Week3-1

1. **학습정리**
    1. 베이즈 통계학
        1. 조건부 확률
            1. B기븐A

조건부 확률이란?

- 베이즈 통계학을 이해하기 위해선 조건부확률의 개념을 이해해야 합니다
- 베이즈 정리는 조건부확률을 이용하여 정보를 갱신하는 방법을 알려줍니다

$$P(A \cap B) = P(B)P(A|B)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = P(B)\frac{P(A|B)}{P(A)}$$

A 라는 새로운 정보가 주어졌을 때 $P(B)$ 로부터 $P(B|A)$ 를 계산하는 방법을 제공한다

2. 베이즈 정리
    1. A라는 정보가 새로 주어졌을 때 B가 일어날 확률을 계산할 수 있음
    2. 예제 문제

베이즈 정리: 예제

$$P(\theta|\mathcal{D}) = P(\theta)\frac{P(\mathcal{D}|\theta)}{P(\mathcal{D})}$$

가능도 (likelihood)

사후확률 (posterior)

사전확률 (prior)

Evidence

- COVID-99 의 발병률이 10% 로 알려져있다. COVID-99 에 실제로 걸렸을 때 검진될 확률은 99%, 실제로 걸리지 않았을 때 오검진될 확률이 1% 라고 할 때,

베이즈 정리: 예제

$$P(\theta|\mathcal{D}) = P(\theta)\frac{P(\mathcal{D}|\theta)}{P(\mathcal{D})}$$

$$P(\theta) = 0.1$$

$$P(\mathcal{D}|\theta) = 0.99$$
$$P(\mathcal{D}|\neg\theta) = 0.01$$

사전확률

가능도

- COVID-99 의 발병률이 10% 로 알려져있다. COVID-99 에 실제로 걸렸을 때 검진될 확률은 99%, 실제로 걸리지 않았을 때 오검진될 확률이 1% 라고 할 때, 어떤 사람이 질병에 걸렸다고 검진결과가 나왔을 때 정말로 COVID-99 에 감염되었을 확률은?

θ 를 COVID-99 발병 사건으로 정의(관찰 불가)하고,
$\mathcal{D}$ 를 테스트 결과라고 정의(관찰 가능)한다

  - 에비던스 구하는 법

$$P(\theta|\mathcal{D}) = P(\theta)\frac{P(\mathcal{D}|\theta)}{P(\mathcal{D})}$$

$$P(\theta) = 0.1$$

$$P(\mathcal{D}|\theta) = 0.99$$
$$P(\mathcal{D}|\neg\theta) = 0.01$$

$P(\neg\theta)$

$$P(\mathcal{D}) = \sum_{\theta} P(\mathcal{D}|\theta)P(\theta) = 0.99 \times 0.1 + 0.01 \times 0.9 = 0.108$$

만일 $P(\mathcal{D}\,|\,\neg\theta)$ 를 모른다면 이 문제는 풀기 어렵다

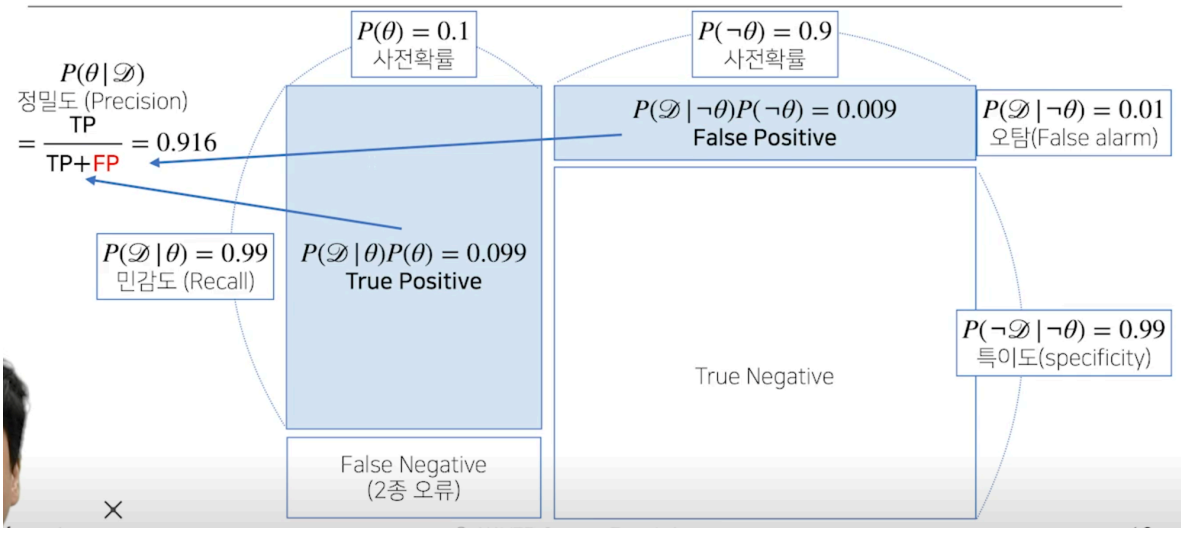
  - 답

$$P(\mathcal{D}) = \sum_{\theta} P(\mathcal{D}|\theta)P(\theta) = 0.99 \times 0.1 + 0.01 \times 0.9 = 0.108$$

