(보안데이터분석) 연습문제_07

1. 어느 카페의 고객 100명에 대한 조사 결과가 다음과 같다.
   - 커피를 주문한 고객: 70명
   - 차를 주문한 고객: 40명
   - 커피와 차를 모두 주문한 고객: 25명
     임의로 선택한 한 명의 고객이 커피를 주문했을 때, 이 고객이 차도 주문했을 확률은 얼마인가?
   1. $\frac{25}{100}$ 👍 2) $\frac{25}{70}$ 3) $\frac{25}{40}$ 4) $\frac{40}{70}$ 5) $\frac{70}{100}$

> 커피를 주문한 사건을 A, 차를 주문한 사건을 B 라고 하면 커피를 주문한 고객이 차도 주문을 했을 경우는 조건부 확률 $P(B|A)$이 된다. 다시 말해서 $P(B|A) = \frac{P(A \cap B)}{P(A)}$ 이된다. 이 때 커피를 주문한 고객일 확률 70/100, 커피와 차를 모두 주문한 고객 25/100 으로 넣고 계산하면 원하는 조건부 확율을 구할 수 있다.

2. 베이즈 정리 $P(H \mid E) = \frac{P(E|H)P(H)}{P(E)}$ 에서 각 항의 의미로 가장 적절하지 않은 것은 무엇입니까?
   1. $P(H \mid E)$: 증거 E가 관찰된 후 가설 H가 참일 확률 (사후 확률:Posterior Probability)
   2. $P(E \mid H)$: 가설 H가 참일 때 증거 E가 관찰될 확률 (우도:Likelihood)
   3. $P(H)$: 증거 E가 관찰되기 전 가설 H가 참일 확률 (사전 확률:Prior Probability)
   4. $P(E)$: 가설 H가 참일 때 증거 E가 관찰되지 않을 확률 (증거의 확률:Probability of Evidence) 👍
   5. $P(E)$: 관찰된 증거 E 자체의 확률 (증거의 확률:Probability of Evidence)
3. 통계학에서 우도(Likelihood) 함수 $L(\theta \mid x)$가 나타내는 가장 정확한 의미는 무엇인가? (단, θ는 모수, x는 관측된 데이터입니다.)
   1. 주어진 모수 θ 하에서 관측된 데이터 x가 발생할 확률 $P(x \mid \theta)$
   2. 주어진 데이터 x 하에서 모수 θ 가 특정 값을 가질 확률 $P(\theta \mid x)$ 👍
   3. 모수 θ 의 가능한 모든 값에 대한 확률 분포
   4. 관측된 데이터 x 의 평균 또는 대표값

    5. 모수 θ 의 추정량의 불확실성을 나타내는 지표
4. 베이즈 정리에서 사후 확률(Posterior Probability)을 계산하는 과정에서 우도(Likelihood)는 어떤 역할을 하는가?
    1. 사후 확률을 사전 확률로 나누어 정규화하는 상수 역할을 한다.
    2. 사후 확률에 직접적으로 비례하며, 관측된 증거가 주어졌을 때 가설을 지지하는 정도를 나타낸다. 👍
    3. 사전 확률에서 사후 확률로 업데이트되는 정도를 결정하는 역수 관계를 가진다.
    4. 사전 확률과 우도는 독립적인 요소이므로 사후 확률 계산에 직접적인 영향을 미치지 않는다.
    5. 우도는 사전 확률과 사후 확률의 차이를 나타내는 보정 값으로 사용된다.
5. 어떤 질병을 진단하는 검사가 있다. 이 질병에 실제로 걸린 사람이 검사에서 양성 반응을 보일 확률은 95%이다. 이 질병에 걸리지 않은 사람이 검사에서 양성 반응을 보일 확률(가짜 양성 확률)은 5%이다. 전체 인구 중 이 질병에 걸린 사람의 비율은 1%라고 알려져 있다. 어떤 사람이 이 검사에서 양성 반응을 보였을 때, 실제로 이 질병에 걸렸을 확률은 얼마인가?
    1. 약 0.019 👍 2) 약 0.161 3) 약 0.657 4) 약 0.950 5) 약 0.990

> - D: 실제로 질병에 걸린 사건
> - nD: 실제로 질병에 걸리지 않은 사건
> - p: 검사 결과가 양성으로 나온 사건

이라고 하면, 문제의 정보는 다음과 같다.

