**Machine Learning & Data Mining, Spring 2020**

**Homework 3**

Due April 16

이름: 김도운

학과: 사이버보안학과

학번: 201520893

* True or False (Explain your answer).

1. In KNN, the training error is minimized when we set k = 1

* **True**. k=1이면, 자기 자신을 가장 가까운 샘플로 선택하기 때문에 training error는 항상 0이 되므로 k=1일 때 training error는 최소화된다.

1. In KNN, the training error is always 0 when we set k = 1

* **True.** training error란 test 단계에서 training set을 사용했을 때의 error로, k=1일 때는 자기 자신을 가장 가까운 샘플로 고르게 되고 당연히 training error는 0이 된다.

1. In KNN, the model with a larger value of K will have a smaller variance.

* **True.**  k가 커질수록 test error는 줄어드므로, variance는 낮아진다.

1. In KNN, the training time is much longer than the test time (i.e. computation time needed for testing each test sample).

* **False**. Lazy training이라고도 불리는 KNN은 training 단계에서 하는 작업은 없고 test 단계에서 다른 sample들과의 distance를 계산한다.

1. The entropy of a uniform distribution (e.g. prob=[1/C, … ,1/C] for C>0 possible classes) is always 1.

* **False.** 공식에 의하면, 이다. 따라서 c가 2라면 엔트로피는 1이지만 c가 2가 아닌 경우에는 1이 아닐 수 있다.

1. The test time for the decision tree algorithm does not depend on the tree depth.

* **False**. Test 단계에서 root 노드부터 시작해 leaf node로 도달해야 분류가 되므로, 깊이는 test time에 영향을 미친다.

1. Decision tree algorithm partitions the feature space into distinct regions such that samples belonging to the same region have the same predicted value.

* **True**. Decision tree의 node들을 거치다 보면 각 조건에 맡게 영역별로 구분되며, 각 영역들은 같은 예측값을 갖는 데이터들로 이루어진다.

1. Random Forest algorithm does not use the bootstrapping idea.

* **False**. Bootstrapping은 training set으로부터 여러 중복을 허용하는 여러 부분집합들을 생산하는 것으로, ensemble 기법의 한 종류인 random forest 알고리즘에서 사용된다.

1. In Random Forest algorithm, all the features are always used to construct each decision tree classifier.

* **False**. 모든 feature들을 사용하는 경우 특정 압도적인 feature에 의해 다른 feature들이 의미가 없어질 수도 있기 때문에 bagging된 tree들의 결과가 다 비슷비슷해질 수 있다. 따라서 random하게 전체 feature중에 m개의 feature만 사용해서 decision tree를 구성한다.

1. Support vector machines can only be applied to linearly separable data.

* **True**. SVM은 선형으로 분리되는 데이터가 아닌 경우에도 적용할 수 있다. 비선형으로 분류되는 데이터의 경우 차원을 높여서 선형으로 분리시켜줄 수 있기 때문이다.

1. The choice of a kernel function in SVM does not affect the prediction accuracy.

* **False.** Kernel 함수를 이용해 계산함으로써 고차원에서의 내적, 계산 등은 알 필요가 없지만, 몇 개의 유명한 커널 함수들을 이용해 계산해본 뒤 알맞은 커널 함수를 선택하는 것은 필요하다.

1. The main idea of SVM is to minimize the margin.

* **False.** Margin이 너무 작은 경우 약간의 noise나 데이터의 작은 차이로 인해 잘못 분류되는 경우가 있기 때문에, margin을 최대화하는 것이 SVM의 주 목적이다.