**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

트레이닝 에러는 트레이닝 데이터로 테스트를 했을 때 일어나는 오류를 일컫는데, KNN알고리즘에서는 트레이닝 단계를 거치지 않고, 바로 테스트 단계로 갑니다. k=1이라면, 가장 가까운 인스턴스는 자기자신이 됩니다. 그래서 traning error는 항상 0일수 밖에 없습니다.

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

T

1번과 같은 이유입니다.

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

T

K가 작을수록 데이터가 가장 가까운 점에 붙는 경향이 있어 전체 데이터에 대해서는 계속 모아지지 않고 멀리 퍼지게 됩니다. 그래서 k가 작을 때는 분산정도가 큽니다. k가 커지면 이런 현상이 누적되면서 편향정도(bias)가 커지게 되는 것입니다.

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

F

KNN알고리즘은 트레이닝 단계에서 아무것도 하지 않습니다.

대신 테스트 단계에서 모든 계산을 수행하게 됩니다.

시간 복잡도는 p를 feature의 수라하고, n을 sample수라 했을 때,

O(np)가 됩니다.

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

F

Entropy(S) = = C \* 이므로 가 엔트로피이고, 이는 c가 2일때만 1입니다 만약 binary classification이 아니라면 엔트로피는 1이 아니라 1이상의 값을 가질 수 있습니다.

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

F

root node부터 branch를 따라 내려가면서 최종으로 도착한 leaf node의 class로 예측을 하기 때문에 depth가 깊을수록 traversal의 시간이 증가하기 때문에, depth에 비례합니다.

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

training phase에서 나누기 전과 나눈 후의 엔트로피 차이가 가장 컸던 feature를 선택하면서 공간을 분할해 나갑니다. 즉, 같은 구역에 있으면 같은 레이블로 예측이 됩니다.

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

F

Random Forest 알고리즘은 2가지 중요한 아이디어를 기반으로 하는데 하나가 bootstrapping이고 다른 하나는 averaging입니다.

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

F

Begging에서는 모든 feature가 트리 구축에 사용되지만, Random Forest에서는 m개 만을 사용하여 트리를 구축하는데 (일반적으로 ) 이유는 지대한 영향을 끼치는 feature가 있을 경우 결과로 나온 tree들의 모양이 비슷할 수 있기 때문입니다. 이는 variance를 줄이는데 도움이 되지못합니다. 왜냐면 모양이 비슷하면 특이한 데이터가 오면 분류기가 바뀔 가능성이 커지기 때문입니다. 결과적으로, m개 만을 사용하여 tree 구축에 사용하여 다양한 모양의 tree를 생성하여 variance를 줄여야 효과를 볼 수 있습니다.

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

F

SVM은 linear separable, linear non-separable, not linear separable에 적용 가능합니다.

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

F

kernel function은 hyper parameter로써 선택하는 것은 예측성능에 영향을 끼칩니다.

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

F

SVM의 main idea는 margin을 최소화하는 것이 아닌 최대화하는 것입니다.