가 고혈압 예측을 요청했다면 어떻게 해야 할까요? 그 사이의 값들에 대한 전처리를 해주어야 하는데 해당 과정에서 시간이 소요되고 또한 전처리한 값은 정확한 수치가 아니므로 정확도가 떨어질 수 있습니다.

2년 동안 꾸준히 앱을 사용해 필요한 정보들을 입력한 사용자가 있고, 이 사람이 고혈압 예측을 요청했다고 가정합시다. 그럼 DB에 저장된 대규모의 시계열 데이터를 전송하고 학습시키는 과정에서 많은 시간이 소요될 것입니다.

(4/3: 실제로 **이진 분류문제에서의 딥러닝 알고리즘의 활용 가능성 평가** 논문에서 2년 6개월 간의 데이터에 대해 학습을 시켜본 결과, 한 훈련 횟수당 MLP 모형은 약 22초가 소요된 반면에 LSTM 모형은 약 24분가량이 소요되었다는 내용이 있습니다.)

대부분의 논문에서는 년 단위로 데이터를 학습시키고 질병을 예측합니다. 이는 1년 이하의 짧은 기록으로 질병을 예측하는 것은 의미 없는 일이라고 말하는 것과 같다고 생각합니다. 그리고 1년 넘게 꾸준히 앱을 사용하면서 기록을 남기는 사람들보다는 그 미만의 기간으로 앱을 사용할 사용자들이 많을 가능성이 높습니다. (Slack – 전민재 님)

위와 같은 문제로 시계열 데이터를 사용하지 않고, 특정 시점 또는 평균 값을 사용한다면 데이터가 어떻게 구성되어야 하는가에 대해 생각해봤습니다.

국민건강영양조사 원시자료 이용지침서 내용 중에 수면 시간은 ‘주중 잠자리에 든 시각과 일어난 시각을 이용하여 분으로 계산된 하루 평균 수면시간’이라고 정의되어 있습니다. 사람의 생활 패턴을 대표하는 값으로 최빈값, 평균값, 중앙값 등이 있을 수 있지만, 앞서 정의된 내용에 따라 혈압(수축기, 이완기), 수면시간 등의 평균값을 계산해 AI 모델로 전송하면 될 듯합니다.

나머지 요소들에 관한 내용도 찾아봤지만 마땅히 평균과 관련된 내용을 찾지 못했습니다. 근거는 빈약하지만 우선은 평균값을 메인으로 정하겠습니다.

1. 유방암 분류를 위한 전이학습 기반 DNN-SVM 하이브리드 모형 = Transfer Learning based DNN-SVM Hybrid Model for Breast Cancer Classification

<https://www.riss.kr/link?id=A108848834>

3p: 본 논문에서는 다루고자 하는 유방암 데이터는 영상 형태가 아닌 Tabular 데이터이다. 여기서 Tabular 데이터란 테이블(table) 형태의 행(row)과 열(column)로 표현되는 데이터를 말하며, 주로 정형 데이터(structured data)라고 한다. 하나의 행은 하나의 데이터 인스턴스를 나타내며, 각 열은 피처(feature)를 나타낸다. CNN에서 다루는 영상은 대표적인 비정형 데이터(unstructured data)이며, 이는 테이블 형식으로 표현할 수 없는 데이터를 말한다. CNN 모형은 영상의 공간정보를 유지한 채 학습이 이루어진다. 하지만 공간적인 관계가 없는 Tabular 데이터에 CNN 모형을 그대로 적용할 수 없다. 따라서 본 논문에서 유방암과 같은 Tabular 데이터 처리는 CNN 모형보다 DNN 모형을 사용하는 것이 더 적절하다.

* 우리 데이터도 Tabular 데이터에 속하니 논문에 따라서 DNN 모델을 사용하는 것이 적절합니다.

1. Heart disease risk prediction using deep learning techniques with feature augmentation

<https://link.springer.com/article/10.1007/s11042-023-14817-z#Sec15>

심혈관계 질환과 관련된 feature들을 선정하고 augmentation해서 Decision tree, Random forest, K-nn, AdaBoost, XGBoost, Gaussian Naïve Bayes, Multilayer Perceptron을 비교하여 MLP의 성능이 미세하게 뛰어나다는 것을 증명하였습니다.

