인공지능 설계 및 실습 과제 답안지

소속(학과)	IoT인공지능융합전공
학번	15****
이름	양승우

질문1.	

1) 위의 원문을 읽고 1번 부터 12번까지의 변수가 의미하는 것이 무엇인지 서술하세요.

1. ID number :
   환자 식별 번호
2. Diagnosis (M = malignant, B = benign) :
   진단 (M = 악성, B = 양성)
3. radius (mean of distances from center to points on the perimeter) :
   반지름 (중심점으로부터 둘레까지의 평균 거리)
4. texture (standard deviation of gray-scale values) :
   질감 ("Gray-Scale Values"의 표준편차)
5. perimeter :
   둘레
6. area :
   면적
7. smoothness (local variation in radius lengths) :
   매끄러움 (반지름 길이의 국소 변화)
8. compactness (perimeter^2 / area - 1.0) :
   조밀도 (稠密度; $둘레^2 \div$ 면적 – 1.0)
9. concavity (severity of concave portions of the contour) :
   오목함 (윤곽의 오목한 부분의 정도)
10. concave points (number of concave portions of the contour) :
    오목한 점의 수 (윤곽의 오목한 부분의 정도의 갯수)
11. symmetry :
    대칭
12. fractal dimension ("coastline approximation" - 1) :
    프랙탈 차원 ("해안선 근사치" - 1)

질문2.	

1) 예측하고자 하는 타겟 변수인 M과 B의 갯수가 균등하지 않을 경우 어떠한 문제점이 발생할 수 있는지 기술하기 바랍니다.

편향되고 부정확한 결과를 초래하게 됩니다.
일반적인 기계학습 기법들은 학습데이터가 범주별로 비슷한 비율로 구성되어 있다고 가정하고 학습을 진행하게 됩니다. 그러나 많은 실세계 문제들이 불균형 데이터(imbalanced data) 문제에 속하게 되고 이러한 경우 소수 범주에 속한 데이터들은 다수 범주에 속한 데이터보다 잘못 분류될 가능성이 높습니다.
이와 같이 데이터의 분포가 불균형한 상태에서 학습을 진행하게 되면 인식기는 훈련 데이터에서 차지하는 **빈도가 높은 데이터에 과적응(overfitting)**하게 되는 문제가 발생하게 됩니다.

이러한 부작용(side effect)은 기계학습 알고리즘의 설계 특성상 각 범주의 상대적인 분포를 고려하는 대신 전반적인 성능을 최적화 시키려하기 때문에 발생하는 것으로 결정트리(decision tree)나 다층 퍼셉트론(multilayer perceptron)과 같은 분류기에서 흔히 나타납니다.

해결 방법
불균형 데이터를 다루는것은 머신 러닝 알고리즘에 넣기 전에,
분류 알고리즘을 향상시키거나 **Training-data 클래스의 균형을 맞추는 작업**이 수반됩니다.

* 균형을 맞추는 작업

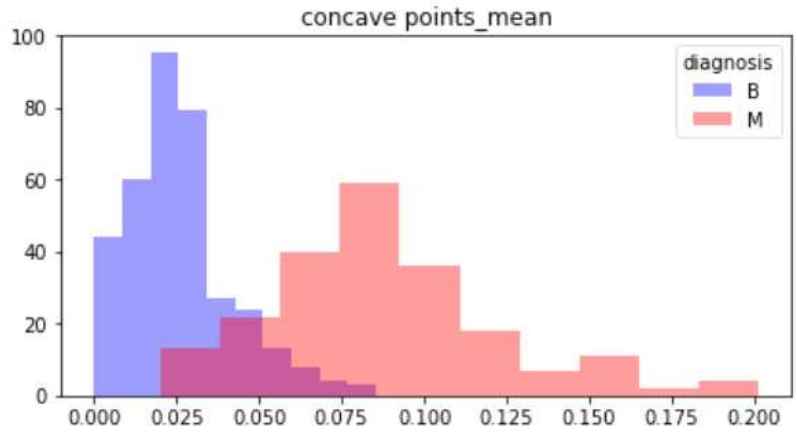
① Class weights
Impose a heavier cost when errors are made in the minority class
② Down-sampling
Randomly remove instances in the majority class
③ Up-sampling
Randomly replicate instances in the minority class
④ Synthetic minority sampling technique (SMOTE):
Down samples the majority class and synthesizes new minority instances by interpolating between existing ones

질문3.	

1) 모든 변수에 히스토그램을 통하여 각 변수 중 악성과 양성을 구분하기에 적절할 것 같은 변수가 무엇인지 그리고 그렇게 생각하는 이유를 서술하세요.

