CIS 4526 Foundations of Machine Learning

Decision Trees

Khai Nguyen

CIS 4526 Foundations of Machine Learning

Part 1

1. **The name of the data set.**
   Unpacking the push-pull system: Assessing the contribution of companion crops along a gradient of landscape complexity

   Filename: Kebede et al 2018b.xlsx

2. **Where the data can be obtained.**
   Harvard Dataverse

3. **A brief (i.e. 1-2 sentences) description of the data set including what the features are and what is being predicted.**
   There are 3 sites that this dataset looks at “Wondo Genet”, “Tula”, and “Hawassa Zuria”, specific keys are provided in the key tabs “Key1”, “Key2”, “Key3”.
   The dataset is used to predict how different “Cropping systems” and “natural enemies” affects landscape complexity.

4. **The number of examples in the data set.**
   Limited, 140

5. **The number of features for each example. If this isn’t concrete, describe it as best as possible.**
   If we look at the “Natural_enemies” tab, there are 12 features for each sample. We have “FarmID”, “DistrictName”, “LandscapeComplexity”, “SamplingPeriod”, “Napier”, “CroppingSystem”, “Formicidae”, “Araneae”, “Coccinellidae”, “Staphylinidae”, “TotalPredators”, “EggPredation”.

   - FarmID: Farm code
   - DistrictName: Name of the district
   - LandscapeComplexity: Gradient of landscape complexity
   - SamplingPeriod: Maize devlopment stage
   - Napier: Presence or absence of napier grass
   - CroppingSystem: Cropping system (M=Sole Maize, MB= Maize-Bean, MD=Maize Desmodium)
   - Formicidae: Number of Formicidae
   - Araneae: Number of Araneae
   - Coccinellidae: Number of Coccinellidae
   - Staphylinidae: Number of Staphylinidae
   - TotalPredators: Number of stemborers predators
   - EggPredation: Area of eggs preadated

Part 2. (in pa2.py)

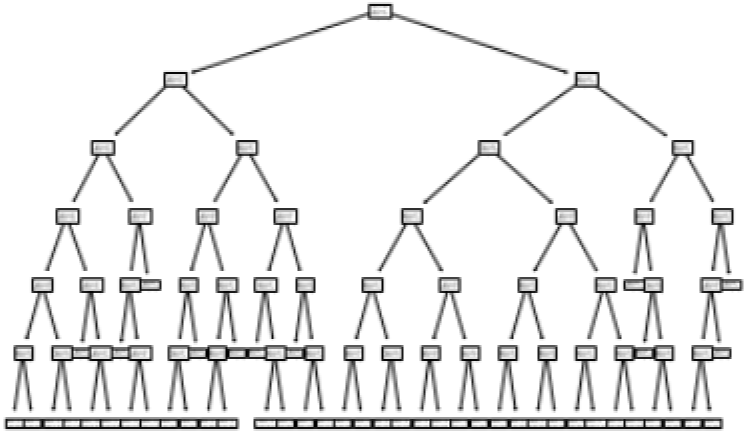
i) titanic_train.csv

Q1)

```
Feature 0 has error: 0.32492997198879553
Feature 1 has error: 0.21988795518207283
Feature 2 has error: 0.4061624649859944
Feature 3 has error: 0.4061624649859944
Feature 4 has error: 0.3851540616246499
Feature 5 has error: 0.38375350140056025
```

Q2)

```
Error for full decision tree: 0.26573426573426573
```



ii) breast_cancer.csv

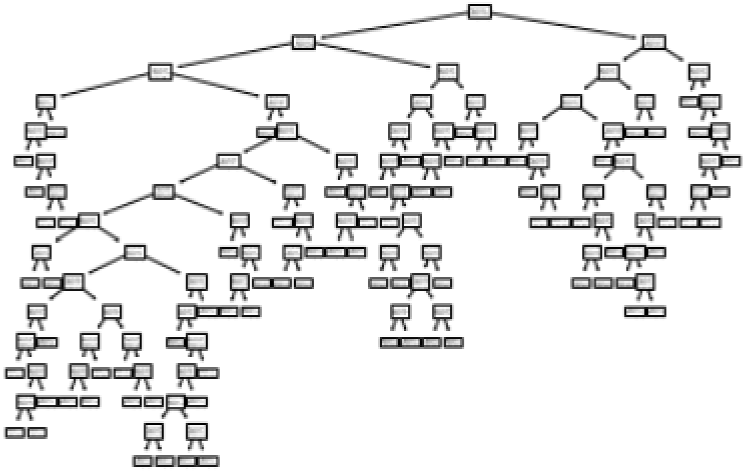
Q1)

```
Feature 0 has error: 0.2972027972027972
Feature 1 has error: 0.2972027972027972
Feature 2 has error: 0.2972027972027972
Feature 3 has error: 0.2727272727272727
Feature 4 has error: 0.2762237762237762
Feature 5 has error: 0.27972027972027974
Feature 6 has error: 0.2972027972027972
Feature 7 has error: 0.2937062937062937
Feature 8 has error: 0.2972027972027972
```

CIS 4526 Foundations of Machine Learning

Q2)

```
Error for full decision tree: 0.3275862068965517
```



Part 3.

We **cannot** use k-NN to train a classifier on these two datasets. For k-NN, we cannot compute the Euclidean distance between datapoints in the “titanic” and “breast_cancer” datasets. The same issue happens to PLA, with the datapoints from the 2 provided datasets, we cannot compute the dot product (w^T.x) in the sign() function.

Therefore, both k-NN and PLA are not good training models for these 2 datasets.