* 고혈압과 관련된 feature들을 선정하고 학습시키는 것에 대한 근거가 되는 논문입니다.

Q. 위험도 분류는 어떻게 하는가?

A. 우선 학습을 시키고 다른 데이터를 input으로 넣었을 때, 나온 시그모이드 함수 값을 확인해 보아야 합니다. 0부터 1 사이의 값으로 mapping 될 것은 확실하지만, 그 수치가 위험도를 대변할 수 있는가에 대한 여부는 확인해봐야 알 수 있습니다.

다른 방법은 일단 고혈압 여부가 1인 사람들을 혈압 순으로 정렬합니다. 그리고 일정 비율로 1~4까지의 번호를 부여합니다. (번호가 높을수록 혈압이 높은 사람, 0은 고혈압 없는 사람) 그리고 0부터 4까지에 대해 같은 비율로 데이터셋을 구축하고, 학습을 시킵니다.

이 방식의 경우 앞과 마찬가지로 소프트맥스를 사용합니다.

Q. 모델은 어떻게 구성하는가?

A. DNN을 사용한다면 우선 데이터 전처리를 해서 0과 1사이의 값으로 정규화해주어야 합니다. (결측치에 대한 것도 처리해야 합니다.) 위에서 말했던 변수들(고혈압 여부 제외)이 DNN 모델에 모두 input으로 들어갈 수 있으며, 출력 함수로는 우선 sigmoid 함수를 사용해보고, 안되면 소프트맥스 함수를 사용합니다.

단순한 mnist 학습 모델을 기반으로 시작해서 점점 초매개변수들과 모델 구성을 바꿔 나가며 발전시킬 예정입니다.

3/28

평균값 등으로 input으로 사용했을 때의 결과값이 위험도로 사용될 때의 정확도에 대해서 생각해보고 있었습니다

논점 1: 시그모이드 함수가 확률 값을 대변할 수 있는가?

1. **고혈압 발생 예측 모형 개발**

<http://www.khep.or.kr/sub/subMain.jsp?menuseq=4010> - 보건교육건강증진학회지 제23권 제4호

2p: 생활습관을 중심으로 고혈압의 발생확률을 예측하여야 할 것이다. (연령, 성별, 신장, 체중, 비만도, 혈압, 음주여부, 흡연여부, 운동여부)

3p: 국민건강보험공단의 건강검진자료

5p: 고혈압 발생을 예측할 수 있는 모형의 개발은 로지스틱회귀분석, 의사결정나무분석, 앙상블 기법을 활용하였고 이를 평가 비교하여 이 중 가장 좋은 모형을 채택하였다.

8p: 로지스틱 회귀모형이 가장 우수한 모형으로 판명됨에 따라 로지스틱 회귀모형을 고혈압 발생을 예측하는 모형으로 채택하였다.

* 이진 분류를 위한 로지스틱 회귀모형에 대해 찾아보자.

1. **로지스틱 회귀모형을 이용한 산불발생확률모형 개발**

<https://www-riss-kr.libproxy.knu.ac.kr/link?id=A60057775>

3p: 로지스틱 회귀모형을 이용하여 산불 발생확률을 도출하였다. ~ 모형의 정확도 평가를 위하여 발생확률 0.5를 기준으로 분할표를 작성하여 분류 정확도를 살펴보았다.

1. **로지스틱회귀모형을 이용한 위험물수송차량 교통사고확률 분석 연구**

<https://www-riss-kr.libproxy.knu.ac.kr/link?id=A108174689>

4p: 로지스틱 회귀모형은 세 가지의 성질을 가진다. 첫째, 각 대안을 선택할 확률은 0과 1 사이의 값을 가진다.

로지스틱 회귀모형은 선형 방정식을 구한 후에 시그모이드 함수를 사용해 확률 값을 예측합니다. 그리고 0~1 사이로 나오는 output은 확률 값이라고 해석할 수 있는 여지를 위의 논문들에서 제공합니다. 따라서 출력 함수로 sigmoid를 사용해도 될 듯합니다.

