Unyleya
EDUCACIONAL

Ganho de informação

Ganho de informação

As árvores são criadas com o propósito de ganhar informação, ou seja, à medida que novos nós vão sendo acrescentados o que se espera é que a árvore saiba mais sobre como tratar o problema que estiver em questão.

Para verificar se a árvore está crescendo no sentido correto, ou seja, se ela está aprendendo, foi criado o conceito de ganho de informação.

O ganho de informação é calculado através de algumas métricas, sendo as principais:

- Entropia;
- GINI.

Fórmula do ganho de informação

$$InfoGain(R, R_e, R_d) = H(R) - (|R_e|*H(R_e) + |R_d|*H(R_d)) / |R|$$

A fórmula de ganho de informação é composta pelos termos identificados como:

- **H**: impureza da região;
- **R** : região atual;
- **R_e** : sub-região à esquerda;
- **R_d** : sub-região à direita;
- **|R|** : quantidade de exemplos da região.

Entropia

$$entropia(R) = -\sum p(c|R) \log (p(c|R))$$

No cálculo da entropia temos as seguintes variáveis:

- **c** : quantidade de ocorrências de uma classe;
- **R** : quantidade de ocorrências dentro da região.

Calculando a entropia

Para uma base de dados com 150 registros, 3 classes, cada classe com 50 registros.

p(c|R) = 50 / 150

p(c|R) ~ 0.33.

Considerando uma separação total entre a classe em questão àquela que está à sua esquerda:

$$entropia(R) = -3 * (0.33\ log\ log\ (0.33)\ ) \sim 0.48 entropia(R_e)$$

$$= -(1.0\ log\ log\ (1.0) + 0.0\ log\ log\ (0.0) + 0.0\ log\ (0.0)) = 0$$

$$entropia(R_d) = -(0.0\ log\ log\ (0.0\ ) + 0.5\ log\ log\ (0.5) + 0.5\ log\ log\ (0.5)\ ) \sim 0.30$$

Ganho pela entropia

$$ganhoInformacao = 0.48 - \frac{50 * 0 + 100 * 0.30}{150} = 0.28)$$

Função para entropia em Python

```
def entropyCriterion(data, labels):
classes = np.unique(labels)
    s = 0
    for c in classes:
        p = np.mean(labels == c)
        s -= p * np.log(p)
    return s


def stoppingCriterion(nClasses, depth, maxDepth):
return (maxDepth is not None and maxDepth == depth) or (nClasses == 1)
```

GINI

$$\text{gini}(R) = \sum p(c|R)\,(1 - p(c|R)),$$

gini(R) = ∑ p(c|R) (1 - p(c|R)) = 3 * (0.33 *(1 - 0.33)) ~0.66

gini(R_e) =(1.0 * (1.0 - 1.0) + 0.0 * (1 - 0.0) + 0.0 * (1 - 0.0)) = 0

gini(R_d) =(0.0 * (1 - 0.0) + 0.5 * (1 - 0.5) + 0.5 * (1 - 0.5)) ~0.5

Calculando ganho

$$0.66 - (50*0 + 100*0.50) / 150 = 0.16$$

Obrigada!

Ana Laurentino