پاسخ تکلیف Decision Tree (ID3)

درس یادگیری ماشین

امیرحسین ابوالحسنی
۴۰۰۴۰۵۰۰۳

1 Suppose there is an attribute, "A," that consists of random values, and these values do not have any correlation with the class labels. Additionally, assume that "A" has a sufficient number of distinct values such that no two instances in the training dataset share the same value for "A." What would be the outcome if a decision tree is built using this attribute? What challenges or issues might arise in this scenario?

پاسخ

قسمت اول

با توجه به الگوریتم ID3، در ابتدا information gain ناشی از هر ویژگی را سنجیده و آن ویژگی که بیشترین gain را دارد انتخاب می‌کنیم (تعداد کلاس‌های هدف = k):

$$Gain(S, A) = Entropy(S) - \sum_{v \in A} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy(S_v) = -\sum_{v} p_v \log(p_v) = P(S_v = 0) \log P(S_v = 0) + P(S_v = 1) \log P(S_v = 1) + \ldots + P(S_v = k) \log p(S_v = k)$$

به علت یکتایی این ویژگی(کلید اصلی بودن) برای هر نمونه، همه ترم‌های $P(S_v = l) \log P(S_v = l)$ برابر با صفر می‌شود. زیرا

$$P(S_v = l) = 0$$

یا

$$P(S_v = l) = 1$$

در نتیجه یکی از مضرب ها ۰ خواهد شد و کل ترم را ۰ خواهد کرد. بدین صورت است که نتیجه می‌گیریم:

$$Entropy(S_v) = 0$$

و این ویژگی برای ریشه انتخاب می‌گردد:

$$\arg\max \{Gain(S_v) | \forall A \in \text{Header}\} = A$$

در نتیجه این کار، ارتفاع درخت ۱ شده و به تعداد مقادیر ویژگی A، شاخه خواهیم داشت.