1. **이진 분류문제에서의 딥러닝 알고리즘의 활용 가능성 평가**

<https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07135528>

5p: 본 연구에서는 0과 1 사이의 값만 입력 받을 수 있는 인공신경망 모형들의 특성상 ~연속형 값들은 0과 1사이로 변환하는 전처리 과정을 수행하였다.

이후 작성한 모형의 성능 평가를 위해 랜덤하게 30%의 데이터를 분리하여 테스트 데이터로 사용하고 나머지 70%의 데이터를 훈련 데이터로 사용하였다. 마지막으로 불균형 학습을 막기 위해 전처리 과정에서 훈련용 데이터에 대해 과대 표본화(Oversampling)를 하여 두 응답의 데이터 건수를 균등하게 맞추었다.

은닉층의 수, 은닉층 내 뉴런의 수 또는 출력되는 데이터(필터)의 개수, Dropout 기법의 적용 유무 조건을 기준으로 모형들 간의 성능 비교를 실시하였다.

6p: CNN 알고리즘의 필터 크기를 한 레코드의 전체 필드 길이만큼 설정하여 데이터의 전체 특징을 한 번에 학습하도록 하였다. 이후 추출한 특징들을 바탕으로 모형이 의사결정을 할 수 있도록 은닉층을 하나 추가하여 모형을 완성하였다.

6~7p: <Figure 1>, <Figure 2> - 이진 분류를 위한 output layer로 sigmoid 함수 사용

9p: Dropout 기법을 적용한 CNN 모형이 가장 좋은 F1 Score를 보였다

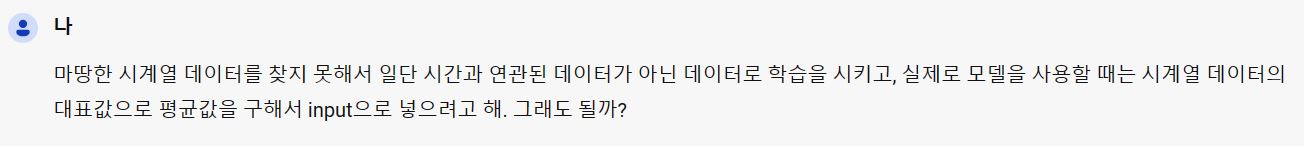
논점 2: 사용자가 하루마다 기록한 데이터들의 평균값을 input으로 사용했을 때, 그 output을 위험도로 사용해도 되는가? (정확한가?)

