**随机森林(决策树)超易懂原理**

**决策树的每一个决策节点向下分类的过程中，所依据的分类原则是找到变量的阈值，使该分类层级的Gini impurity（分类得到的子集中任选一个点，这个点是错误分类的概率）最小：**

**蓝色代表牛瘟、红色代表口蹄疫、绿色代表呼吸道疾病**

**有个数据框：**

|  |  |  |
| --- | --- | --- |
| **x** | **y** | **color** |
| **0** | **1** | **blue** |
| **0.5** | **0.5** | **blue** |
| **1.1** | **1.5** | **blue** |
| **1.8** | **2.1** | **red** |
| **1.9** | **2.8** | **red** |
| **2** | **2** | **green** |
| **2.5** | **2.2** | **green** |
| **3** | **3** | **green** |
| **3.6** | **3.3** | **green** |
| **3.7** | **3.5** | **green** |

**第一个决策节点为整个数据集，假设分类变量的阈值为y=2，则整个数据集可被分为：**

|  |  |  |
| --- | --- | --- |
| **0** | **1** | **blue** |
| **0.5** | **0.5** | **blue** |
| **1.1** | **1.5** | **blue** |

**和**

|  |  |  |
| --- | --- | --- |
| **1.8** | **2.1** | **red** |
| **1.9** | **2.8** | **red** |
| **2** | **2** | **green** |
| **2.5** | **2.2** | **green** |
| **3** | **3** | **green** |
| **3.6** | **3.3** | **green** |
| **3.7** | **3.5** | **green** |

**右边数据框有2红5绿，任取一点判为错误的概率(即Gini impurity)为：**

**P抽到红\*P判为绿 + P抽到绿\*P判为红 = 2/7\*5/7 + 5/7\*2/7 = 20/49**

**左边数据框全部为3蓝，任取一点判为错误的概率(即Gini impurity)为：**

**0**

**进行加权：Gy-root = 0.7\*20/49 + 0.3\*0 = 0.286**

**假设分类变量的阈值为x=2，则整个数据集可被分为：**

|  |  |  |
| --- | --- | --- |
| **0** | **1** | **blue** |
| **0.5** | **0.5** | **blue** |
| **1.1** | **1.5** | **blue** |
| **1.8** | **2.1** | **red** |
| **1.9** | **2.8** | **red** |

**和**

|  |  |  |
| --- | --- | --- |
| **2** | **2** | **green** |
| **2.5** | **2.2** | **green** |
| **3** | **3** | **green** |
| **3.6** | **3.3** | **green** |
| **3.7** | **3.5** | **green** |

**右边数据框全部为5绿，任取一点判为错误的概率(即Gini impurity)为：0**

**左边数据框为3蓝2红，任取一点判为错误的概率(即Gini impurity)为：**

**P抽到蓝\*P判为红 + P抽到红\*P判为蓝 = 3/5\*2/5 + 2/5\*3/5 = 12/26**

**进行加权：Gx-root = 0.5\*12/26 + 0.5\*0 = 0.231**

**以x=2为分类变量的阈值可以获得更小的Gini impurity，故舍弃y=2这一分类变量**

**https://mp.weixin.qq.com/s/lBRvQJwix2H8Up29vSI0Ig**