Pruning Trees

Purity, Gini Index and Cross-Entropy

- **Purity** is the extent to which observations **within node/regions** contain data predominantly **from 1 class** k, as opposed to being widely scattered across classes.
- **Purity** is a useful measure to evaluate **tree-pruning** by evaluating how **"pure"** a **node** is → 2 popular measures:

$$Gini\ Index = G = \sum_{k=1}^{K} p_{mk}(1 - p_{mk})$$

$$Cross - Entropy\ = D = -\sum_{k=1}^{K} p_{mk} * Log(p_{mk})$$

- p_{mk} is the proportion of observations in the m^{th} tree **region** in the k^{th} (most common) class (in that region).
- G and D are very similar quantitatively and both take a **small** value when p_{mk} is close to 0 or 1 → **more pure** and a good indicator of the **quality** of a **split**

R Tips

`prune.regtree.fit=prune.tree(regtree.fit,best=5)` → Prunes the `regtree.fit` regression tree object to `5` terminal nodes

```
prune.classtree.fit=
    prune.misclass(classtree.fit,best=9)
```
→ Prunes the `classtree.fit` object to `9` terminal nodes

KOGOD SCHOOL
of
BUSINESS