Decision Trees

Useful formulas

- Entropy of a set S with elements from C classes: $H(S)=\sum_{i=1}^{C} -p_i \log p_i$.
- Gini impurity of a set S with elements from C classes: $G(S)=1-\sum_{i=1}^{C} p_i^2$
- Gini/Information gain: $G(S,F)=M(S)-\sum_{f \in values(F)} \frac{|S_f|}{|S|} M(S_f)$, where M is either the Gini impurity or the Entropy.

Questions

1. What is the entropy of a dataset? How do you compute it?

   The entropy is a measure of how much surprise I can expect, if I extracted an element from a set. If the set contains all elements of the same type, my surprise will be minimal. If the set contains an equal number of elements of each type I don't know what to expect, and my surprise will be maximal.

   Technically, it is the expected value of the information of a message that can take n values each one with probability p_i : $H=E[I]=\sum_i -p_i \log_2 p_i$.

2. How does the algorithm ID3 decide what the next feature to split on is?

   It computes the information gain of each feature, and splits with respect to the feature with highest information gain.

3. What does ID3 do when there are no more features left to split on?

   Assigns to the leaf of the tree the class of the majority of points that have the feature values corresponding to that leaf.

4. What type of data can decision trees classify which MLPs cannot?

   Non-metric data, that is, data points whose features are not numbers.

5. What is a random forest? What are the sources of randomness that diversify the trees in the forest?

   A random forest is a collection of decision trees created from the same dataset. Randomness is obtained by using only a random fraction of the features, or a random fraction of the data.

6. What is the main difference between the CART and the ID3 algorithms?

   CART uses the Gini impurity, while ID3 uses information gain.

7. Consider the following dataset, where data have two features, each of which has three values (A, B, or C), and the last element is the class: <A, B,0>, <A, C,1>, <A, B,0>, <B, B,0>, <B, B, 0>, <B, C,1>, <C, A,1>, <C, B,1>, <C, B,1>, <C, C,0>. Construct a decision tree on the dataset with ID3.

   The information gain for a set S and feature F is: $G(S,F)=\mathrm{H}(S)-\sum_{f\in values(F)}\frac{|S_f|}{|S|}\mathrm{H}(S_f)$

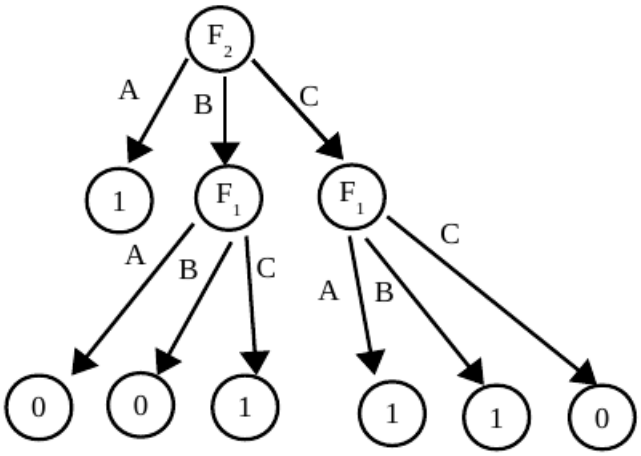
   So, for Feature 1:

   $$G(S,F_1)=\mathrm{H}(S)-\sum_{f\in\{A,B,C\}}\frac{|S_f|}{|S|}\mathrm{H}(S_f)=$$

   $$=1-2\cdot\frac{3}{10}(-\frac{2}{3}\log_2(\frac{2}{3})-\frac{1}{3}\log_2(\frac{1}{3}))-\frac{4}{10}(-\frac{3}{4}\log_2(\frac{3}{4})-\frac{1}{4}\log_2(\frac{1}{4}))=0.125$$

   while, for Feature 2:

   $$G(S,F_2)=1-\frac{1}{10}\cdot 0-\frac{6}{10}(-\frac{4}{6}\log_2(\frac{4}{6})-\frac{2}{6}\log_2(\frac{2}{6}))-\frac{3}{10}(-\frac{2}{3}\log_2(\frac{2}{3})-\frac{1}{3}\log_2(\frac{1}{3}))=0.174$$



