HW #3 Due: 4/24/2020

1. We mentioned the "play/no play" example in the lecture notes. Compute all gains for the subtree in the "rainy" branch by hands and draw the corresponding subtree based on ID 3 criteria.
2. We have $G = 0.048$ for $\mathrm{H}_0 = 65$ and $G = 0.102$ for $\mathrm{H}_0 = 80$ on pp. 45 of the PPT file. Perform hand calculation to confirm that these G values are correct.
3. Read the reference: http://www.stats.ox.ac.uk/~flaxman/HT17_lecture13.pdf, and then use the algorithm on pp. 31 to find the optimal split on the root node in a CART tree with the following dataset: $\{(x, y)|y = 2x, x = 1,2,3,4\}$. In this problem, x is an attribute and y is a dependent variable.
4. In this problem, you are asked to use the cancer dataset "Breast Cancer Wisconsin (Original) Data Set" to perform PCA (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29). The full features are from attribute 2 to attribute 10, and the classes are benign and malignant. There are some missing attributes, so you need to use the in-class average to replace the missing attributes. Use 70% of the dataset as the training set to (a) plot PoV(k) for k from 1 to 9 for the training set, and (b) to find the value of k_0 such that $\mathrm{PoV}(k_0) > 0.9$. Remove mean values before computing eigenvalues.
5. Perform 5-NN classification for the dataset given in problem 4 with (a) original 9 attributes, and (b) k_0 attributes obtained after PCA transformation. Repeat the experiments 10 times (10 trials) and report the average accuracy for both types of attributes. As the contents of the training set vary, the PCA and k_0 must be re-calculated for each trial. Do you observe large accuracy difference in both accuracy results?