Mesterséges intelligencia II.

Döntési fák

Legyenek adottak az $F_1, \ldots, F_n$ jellemzők. Az elmlített jellemző készlethez tartozó döntési fának nevezzük azokat a fákat, melyek:

- Minden belső csomópontjához pontosan egy F_i jellemző van rendelve.
- Minden belső csomópontnak pontosan annyi leszármazottja van, ahány különböző értéket felvehet az adott csomóponthoz rendelt jellemző (A jellemzők véges értékkészletűek).
- Minden levél csomóponthoz hozzárendeljük a + vagy − címkét.

Döntési fák tanulása ID3 algoritmussal. Legyen adott $S = \{(\bar{x}_1, y_1), \ldots, (\bar{x}_l, y_l)\} \subseteq F_1 \times \ldots \times F_n \times \{+, -\}$ tanítóhalmaz. Valamint a kezdeti jellemzők halmaza legyen: $F = \{F_1, \ldots, F_n\}$.

root ID3(S, F):
  root ← új csomópont
  if minden példa címkéje + az S halmazban **then**
    root.label ← +
    return *root*
  end if
  if minden példa címkéje - az S halmazban **then**
    root.label ← −
    return *root*
  end if
  if $F = \emptyset$ **then**
    root.label ← a leggyakoribb osztálycímke az S halmazban
    return *root*
  end if
  $F_{best} \leftarrow argmax_{F_i \in F} Gain(S, F_i)$
  root.attribute ← F_{best}
  for all v_i lehetséges értékére az F_{best} attributumnak **do**
    S_{v_i} ← S halmaz olyan részhalmaza, hogy minden elem esetén F_{best} értéke v_i
    if $S_{v_i} = \emptyset$ **then**
      node ← a *root* csomópont új, levél csomópontja, v_i élcmímkével
      node.label ← a leggyakoribb osztálycímke az S halmazban
    else
      node ← ID3(S_{v_i}, $F - \{F_{best}\}$)
      node legyen a *root* új gyerek csomóponja, v_i élcímkével
    end if
  end for
  return *root*

Az algoritmusban látott $Gain(S, F)$ függvény az entrópia változás:

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

ahol R_A az A attributum értékkészletét, S_v pedig S azon részhalmazát jelöli, melynek minden elemére az A attributum értéke éppen v, az Entropy(S) pedig a következő függvény:

$$Entropy(S) = -\sum_{i \in \{+,-\}} p_i log(p_i),$$

ahol p_+ a pozitív példák, p_- a negatív példák előfordulási valószínűsége S-ben.

Általános esetben az entrópia:

$$Entropy(S) = -\sum_{v_i \in R_S} p_{v_i} log(p_{v_i}),$$

ahol R_S jelöli az S halmaz véges értékkészletét, p_{v_i} pedig a v_i érték előfordulási valószínűségét.

Megjegyzés: A fent látott entrópia a Shannon entrópia. Az ID3 algoritmusban használható más típusú entrópia is.