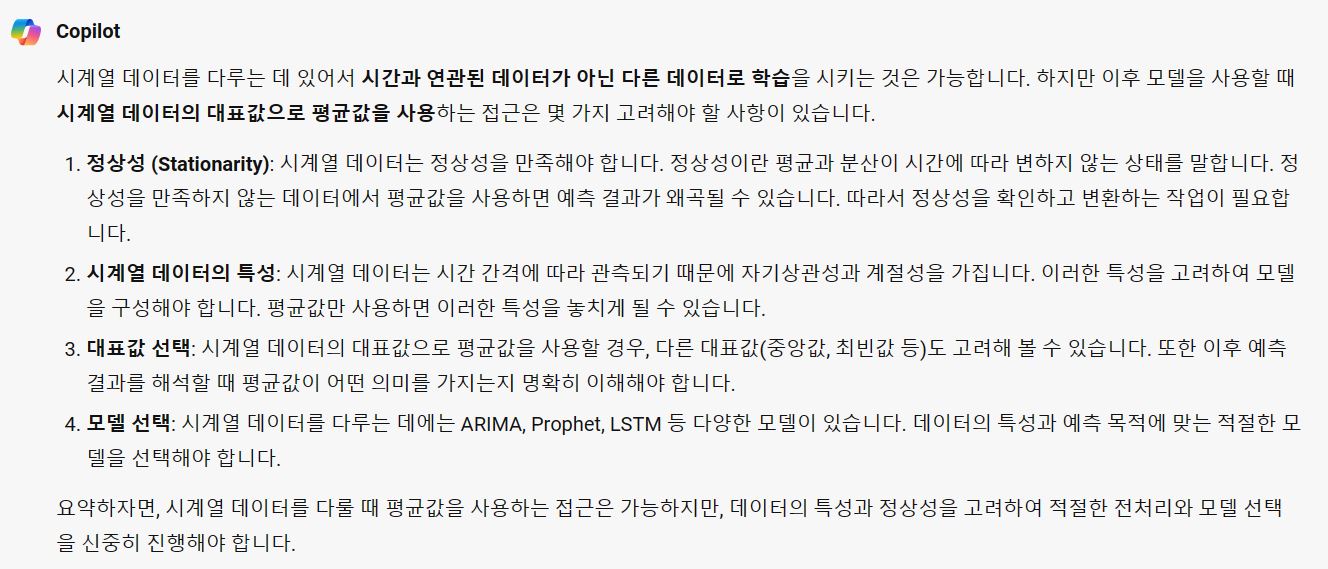
1. **한국한의학연구원 논문의 통계적 오류에 관한 연구**

<https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=JAKO200800557084641>

4p: 중심경향과 산포도는 모수적 방법을 사용한 경우에는 일반적으로 평균 표준편차를 제시하며, 비모수적 방법을 사용한 경우에는 중앙값과 범위를 주로 사용한다.

* 모수적 방법은 수집한 데이터가 분포를 따른다는 가정하에 접근하는 방법을 말하며, 비모수적 방법은 분포를 가정하지 않고 접근하는 방법을 말한다. 그리고 중심경향을 나타내는 평균값만을 사용하는 것은 데이터의 분포에 따라 오류가 있을 수 있다.
* 따라서 평균값만 사용하지 않고 일일 습관 데이터가 일정한 분포를 따른다고 가정한 후에, 각각의 변수(혈압, 수면시간 등)의 평균값과 표준편차를 input 값으로 사용한 후, 1d-cnn으로 kernel\_size는 2, strides도 2로 설정하여 평균과 표준편차의 특징을 추출하는 방향으로 학습시켜보자.
* 학습 데이터의 표준편차는? 0과 1(고혈압 여부)로 나눈 데이터셋에서 각각 평균값을 구하고, 그에 속하는 하나의 데이터에 대해 표준편차를 구해보자. (보류)





일단 코드를 작성해보고 평가 기준을 만족하지 못한다면, 평균값 또는 평균값과 표준편차가 적절하지 않다고 판단한 후, 새로운 모델을 구성하는 방향으로 진행해야 할 것 같습니다.

그 외 참고할 수 있는 사이트

국민건강보험 건강검진 코호트 데이터를 이용한 질병 예측 딥러닝 기반 모델 개발

<https://github.com/Ahn-HS/Survival-Model-for-Neural-Networks>

발병 확률 예측 모델인 것으로 추정됩니다. 마찬가지로 출력함수에 시그모이드 함수를 사용했습니다.

Selvy Checkup

<https://www.selvycheckup.com/overview>

건강검진 정보를 기반으로 근 미래의 주요 질환에 대한 발병 위험도를 예측해주는 인공지능 스마트 헬스케어 제품입니다.

4/1

고혈압의 위험도는 모델의 출력 함수를 시그모이드 함수로 정해서 대변하는 것이 최선일 듯합니다.

그리고 시계열 데이터를 사용하는 것보다는 일반적인 데이터를 사용하여 고혈압 위험도를 예측하는 방향으로 진행하려고 합니다.

<sigmoid를 이진 분류로 사용한 테스트 예제>

<https://colab.research.google.com/drive/1cdHjVWKM35V6ujmzIiLRso9haIjSrCFw#scrollTo=DdWzFRiNrjSq>

아래 블로그의 글을 참고해 예측값을 도출하는 테스트를 진행했습니다.

