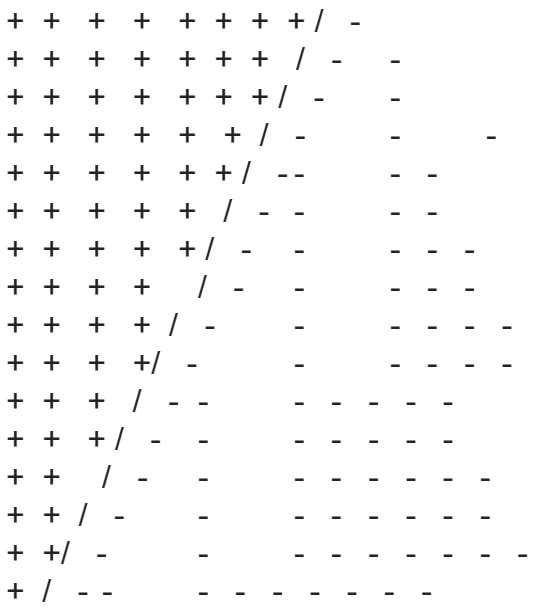
Session 12 Lab Abstract

Here is what we have planned for Session 12 - Lab. There are 3 experiments. Those are:

Experiment 0 (<15 minutes)
A Math problem for evaluating your understanding of the Entropy and Gini Index (might be handy to bring a scientific calculator or use an online scientific calculator) (Released during lab)

Experiment 1 (<75 minutes)
Part 1:
* Take a small (synthetic) dataset of 2 classes, well separable by a 45-degree diagonal line passing through origin (X=Y),
  but not separable by a horizontal or vertical line (similar to the plot below).

* Plot the data and see its nature



Part 2:
* Train a linear classifier (perceptron)
* Plot the decision boundary (one should see a diagonal line as decision boundary).

Part 3:
* Train a decision tree classifier on it (set the depth so that the leaf nodes are pure).
* Plot the decision boundary and see (one should see a stair-case like boundary).

Part 4: Task 1 for the students:

* Find the effect for varying the depth of the tree on the decision boundary by plotting it for various values of depth.

Part 5: Task 2 for students:
* Add a new feature (Y-X) that makes the samples separable by a horizontal plane. How does the decision tree change?

* Question 1: What is the nature of decision boundary? Question 2: How does this affect regression of continuous values?

Experiment 2 (>180 minutes):
End to end accident severity prediction problem. (A classification problem)

Step 0) Kindly make sure you make a Kaggle account. There is a template presentation under the “Data” tab of Kaggle. Please download it. Please spend 30 minutes with your group to strategize your plan of approach. Please try to create your own such presentation (in .pdf format only)

Step 1) Please add your team-mates to your Kaggle teams only

Step 2) Please read all the entire write up in Kaggle thoroughly and try to submit a sample submission (just to familiarise with the system)

Step 3) The Attributes description, head of the train set is available (first 5 records), and the test set is available right now.

The rest of the data set would be released at 4:00 pm IST April 21/2018. Build your models in the 3 sequences provided. There is a deadline for the Kaggle leaderboard to be frozen.

Link to Kaggle: https://www.kaggle.com/t/416d85aa866c4a038aa6e5e0029157b3

The leaderboard takes and reflects your best submission until the specified deadline (maximum of 20 submissions/day), however you can submit post the deadline to see where you would have featured. At this time the Kaggle end to end Accident Severity is not going to be evaluated.

Note: This data is proprietary please do not share the dataset with anyone. The Solution Test file and the solution python notebook will not be provided.

All the best!

Thanks, Jay