Variable of **Concave Points**

겹치는 히스토그램의 면적 값이 가장 작기 때문이라고 생각합니다.
다시 말하면, 겹치는 부분이 적을수록 분류가 더 잘 됐다고도 말할 수 있습니다.



질문4.	

1) 알고리즘을 통한 데이터 분석에서 다른 변수가 나머지 변수에 비해 값이 압도적으로 클 경우 어떠한 문제가 발생할 수 있는지에 대해 기술하세요.

일반적으로 정확한 비교가 힘들고, 특징의 범위가 크게 다르면 모델은 각 특징의 중요성을 오인하기 마련입니다.
또한, 데이터의 값이 너무 크거나 혹은 작은 경우에 모델 알고리즘 학습 과정에서 0으로 수렴하거나 무한으로 발산해버릴 수 있습니다.

Data Scaling
특징 값의 범위를 균일하게 맞춰주는 작업

(1) StandardScaler
각 feature의 평균을 0, 분산을 1로 변경합니다.
모든 특성들이 같은 스케일을 갖게 됩니다.

(2) RobustScaler
모든 특성들이 같은 크기를 갖는다는 점에서 StandardScaler와 비슷하지만,
평균과 분산 대신 median과 quartile을 사용합니다.
RobustScaler는 이상치에 영향을 받지 않습니다.

(3) MinMaxScaler
모든 feature가 0과 1사이에 위치하게 만듭니다.
데이터가 2차원 셋일 경우, 모든 데이터는 x축의 0과 1 사이에,
y축의 0과 1사이에 위치하게 됩니다.

(4) Normalizer
StandardScaler, RobustScaler, MinMaxScaler가 각 columns의 통계치를
이용한다면 Normalizer는 row마다 각각 정규화됩니다.
Normalizer는 유클리드 거리가 1이 되도록 데이터를 조정합니다.
(유클리드 거리는 두 점 사이의 거리를 계산할 때 쓰는 방법, L2 Distance)

질문5.	

1) 상관계수란 무엇이고, 그 값이 의미하는 것이 무엇인지 서술하세요.

상관계수(correlation coefficient, 相關係數)

두 변수간의 연관성을 보여주는 지표입니다.
-1에서 1사이의 값을 가지며, 변수와의 방향은 (-)와 (+)로 표현합니다.
값이 1에 가깝다면 두 변수의 움직임이 완전히 같다는 뜻이며 –1이면 움직임이 완전히 역방향임을 의미합니다.
0에 가까울수록 두 변수 사이에 관련이 없다고 볼 수 있습니다.

2) 또한 상관계수가 높다고 해서 좋은 변수인지,
낮다고 해서 안좋은 변수인지에 대해 생각해 보세요.

"좋다는 것"이 "중요하다는 것"이라고 생각했을 때,
상관계수가 낮다고 하더라도 다른 관점, 분석기법을 적용한다면 충분히 새롭고 중요한 가치를 창출해 낼 수 있다고 생각합니다.

3) 상관계수가 높다하여 인과관계가 있다고 할 수 있을지에 대해 생각해 보세요.

없을 가능성이 높지만, 인과관계가 있다고도 말할 수 있다.

예를 들어 생각해보면, 나는 선글라스를 쓰면 멋있다고 생각한다.
그래서인지 선글라스를 쓰는 날에는 여자들이 쳐다보는 횟수가 증가한다.
그래서 나는 선글라스를 쓴 나의 멋진 모습이 여자들이 쳐다보는 원인이라고 생각했지만, 둘리의 마이콜을 닮아 신기해서 쳐다볼 수 있고, 수상해 보여 경계하느라 쳐다볼 수도 있고, 혹은 정말로 멋있어서 쳐다볼 수도 있기 때문이다.

4) 마지막으로 현재 주어진 0.9 이상인 4개의 상관계수가 높게 나왔는지를
변수들을 이용하여 해석하세요.

	Row	Column	Value	abs_Value
0	radius_mean	perimeter_mean	0.997855	0.997855
1	radius_mean	area_mean	0.987357	0.987357
2	perimeter_mean	area_mean	0.986507	0.986507
3	concavity_mean	concave points_mean	0.921391	0.921391

모두 강한 양의 상관관계를 가지고 있습니다.

질문6.	

1) 데이터 분석에 들어가기 이전에 데이터 학습에 따라
'과대적합(overfitting)'과 '과소적합(underfitting)'이라는 문제가 발생할 수 있습니다.
각각의 문제가 무엇이고, 어떻게 해석해야 하는지에 대해 서술하시길 바랍니다.