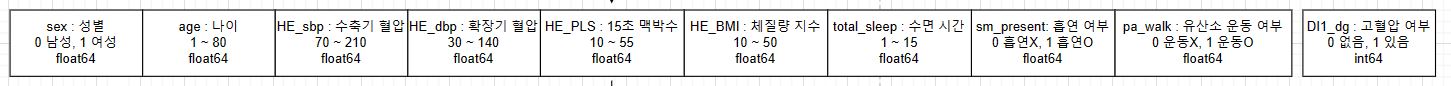
<https://tensorflow.blog/%ec%bc%80%eb%9d%bc%ec%8a%a4-%eb%94%a5%eb%9f%ac%eb%8b%9d/3-4-%ec%98%81%ed%99%94-%eb%a6%ac%eb%b7%b0-%eb%b6%84%eb%a5%98-%ec%9d%b4%ec%a7%84-%eb%b6%84%eb%a5%98-%ec%98%88%ec%a0%9c/>

학습시킨 모델에 predict 함수를 사용하면 output으로 예측값(확률)이 나오는 것을 확인했습니다. 모델을 저장하고 불러오는 것도 Keras를 통해 구현해 봤습니다.

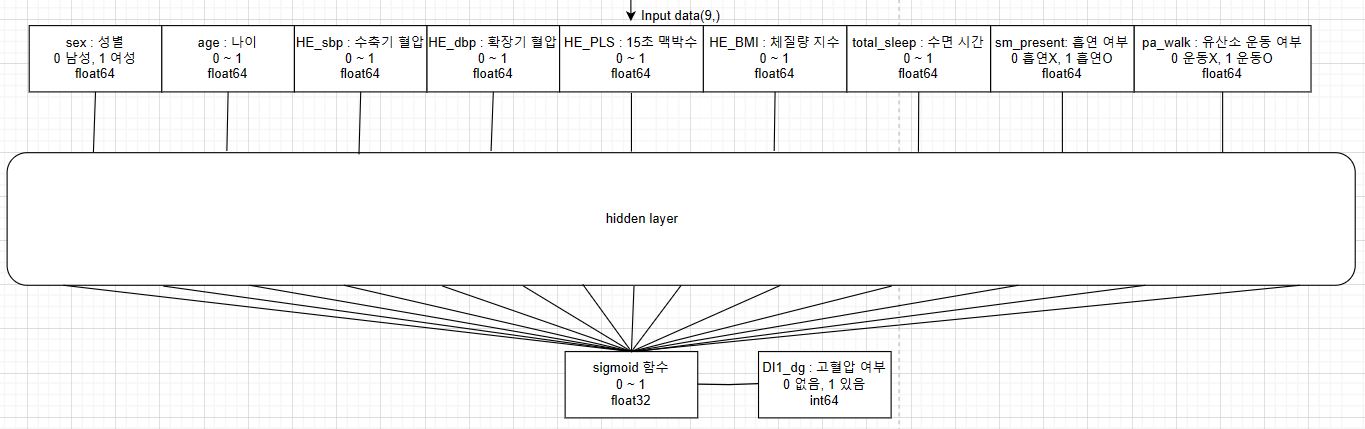
4/2

전체 과정 정리

1. 국민건강영양조사 원시 자료를 사용할 수 있는 데이터로 가공합니다.
2. 결측치를 처리합니다. (결측치가 존재하면 해당 행을 제거합니다. 7992개의 데이터에서 2063개의 데이터가 제거되어 5929개의 데이터가 남았습니다.)
3. 아래는 위의 과정을 수행하고 얻은 변수들과 대략적인 값의 범위입니다.



1. 모델에 입력할 데이터들을 0 ~ 1 사이의 값으로 정규화 시켜줍니다.



1. 입력층에 입력될 데이터 9개와 출력층에 고혈압 위험도를 대변할 시그모이드 함수를 사용합니다(고정). 그리고 은닉층의 구조와 하이퍼파라미터 변경, 추가 데이터셋 수집 등의 작업이 필요합니다.

우선 딥러닝의 기본이라 할 수 있는 mnist 숫자 인식 모델의 구조를 참고해 시뮬레이션 해보았습니다. 수행한 결과, 정확률이 0.63으로 현재는 SVM보다 못한 성능을 보여줍니다. 모델 개선이 필요하며 또한, loss 값이 왜 음의 방향으로 커지는 지에 대해서도 추가적인 조사가 필요해 보입니다.

추후 계획으로는 모델 구조 개선이나 하이퍼 파라미터 변경, 추가 데이터셋 수집, 시계열 데이터 조사 등이 있습니다.

 사진이 잘 보이지 않아 추가적으로 첨부합니다.