$$P(\theta|\mathcal{D}) = 0.1 \times \frac{0.99}{0.108} \approx 0.916$$

3. 조건부 확률 시각화

조건부 확률의 시각화



4. 베이즈 정리를 통한 정보 갱신

베이즈 정리를 통한 정보의 갱신

- 베이즈 정리를 통해 새로운 데이터가 들어왔을 때 앞서 계산한 사후확률을 사전확률로 사용하여 갱신된 사후확률을 계산할 수 있습니다

$$P(\theta|\mathcal{D}) = P(\theta)\frac{P(\mathcal{D}|\theta)}{P(\mathcal{D})}$$

사후확률
(posterior)

$$P(\theta|\mathcal{D}) = P(\theta)\frac{P(\mathcal{D}|\theta)}{P(\mathcal{D})}$$

갱신된 사후확률
(posterior)

$$P(\theta|\mathcal{D}) = 0.1 \times \frac{0.99}{0.189} \approx 0.524$$

- 앞서 COVID-99 판정을 받은 사람이 두 번째 검진을 받았을 때도 양성이 나왔을 때 진짜 COVID-99 에 걸렸을 확률은?

$$P(\theta|\mathcal{D}) = 0.1 \times \frac{0.99}{0.189} \approx 0.524$$

$$P(\mathcal{D}|\theta) = 0.99$$
$$P(\mathcal{D}|\neg\theta) = 0.1$$

$$P(\mathcal{D}^*) = 0.99 \times 0.524 + 0.1 \times 0.476 \approx 0.566$$

갱신된 사후확률
(posterior)

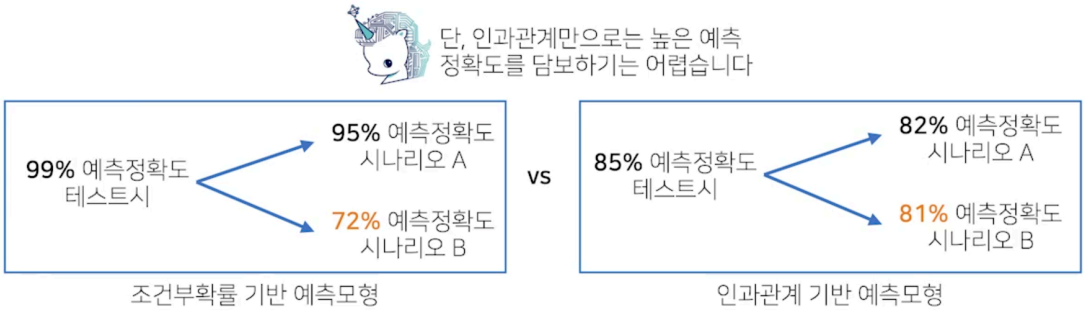
$$P(\theta|\mathcal{D}^*) = 0.524 \times \frac{0.99}{0.566} \approx 0.917$$

5. 조건부 확률만으로 인과관계를 추론할 수 없음
    1. 인과관계 없이 예측모형을 만드면 테스트시에는 높은 정확도를 보여줄 수 있지만 실제 시나리오에서는 데이터 분포가 달라져 다양한 정확도가 나올 수 있음
    2. 인과관계만으로 예측모형을 만들면 정확도는 별로 높지 않지만 데이터

분포 변화에 강건함

조건부 확률 → 인과관계?

- 조건부 확률은 유용한 통계적 해석을 제공하지만 인과관계(causality)를 추론할 때 함부로 사용해서는 안 됩니다
- 인과관계는 데이터 분포의 변화에 강건한 예측모형을 만들 때 필요합니다



3. 중첩요인 제거

- 인과관계를 알아내기 위해서는 중첩요인(confounding factor)의 효과를 제거하고 원인에 해당하는 변수만의 인과관계를 계산해야 합니다