과대적합 (overfitting)
제한된 샘플에 너무 특화가 되어 새로운 샘플에 대한 예측 결과가 오히려
나빠지거나 학습의 효과가 나타나지 않는 경우

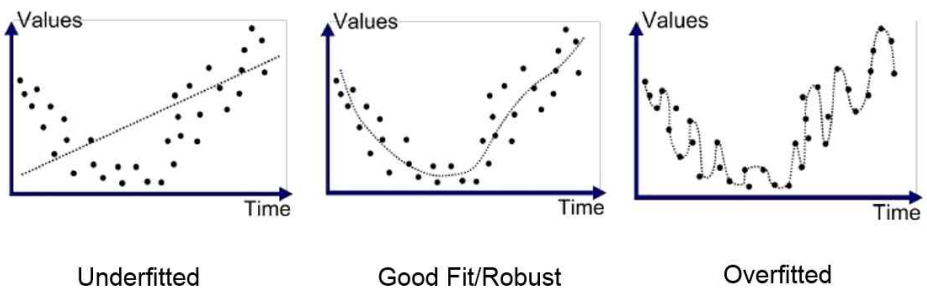
@ 해결/해석
파라미터 수가 적은 모델을 선택하거나, 모델에 제약을 가하여 단순화시킵니다.
훈련 데이터를 더 많이 확보합니다. 훈련 데이터의 잡음을 줄입니다.

과소적합 (underfitting)
제한된 샘플에 의해 형성된 모델이 너무 단순해서 새로운 샘플에 대한 예측 결과가
나쁘거나 학습의 효과가 나타나지 않는 경우

@ 해결/해석
파라미터가 더 많은 강력한 모델을 선택합니다.
학습 알고리즘에 더 좋은 특성을 제공합니다. 모델의 제약을 줄입니다.

공통적으로는 더 많은 데이터를 사용합니다.

이 외에,
① Early stopping
① Data Argumentation (Affine transform)
② Drop-out, Ensemble
③ Tuning Hyper-parameters



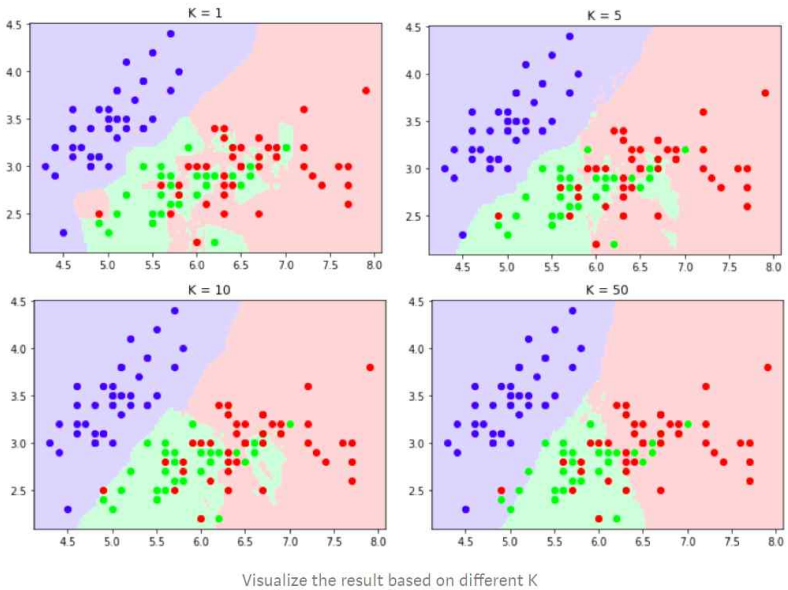
질문7.	

1. 위의 이웃의 갯수에 따른 그래프와 표를 과대적합과 과소적합의 관점에서 결과를 해석하세요.

k 값이 작을 때 : Overfitting
training-data에 대한 정확도는 높지만, test-data에 대한 정확도는 낮습니다.

k 값이 클 때 : Underfitting
모델이 단순화되어 training-data와 test-data 모두에 대하여 정확도가 떨어집니다.

제시된 데이터는 k=6 일 때, 최적점을 갖는다고 할 수 있습니다.



Visualize the result based on different K

(참고 그림)

질문8.	

1. 지니계수와 엔트로피 계수가 갖는 의미를 찾아보고 서술하세요.

분류모델의 비용함수(Cost function) → 불순도(heterogeneity) 측정

Decision tree는 구분 뒤 각 영역의 순도(homogeneity)가 증가,
불순도(heterogeneity)가 최대한 감소하는
방향으로 학습합니다.

불순도의 척도 → ① Entropy ② Gini ③ Misclassification

모든 샘플의 부류가 같으면 불순도는 0이고
서로 다른 부류가 많이 섞여 있을수록 불순도가 높습니다.

① Entropy

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

② Gini

$$Gini(S) = 1 - \sum_{i=1}^{c} {p_i}^2$$