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

**Homework 3**

Due April 16

* True or False (Explain your answer).

1. In KNN, the training error is minimized when we set k = 1 **(T)**

KNN알고리즘에서 K=1 이면 가장 가까운 data를 선택하게 되므로 training data로 평가하게 되면 training error는 0이다. 즉, training error가 최소화된다.

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

KNN알고리즘에서 training data로 평가할때, K=1 이면 가장 가까운 data인 자기 자신을 선택하게 되므로 training error는 0 이다.

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

KNN에서 k값이 클수록, variance(분산)는 감소하지만 bias(편향정도)는 증가한다. 즉, 안정적이게 된다.

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

KNN은 학습시키는 알고리즘이 아니기때문에, training time과 test time이 동일하다.

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

확률 또는 확률밀도가 특정 값에 몰려 있으면 엔트로피가 작다고 하고, 반대로 여러가지 값에 골고루 퍼져 있다면 엔트로피가 크다고 한다. 하지만 연속 균등 분포의 entropy가 항상 1이 된다고 할 수는 없다. 또한 엔트로피를 구하는 공식에서 엔트로피는 (1/C)\*log2(C) + … +(1/C)\*log2(C)이고, 연속 균등 분포에서 각 값의 확률은 1/n이므로 각 값을 더하면 log2(C)가 된다. 즉, 연속 균등분포의 entropy는 항상 1이 아니다.

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

**(F)**

Decision tree는 각 노드에서 feature에 대한 조건으로 data 평가한다. 그래서 단말 node에서 분류를 마치게 된다. 즉, tree depth에 따라서 test time이 증가한다.

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

Decision tree는 decision boundary로 각각의 구역을 분리하는데, 이때, 나누기 전과 나눈 후의 차이가 가장 컸던 feature를 선택하면서 공간을 분할한다. 즉, 같은 구역에 있으면 같은 레이블로 예측된다.

1. Random Forest algorithm does not use the bootstrapping idea. **(F)**

Bootstrapping idea는 전체 feature중 일부 feature를 뽑아서 tree의 입력 값으로 넣어서 학습시키는 방식이다. Random forest는 bootstrapping idea를 사용해서 여러 개의 decision tree를 만들고, 그 결과값들 중에 가장 흔하게 발생하는 결과를 선택하는 알고리즘이다.

즉, bootstrapping idea를 사용한다.

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

8번에서 설명했듯이, random forest는 bootstrapping idea를 사용해서 모든 feature중에 일부 feature를 선택해서 사용한다.

1. Support vector machines can only be applied to linearly separable data. **(F)**

SVM은 선형적으로 분리된 data뿐만 아니라, 선형적으로 분리되지 않은 data에도 적용된다. 이때, slack변수를 사용하는 방식 또는 kernel Trick을 사용한다.

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

적절하게 커널 함수를 선택할수록, 적절한 decision boundary를 결정할 수 있기때문에 kernel함수의 선택이 prediction accuracy에 영향을 미친다.

1. The main idea of SVM is to minimize the margin. **(F)**

SVM은 margin을 최대화해서 잘못된 test 결과가 발생하지 않도록 하는 것이main idea이다.