[AI 머신러닝, 딥러닝 이론]

1. **베이즈 이론 및 베이즈 분류기에 대해 설명하시오.**
   - 베이즈 이론 : 훈련데이터를 이용해서 특징 값이 제공하는 증거를 기반으로 결과가 관측될 확률을 계산.

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap B^c)}$$

$$= \frac{P(B) \cdot P(A \mid B)}{P(B) \cdot P(A \mid B) + P(B^c) \cdot P(A \mid B^c)}$$

   - 베이즈 분류기 : 미래의 사건을 추정하기 위해 과거의 사건 데이터를 사용 하는 **확률적 학습** 기반 필터기. 결과에 대한 전체 확률을 알기 위해 동시에 여러가지 속성에 대한 정보를 고려해야 하는 문제에 베이지안 분류기 사용이 적합하다

2. **이메일에 '나이트'라는 단어가 검출되었을 때의 스팸 확률을 수식으로 작성하시오. (베이즈 이론 수식에서 사후확률을 기술하시오)**
   - P(스팸|나이트) = P(나이트|스팸) * P(스팸) / P(나이트)
   - 사후확률 = (우도 * 사전확률) / 주변우도

3. **나이브 베이지안 분류기가 응용되는 예를 드시오.**
   - 스팸 메일 필터링(텍스트 분류)
   - 네트웍 침입/비정상행위 탐지(IDS)
   - 일련의 관찰된 증상에 대한 의학적 질병 진단

[Tensorflow 머신러닝 (DNN,CNN,RNN,GAN)]

1. **recall, f-measure, precision, accuracy에 대해 설명하시오.**

   위의 4가지는 confusion matrix를 사용하여 분류성능을 평가하는 지표이다.
   - recall(재현율) : 실제 True인 것 중에서 True라고 예측한 것의 비율

$$(Recall) = \frac{TP}{TP + FN}$$

   - f-measure : F1-Score라고도 부름. Precision과 Recall의 조화평균. 모델의 성능 평가 지표

$$(F1\text{-}score) = 2 \times \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- precision(정밀도) : True라고 예측한 것 중에서 실제 True인 것의 비율

$$(Precision) = \frac{TP}{TP + FP}$$

- accuracy(정확도) : 실제 True를 True라고 예측한 것, 실제 False를 False라고 예측한 것의 비율

$$(Accuracy) = \frac{TP + TN}{TP + FN + FP + TN}$$

2. **cross-validation에 대해 설명하시오.**
   - 정의 : 교차 검증.
   - 활용 :
     - 모델이 특정 데이터 셋에 overfitting 되는 것을 방지
     - 데이터가 부족할 때 underfitting되는 것을 방지
     - 일반화 된 모델을 만들어 정확도를 향상 시킴

3. **과적합이 발생되는 이유와 해결 방법을 기술하시오.**
   - 정의 : 특정 데이터에 대해 모델훈련이 아주 잘 되어 그 외의 데이터에는 정확도가 떨어지는 경우
   - 발생 원인
     - 데이터 편향 : 학습 데이터가 실제 데이터의 편향된 부분만 가지고 있을 때
     - 데이터 오류 : 학습 데이터에 오류가 포함 되어 있을 때
     - 변수가 지나치게 많을 때
     - 모델이 너무 복잡할 때
   - 해결 방법 :
     - 데이터 양을 늘리기
     - 변수 개수 줄이기
     - 모델 복잡도 줄이기
     - 가중치 규제
     - 드롭아웃 : 딥러닝의 경우 신경망 일부를 사용하지 않기(랜덤 노드를 제거하고 학습한다.)

[Keras 머신러닝 (DNN,CNN,RNN,GAN)]

선형 회귀 분석 수행시, 다음 함수 및 기법에 대해 설명하시오.

1. **가설 함수** : 어떤 모집단의 모수에 대한 잠정적인 가설을 세워 예측한 값. 선형회귀의 가설함수는 '데이터들을 표현할 수 있는 직선이 존재한다'이다.

$$H(x) = \omega x + b$$

2. **분석 알고리즘** : 일차함수의 개념인 직선을 바탕으로 예측하는 알고리즘. 데이터 분포를 가장 잘 설명할 수 있는 직선 함수르 찾아내는 알고리즘이다.

3. **cost 함수** : 비용함수. 가설과 실제 값의 차이를 제곱하여 평균 낸 것. cost를 최소화하는 weight를 찾아야 한다.

$$cost(W) = \frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

4. **경사하강법** : 함수값이 최소가 되게 하는 w와 b를 찾는 방법. 함수의 기울기(경사)를 구하고 경사의 절댓값이 낮은 쪽으로 계속 이동시켜 극값에 이를 때까지 반복힌다.

$$W := W - \alpha\frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})x^{(i)}$$

> *기울기 = cost 함수를 미분한 값*
> *α = 한 번에 움직이는 step size = learninng rate*