Set Diversity (Gini Impurity)

In this question we are going to asks a simple question: what is the probability that two chosen objects from a set are different. This statistic, formally called the Gini Impurity is used both in Random Forest algorithms and in social science. This was a problem from the Stanford Midterm in Fall 2023.

a) Consider the following set of shapes. If you chose two shapes **with replacement** what is the probability that the two shapes are **the same**? Note that it is possible to get two triangles: after you pick the first triangle, you put it back into the set of shapes and it can be chosen again.

Define the Sample Space to be the 49 different outcomes of chosing the two shapes with replacement (7 choices for the first shape and 7 choices for the second shape). Notice that in our construction of the sample space we have chosen for the shapes to be treated as all being distinct from one another. An outcome in S ($\in S$) is thus an ordered tuple of distinct shapes. For example one outcome in S is (Shape_4, Shape_2). $|S| = 7 \times 7 = 49$. Note that all the outcomes in S are equally likely (that was the reason why we treated shapes as distinct).

Let E be the subset of S where the two shapes match. Let A be the event that you have two squares and let B be the event that you have two trianges. Note that A and B are mutually exclusive. By the step rule of counting $|A| = 6 \cdot 6 = 36$ since there are two steps to creating an outcome in A: chose a square then chose another square. Similarly $|B| = 1 \cdot 1 = 1$.

$$\begin{aligned} P(\text{same}) &= \frac{|E|}{|S|} \\ &= \frac{|A| + |B|}{|S|} \\ &= \frac{36 + 1}{49} \\ &= \frac{37}{49} \end{aligned}$$

b) Consider the following set of shapes. If you chose two shapes **with replacement** what is the probability that the two shapes are **different**? Notice that the previous question asked for the probability that the two shapes are the same. The probability that two items are different is called the Gini Impurity of a set.

Define the same Sample Space to be the 49 different outcomes of chosing the two shapes with replacement (7 choices for the first shape and 7 choices for the second shape). $|S| = 7 \times 7 = 49$. Note that all the outcomes in S are equally likely.

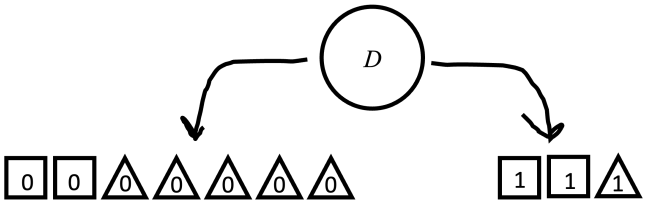
Let E be the subset of S where the two shapes match.
Let A be the event that you have two squares.
Let B be the event that you have two trianges.
Let C be the event that you have two stars. Note that A, B and C are mutually exclusive. By the step rule of counting
$|A| = 4 \cdot 4 = 16$
$|B| = 2 \cdot 2 = 4$
$|C| = 1 \cdot 1 = 1$

$$\begin{aligned} P(\text{different}) &= 1 - P(\text{same}) \\ &= 1 - \frac{|E|}{|S|} \\ &= 1 - \frac{|A| + |B| + |C|}{|S|} \\ &= 1 - \frac{16 + 4 + 1}{49} \\ &= 1 - \frac{21}{49} = \frac{4}{7} \end{aligned}$$

Gini Imputiy in Decision Trees

Note: This next problem isn't remarkably different to the problems above. The point of this problem is to show you how the concept of Gini Impurity connects to Decision Trees.

Decision Trees (and their big brother Random Forests) are some of the most popular classification AI algorithms that don't use deep nerual networks. They are built based off data, node-by-node. Once built, they can be used to make classification decisions. The critical decision when making a decision tree is to decide which node to add next. One way to make that decision is to chose the node which choice of new node leads to the largest decrease in Gini Impurity of the shapes that end up together at the end of the decision tree.



Aside: this particular node has split the shapes based off a value (which is slightly related to the shape). Shapes with value 0 go to the left child, shapes with value 1 go to the right.

c) Consider the node in the picture above. Let G_L be the probability that two shapes, selected from the shapes in the left side, are different (Gini Impurity) and let G_R be the probability that two shapes, selected from the shapes in the right side, are different (again the Gini Impurity). What is the value of $\max(G_L, G_R)$? We only use the shape type to calculate the Gini Impurity. The value represents the features that were used to sort them between the left and right node and can be ignored in this question.

First lets generalize our calculation of G for a set of shapes. Let n be the number of shapes in the set. Let n_i be the number of shapes of type i in the set. It is going to be easier to calculate the probability that two share ares are the same than the probability that two shapes are different. If we set up the sample space S to be the set of all ways to pick two shapes (treating each shape as distinct, and treating the selection as ordered). The event space E is then the subset of events where the shapes are the same. Because chosing the same shape of one type is mutually exclusive with chosing the same shape of another type, we can calculate the probability that two shapes are the same by calculating the probability that two shapes are the same for each type of shape and summing them up:

$$\begin{aligned} G &= 1 - P(\text{same}) \\ &= 1 - \frac{|E|}{|S|} \\ &= 1 - \frac{\sum_i^n n_i^2}{|S|} \end{aligned}$$

$G_L = 1 - \frac{2^2+5^2}{7^2}$ and $G_R = 1 - \frac{2^2+1^2}{3^2}$. For those curious G_R is the larger of the two.

Going Further

Note: in the midterm students were asked to calculate the "Expected" Gini Score, instead of the max. That uses Expecataion, a concept we will learn in the next section. We also asked for the Gini Impurity of a Poisson random variable, a concept from Part 3.