2. 딥러닝 기본 용어
    1. 딥 러닝 4 가지 키 컴포넌트
        1. 데이터
        2. 모델
        3. 손실함수
        4. 알고리즘
    2. 딥러닝 발전과정
3. 파이토치 시작하기
    1. 파이토치 장점
        1. numpy(행렬연산)
        2. autograd(자동미분)
        3. 딥러닝관련 다양한 함수, 모델
    2. vscode와 colab 연동
    3. tensor 타입
        1. torch.FloatTensor(데이터)
        2. 데이터에는 리스트나 numpy 데이터 타입
        3. .size() 하면 쉐잎이 나옴
    4. numpy 에 있는 함수가 거의다 대응 됨
    5. squeeze()

```
[27] t1.view(-1, 10)

     tensor([[0., 1., 2., 3., 4., 5., 6., 7., 8., 9.]])
```

```
[26] t1.view(-1, 10).squeeze()

     tensor([0., 1., 2., 3., 4., 5., 6., 7., 8., 9.])
```

```
t1.view(-1, 10).squeeze().unsqueeze(dim=1)

tensor([[0.],
        [1.],
        [2.],
        [3.],
        [4.],
        [5.],
        [6.],
        [7.],
        [8.],
        [9.]])
```

4. 뉴럴 네트워크 - MLP
    1. 히든 레이어가 한 개만 있는 뉴럴 네트워크의 표현력은
    2. MLP
        1. 가장 간단한 뉴럴 네트워크
    3. MLP 로스 펑션
        1. 리그레션
            1. 에러 최소화
            2. 입력값의 타겟과 모델의 출력값의 차이의 제곱을 출임
        2. 분류
            1. 크로스 엔트로피 최대화
            2. 모델의 아웃풋은 원-핫 벡터 (하나만 1 나머진 0)
            3. 모델은 뉴럴 네트워크의 출력값에서 해당하는 클래스의 값만 높임
        3. 확률적 모델

Multi-Layer Perceptron

- What about the loss functions?

True target

Predicted output

Regression Task $$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D}(\boxed{y_i^{(d)}} - \boxed{\hat{y}_i^{(d)}})^2$$

Classification Task $$\mathrm{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D} y_i^{(d)} \log \hat{y}_i^{(d)}$$

Probabilistic Task $$\mathrm{MLE} = \frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D} \log \mathcal{N}(y_i^{(d)}; \hat{y}_i^{(d)}, 1)$$ (=MSE)

5. 데이터셋 다루기
    1. 데이터 셋 만들기
        1. 기본적으로 세 기능
            1. init
            2. len
            3. getitem
    2. 데이터 로더 만들기
        1. 이터레이터
        2. iter=한번에 가져올 수
    3. 이미지 처리
        1. from PIL import image
        2. 이미지.getdata()
            1. 이미지를 벡터로 변환
    4. 나만의 데이터 셋 만들어보기
        1. 클래스 만들기
        2. 데이터 셋을 다운로드 받기
        3. 어딘가에 저장
        4. 거기에서 텐서를 뽑기
        5. MLP에 넣고 돌려보기

2. **피어세션**
    1. 강의리뷰
        1. 강의 소감
            1. 조건부 확률에 대해
            2. 인과관계
        2. 강건한 모델에 대해
            1. 변화에 강건하지 않아야 하는 모델은?
        3. 조교 : 인터뷰
            1. 인터뷰 문서
                1. https://docs.google.com/document/d/1F3ZWNVTLRPncF30iW10cup5R40ypXBQHlFi61CJk6Q0/edit
            2. 인터뷰 순서

            1. 문제읽고 문제 정의
                1. 인터뷰어와 인터뷰이가 서로 같은 선상으로 이해하고 있는지
                2. 이해하고 있는 것에 대해 인터뷰어에게 질문
            2. 테스트케이스 만들고
                1. 결과 물어보기
            3. 자기 아이디어 말하기
            4. 코드 작성
        3. 인터뷰 방법 정리
            1. 문제에 대해 clarify하는 것부터 시작
                1. 같은 선상의 이해도를 가지고 있는지 유도질문
                2. ex)제가 이해하는 것이 ~~한데 이렇게 생각하면 될까요?
            2. 테스트케이스를 만듦
                1. 엣지케이스 나열 + 동의 확인
            3. 자기 아이디어
                1. ex) 아이디어 구상에 대해서도 확인
            4. 코드를 작성
                1. ex) 인터뷰어가 코드에 대해 추가질문
            5. 팁 : 질문 대답하면서 계속 코딩할 것
                1. 여러 접근 방법에 대해 맞다/틀리다 보다는 괜찮을까요? 라고 물어보기
                2. 간단한 psuedocode 적으면서 물어보기
                3. 오디오가 끊기지 않도록 코드를 쓰면서도 말하기
        4. 엘리베이터 피치
            1. 인생을 쭉 읊는 것은 금지
            2. 워딩은 깔끔하게
            3. 프로젝트 경험은 자신의 프로젝트 경험이 회사에 어떻게 도움이 될 것인지 말하기
    4. 파이썬 딕셔너리는 들어간 순서가 보장됨
3. **과제**
    1. 완료