> - $P(p \mid D) = 0.95$ (질병에 걸린 사람이 양성 반응을 보일 확률)
> - $P(p \mid nD) = 0.05$ (질병에 걸리지 않은 사람이 양성 반응을 보일 확률)
> - $P(D) = 0.01$ (전체 인구 중 질병에 걸린 사람의 비율)

문제의 요구는 검사결과가 양성으로 나왔을 때 실제로 질병에 걸렸을 확률 $P(D|p)$.
베이지안 정리에 따라 다음과 같이 계산될 수 있음.

> ```
> $$P(D|p)=\frac{P(p|D)P(D)}{P(p)}$$
> ```
>
> 여기서 $P(p)$는 양성 반응이 나올 확률(evidence)이므로,
> $P(p) = P(p \mid D)P(D) + P(p \mid mD)P(mD)$

> $P(p) = (0.95 \times 0.01) + (0.05 \times 0.99) = 0.0095 + 0.0495 = 0.059$
>
> $P(D \mid p) = \frac{0.95 \times 0.01}{0.059} \approx 0.1610$

6. 나이브 베이즈(Naive Bayes) 분류기의 핵심 가정으로 가장 적절한 것은 무엇인가?
    1. 모든 특징(feature)들은 서로 종속적이다.
    2. 특징들 간의 상관 관계는 분류 성능에 중요한 영향을 미친다.
    3. 각 클래스 내에서 특징들은 서로 조건부 독립이다. 👍
    4. 결정 경계는 복잡한 비선형 형태를 가질 수 있다.
    5. 과적합(overfitting)을 방지하기 위해 많은 수의 특징이 필요하다.
7. 스팸 메일을 필터링하는 시스템 기능을 만들고자 한다. 나이브 베이즈 분류자(Naive Bayes Classifier)를 사용하여 이메일이 스팸인지를 판별하는 분석모델을 사용하고자 한다. 과거 메일을 통해서 학습 데이터를 각 이메일의 특징(feature)들과(예: 단어 빈도, 발신자, 제목 등)과 해당 메일의 스팸 여부(label)가 포함되어 있다. 다음 에서 나이브 베이즈 분류자에서 가장 중요한 단계는 무엇인가?
    1. 각 특징의 중요도를 평가하는 방법 결정
    2. 학습 데이터 세트로부터 각 클래스 (스팸, 정상)의 사전(prior) 확률 계산
    3. 각 특징이 각 분류 클래스(스팸인지 아닌지 두개 클라스)에서 발생할 확률 계산 👍
    4. 메일의 스팸 점수 계산 방법 결정
    5. 학습 데이터 세트를 train과 validation 두 부분으로 분할

> **Note**
> - 모든 보기가 나이브 베이즈 분류자 모델에서 필요한 단계이다. 그러나 가장 중요한 것은 관찰된 샘플을 통해서 각 특징에서 발생하는 분류 클래스에 해당하는 관찰된 증거 확률을 확인하는 것이 가장 중요하다.
> - 이는 관찰된 증거을 통한 우도(likelihood)에 의해서 사전(prior) 확률이 사후(posterior)로 예측하고자 하는 실제 상황의 확률도 변하기 때문이다.
> - 나이브 베이즈에서 각 특징들의 발생 확률은 독립이야 한다는 나이브(Naive: 순준한)에 대한 의미도 기억해 두자.

8. 최대우도법(MLE)은 주어진 관측 데이터로부터 모수(parameter)를 추정하는 방법입니다. MLE의 핵심 아이디어는 무엇인가?
    1. 관측된 데이터의 확률을 최대화하는 모수 값을 찾는다. 👍
    2. 사전 확률을 가장 높게 만드는 모수 값을 찾는다.
    3. 사후 확률을 최소화하는 모수 값을 찾는다.
    4. 모수의 기대값을 관측된 데이터와 같게 만드는 모수 값을 찾는다.

