Machine Learning, Fall 2022 - Homework 2

- 이름 (학번):
- 학과:

1 Short questions

1) Supervised learning과 Unsupervised learning 방식의 차이에 대해 설명하고 classification은 어느 쪽에 속하는지 쓰시오.

2) 기계학습에서 데이터를 training/validation/test set 세 부분으로 나누어야 하는 이유에 대해 설명하시오 (즉, 각 데이터 부분의 용도)

3) 기계학습에서 overfitting(과적합)이 의미하는 바를 쓰시오.

2 2차원 상의 데이터 $\mathbf{x} = (x_1, x_2)^T$ 를 고차원으로 매핑하는 다음의 함수를 가정하자.

$$\varphi(\mathbf{x}) = \begin{pmatrix} \varphi_1(\mathbf{x}) \\ \varphi_2(\mathbf{x}) \\ \varphi_3(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix}$$

1) 두 벡터 $\mathbf{u} = (u_1, u_2)^T$와 $\mathbf{v} = (v_1, v_2)^T$를 φ로 각각 매핑한 벡터의 내적에 해당하는 kernel function $k(\mathbf{u}, \mathbf{v})$를 구하시오.

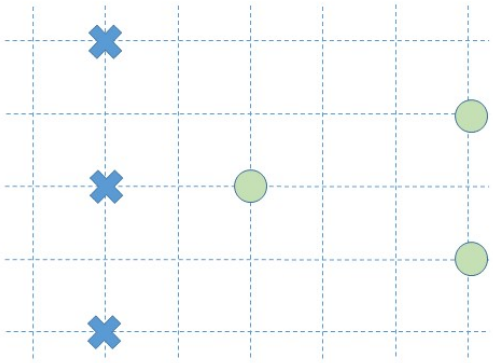
2) SVM에서 kernel trick이 무엇인지 그 장점의 측면에서 서술하시오.

3 Random Forest

1) Random forest의 주요 하이퍼마라미터 두 가지를 골라 그 의미를 각각 설명하시오.

2) Random forest가 기본 decision tree ensemble(bagged tree) 모델과 구별되는 가장 큰 특징을 한 가지 서술하고 그 특징으로 인한 장점을 설명하시오.

4 다음 2차원 평면 상의 6개의 데이터 샘플에 대해 분류 알고리즘을 적용하기로 하자. 각 샘플의 모양은 두 클래스(o, x) 중 어디에 속하는지를 나타낸다. KNN 알고리즘에서 흔히 사용되는 유클리드 거리를 사용한다고 가정하자.



1) 1-NN을 사용하는 경우 training error를 쓰고 이유를 간략히 설명하시오.

2) 3-NN을 사용하는 경우 training error를 쓰고 이유를 간략히 설명하시오.

3) 5-NN을 사용하는 경우 LOOCV error를 쓰고 이유를 간략히 설명하시오.

4) Linear SVM을 사용하는 경우 LOOCV error를 쓰고 그 이유를 생성되는 Decision boundary들을 그림에 표시하여 설명하시오.

5 0과 1의 두 가지 클래스로 이루어진 이진 분류(binary classification) 문제를 풀기 위한 모델을 학습하여 다섯 개의 테스트 샘플에 적용한 결과, 각 샘플별로 클래스 1에 속할 확률이 아래와 같이 구해졌다고 가정하자.

ID	P(y=1|x)	True Label
1	0.95	1
2	0.85	0
3	0.75	1
4	0.65	1
5	0.45	0
6	0.35	1
7	0.15	0

1) P(y=1|x)>0.7인 경우에 클래스 1로, 아닌 경우 0으로 예측하기로 할 경우의 Sensitivity, Specificity를 구하시오(단, 정답 클래스는 위 테이블의 'True Label' 열에 주어져 있다).

2) 위 표를 기반으로, 클래스 1로 예측하는 기준이 되는 threshold (예를 들면 1)번에서는 0.7)를 바꾸어 가면서 반복적으로 Sensitivity, Specificity를 구할 경우 ROC curve를 그릴 수 있다. 주어진 결과에 해당하는 ROC curve 상의 (0, 0)과 (1, 1)을 제외한 주요 좌표 6개를 적고 AUC(Area under the ROC curve) 값을 구하시오.

3) (True or False) ROC curve가 (0,1)점을 지나는 경우 그 분류기는 ROC curve를 만드는 데 사용된 테스트 데이터들을 항상 완벽하게 분류한다 (즉, error = 0).