   Since the second feature has a larger information gain, ID3 splits with respect to it. The full tree is as follows:

8. We want to learn a classifier for car diagnosis. The classes are: OK (O); go to a garage (G); severe failure, don't drive (F). The features are: makes a strange noise (N) or not (nN); emits black smoke (S) or not (nS); going straight, the car drifts on a side (D), or doesn't (nD). We ask a mechanic, and build the following (very extensive) dataset: <N, nS,nD,G>, <nN,nS,nD,O>, <nN,S,nD,F>, <nN,S,D,F>, <N,S,nD,F>, <nN,nS,D,G>, <N,nS,D,G>, <N,nS,nD,G>. Construct a decision tree on the dataset with ID3. My car makes a strange noise, what should I do?

   The mechanic classifies the fault based on three binary features. If a client calls on the phone, which question should the mechanic ask first? We choose the first feature to split on by computing the information gain, as before (note that in this case we have three classes!):

$$H(S)=-\frac{1}{8}\log_2(\frac{1}{8})-\frac{4}{8}\log_2(\frac{4}{8})-\frac{3}{8}\log_2(\frac{3}{8})=1.406$$

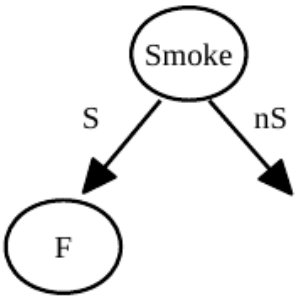
$$G(S,\text{Noise})=1.406-\frac{4}{8}\overbrace{(\underbrace{-\frac{3}{4}\log_2(\frac{3}{4})}_{\text{garage}}\underbrace{-\frac{1}{4}\log_2(\frac{1}{4})}_{\text{dont' drive (F)}})}^{\text{noise}}-\frac{4}{8}\overbrace{(\underbrace{-\frac{1}{4}\log_2(\frac{1}{4})}_{\text{OK}}\underbrace{-\frac{2}{4}\log_2(\frac{2}{4})}_{\text{don't drive}}\underbrace{-\frac{1}{4}\log_2(\frac{1}{4})}_{\text{garage}})}^{\text{no noise}}=0.25$$

$$G(S,\text{Smoke})=1.406-\frac{3}{8}\cdot 0-\frac{5}{8}(-\frac{4}{5}\log_2(\frac{4}{5})-\frac{1}{5}\log_2(\frac{1}{5}))=0.955$$

$$G(S,\text{Drifting})=1.406-\frac{3}{8}(-\frac{1}{3}\log_2(\frac{1}{3})-\frac{2}{3}\log_2(\frac{2}{3}))-\frac{5}{8}(-2\cdot\frac{2}{5}\log_2(\frac{2}{5})-\frac{1}{5}\log_2(\frac{1}{5}))=0.11$$

The most informative feature is Smoke, and therefore the first question the mechanic should ask is: is the car emitting smoke?
The first split is with respect to Smoke, and all cars that smoke must stop, so this is what we know for now:

Now we need to focus of all the points that have “nS” as a value of the Smoke feature. This