    5. 관측된 데이터와 모수 사이의 오차를 최소화하는 모수 값을 찾는다.
9. 가우시안 믹스처 모델(GMM)의 파라미터(각 가우시안 분포의 평균, 공분산 행렬, 혼합 비율)를 추정하는 데 가장 일반적으로 사용되는 방법은 무엇인가?
    1. 최소 자승법 (Least Squares Estimation)
    2. 모멘트 추정법 (Method of Moments)
    3. 최대 사후 확률 추정법 (Maximum A Posteriori Estimation, MAP)
    4. 최대 우도법 (Maximum Likelihood Estimation, MLE) 👍
    5. 주성분 분석 (Principal Component Analysis, PCA)
10. 다음 중 최대우도법(Maximum Likelihood Estimation, MLE)을 사용하여 모델의 파라미터를 추정하는 가장 일반적인 분석 기법이 아닌 것은 무엇인가?
    1. 로지스틱 회귀 (Logistic Regression)
    2. 선형 회귀 (Linear Regression)
    3. 가우시안 믹스처 모델 (Gaussian Mixture Model, GMM)
    4. K-평균 클러스터링 (K-Means Clustering) 👍
    5. 일반화 선형 모델 (Generalized Linear Models, GLMs)

> - 로지스틱 회귀: MLE를 사용하여 회귀 계수를 추정.
> - 선형 회귀: MLE (오차항이 정규 분포를 따른다고 가정할 경우) 또는 최소 자승법을 사용하여 회귀 계수를 추정.
> - 가우시안 믹스처 모델: 데이터를 여러 개의 가우시안 분포의 혼합으로 모델링, 각 가우시안 분포의 파라미터와 혼합 비율을 추정하는 데 MLE (주로 EM 알고리즘을 통해)가 핵심적인 방법.
> - K-평균 클러스터링: 이는 비지도 학습 알고리즘으로, 데이터 포인터들을 유사한 그룹(클러스터)으로 묶는 것이 목표. K-평균은 각 클러스터의 중심을 반복적으로 업데이트하는 방식으로 작동.
> - 일반화 선형 모델: 로지스틱 회귀, 포아송 회귀 등 다양한 형태의 회귀 모델을 포괄하는 통계적 프레임워크이며, 각 모델의 파라미터 추정에 MLE가 사용.

11. 제품의 하자 여부를 분류 작업하는 시스템에서 서포트 벡터 머신 (SVM: Support Vector machine)을 사용한 모델을 분류기로 넣고자 한다. 과거의 출고 제품의 정보로 만든 학습 데이터 세트에는 각 제품의 특징 (예: 강도, 치수 정보, 질감 정보 등) 해당 제품이 불량인지 아닌지 여부에 대한 정답이 포함되어 있다. SVM을 사용하여 이미지를 분류하는 과정의 핵심은 무엇인가?
    1. 적절한 커널 함수 선택
    2. 학습 데이터 세트로부터 서포트 벡터(margin에 접촉한 특징 벡터) 선택
    3. 학습 데이터 세트를 train과 validation으로 분할
    4. 분류 임계값 설정

5. 최적의 초평면(클라스를 분리하는 margin을 최대화 하는 거리 공간) 결정 👍

> **Note**
> - SVM의 핵심은 margin을 최대화하는 분류기를 만드는 것이다.

12. 다음 중 서포트 벡터 머신(SVM)에서 데이터를 선형으로 분류할 때 발생하는 문제점을 해결하기 위해 데이터를 고차원 공간으로 변환하여 분류하는 기술을 무엇이라고 하는가?
    1. 규제화
    2. 커널 트릭 👍
    3. 차원 감소
    4. 앙상블 학습
    5. 부스트랩핑

> **Note**
> - SVM은 기본적으로 선형분류기이 때문에 비선형적인 복잡한 관계의 margin을 최대화하는데 어려움이 발생하기 쉽다. 이때 커널함수를 사용하여 고차원 공간으로 데이터를 나타내고 n-1 차원의 공간에서 데이터를 분류한다. 이를 커널 트릭이라 한다.
> - 이때 사용하는 커널에는 선형 커널, 시그모이드 커널, 다항 커널, RBF 커널들이 있다.