CMU 10-715: Homework 8

Decision Trees

DUE: Nov. 16, 2020, 11:59 PM.

Instructions:

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books, again after you have thought about the problems on your own. Please don't search for answers on the web, previous years' homeworks, etc. (please ask the TAs if you are not sure if you can use a particular reference). There are two requirements: first, cite your collaborators fully and completely (e.g., "Alice explained to me what is asked in Question 4.3"). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only.
- **Submitting your work:** Assignments should be submitted as PDFs using Gradescope unless explicitly stated otherwise. Each derivation/proof should be completed on a separate page. Submissions can be handwritten, but should be labeled and clearly legible. Else, submission can be written in LaTeX.
- **Late days:** For each homework you get three late days to be used only when anything urgent comes up. No points will be deducted for using these late days. We will consider an honor system where we will rely on you to use the late days appropriately.

1 Decision Trees

In this homework you will implement a decision tree and train it to detect recurrence of breast cancer on the UCI breast cancer dataset. In particular, you need to implement a variant of the ID3 algorithm learnt in class.

1.1 Modified ID3 algorithm

The ID3 algorithm learnt in class only allows for binary target and binary features. The modified version presented can use multivariate features.

Algorithm 1: Modified ID3 algorithm

input: training set S, features A, max depth k, node depth d, `Gain` measure
Create a Root node for the tree
if *all examples in S are labeled 1* **then**
  return a leaf 1
else if *all examples in S are labeled 0* **then**
  return a leaf 0
else if $k = d$ **then**
  return a leaf whose value is majority of labels in S
else
  $j \leftarrow argmax_{i \in A}$`Gain`$(S, i)$
end
Create a subtree with root node having attribute j
for *each unique value* v_i *of feature* j **do**
  Add a new branch corresponding to $j = v_i$
  if $|S_{v_i}| = 0$ **then**
    Add a leaf whose value is majority of labels in S below the branch
  else
    Add the subtree returned by $ID3(S_{v_i}, A \setminus \{j\}, k, d+1,$ `Gain`$)$
  end
end
Return Root

1.2 Instructions

- Download the data from `https://github.com/ShenghaoWu/10715/tree/master/hw8` and read the dataset description in breast-cancer.names.txt. *Hint: you may want to create a dictionary will all the different values for each feature in order to create all the possible branches for each variable.*
- Implement the modified ID3 algorithm stated above. For this hw we do not provide you with a base code.
- Use the Gini Index as Gain measure

- The target variable is the presence of recurrence-events (1: recurrence-events, 0: no-recurrence-events)
- **Include all your code at the end of the pdf.**

1.3 Results

a (50 points) Perform k-fold cross-validation with $k = 3$. Try with max depth 1, 3, 5, 7 and 9. Plot the training and validation set accuracy for each value of max depth.

b (30 points) Train the model again with ALL the training samples on the best max depth based on the previous question. Report train and test accuracy.

c (20 points) What can you say about the relation between train accuracy and max depth? What about validation accuracy and max depth?