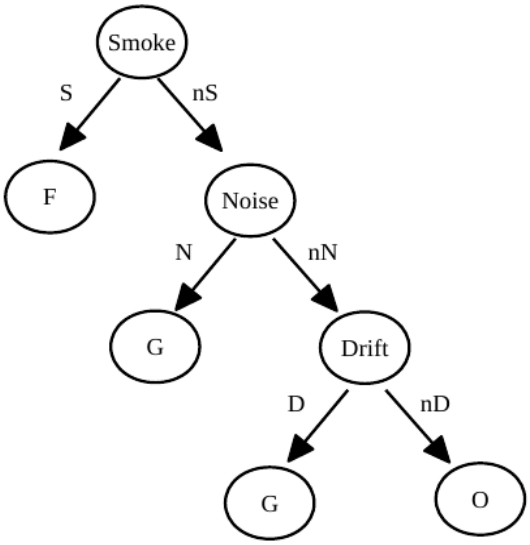
is the set S_{nS} ={<N, nS,nD,G>, <nN,nS,nD,O>, <nN,nS,D,G>, <N,nS,D,G>, <N,nS,nD,G>}. Its entropy is:

$$H(S_{nS})=-\frac{4}{5}\log_2(\frac{4}{5})-\frac{1}{5}\log_2(\frac{1}{5})=0.722 \quad .$$

$$G(S_{nS},\text{Noise})=0.722-\frac{3}{5}\cdot 0-\frac{2}{5}\cdot 1=0.322 \quad .$$

$$G(S_{nS},\text{Drifting})=0.722-\frac{2}{5}\cdot 0-\frac{3}{5}(-\frac{2}{3}\log_2(\frac{2}{3})-\frac{1}{3}\log_2(\frac{1}{3}))=0.171 \quad .$$

The most informative of the two features left is Noise, and this is the resulting tree:



9. Same as the last two questions, but with CART.

CART uses the Gini Impurity. The previous calculations maintain the same structure, but we need to substitute entropy with Gini impurity, and see what happens. Starting with question 7:

$$G(S,F_1)= \overbrace{\left(1-\frac{1}{4}-\frac{1}{4}\right)}^{\text{gini impurity of the whole set}} -2\cdot\frac{3}{10}\left(1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2\right)-\frac{4}{10}\left(1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right)=0.083$$

$$G(S,F_2)=0.5-\frac{1}{10}\cdot 0-\frac{6}{10}\left(1-\left(\frac{4}{6}\right)^2-\left(\frac{2}{6}\right)^2\right)-\frac{3}{10}\left(1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2\right)=0.1$$

In this case as well, feature 2 is the better one, since the average impurity after the split is lower. The resulting tree is the same as for ID3.

Let's now look at Question 8:

$$G(S,\text{Noise})=0.594-\overbrace{\frac{4}{8}\left(1-\underbrace{\left(\frac{3}{4}\right)^2}_{\text{garage}}-\underbrace{\left(\frac{1}{4}\right)^2}_{\text{dont' drive (F)}}\right)}^{\text{noise}}-\overbrace{\frac{4}{8}\left(1-\underbrace{\left(\frac{1}{4}\right)^2}_{\text{OK}}-\underbrace{\left(\frac{2}{4}\right)^2}_{\text{don't drive}}-\underbrace{\left(\frac{1}{4}\right)^2}_{\text{garage}}\right)}^{\text{no noise}}=0.094$$

$$G(S,\text{Smoke})=0.594-\frac{3}{8}\cdot 0-\frac{5}{8}\left(1-\left(\frac{4}{5}\right)^2-\left(\frac{1}{5}\right)^2\right)=0.394$$

$$G(S,\text{Drifting})=0.594-\frac{3}{8}\left(1-\left(\frac{1}{3}\right)^2-\left(\frac{2}{3}\right)^2\right)-\frac{5}{8}\left(1-2\cdot\left(\frac{2}{5}\right)^2-\left(\frac{1}{5}\right)^2\right)=0.027$$

again, Smoke is the first feature. Let's see if the second one is also the same…

$$G(S_{nS},\text{Noise})=0.32-\frac{3}{5}\cdot 0-\frac{2}{5}\cdot 0.5=0.12$$

$$G(S_{nS},\text{Drifting})=0.32-\frac{2}{5}\cdot 0-\frac{3}{5}\left(1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2\right)=0.053$$

yes. This tree is also the same as the previous one. This happens quite often, entropy and Gini impurity tend to give the same splits.