Tik-79.230 **Kevät 2000**
Agenttipohjaisen tietojenkäsittelyn perusteet
Laskuharjoitus 5
Ratkaisut

1. Päätöspuu muodostetaan opetusdatasta siten, että jokaisella askeleella valitaan yksi valintamuuttujista ja lisätään se puuhun. Muuttuja pyritään valitsemaan siten, että se antaa mahdollisimman paljon tietoa lopputuloksesta. Tosin sanoen, valitaan muuttuja, joka pienentää esimerkkidatan entropiaa mahdollisimman paljon.

   Todennäköisyysjakauman informaatiosisältö (entropia) I saadaan laskemalla

   $$I(P(v_1), \ldots, P(v_n)) = \sum_{i=1}^{n} -P(v_i) \log_2 P(v_i)$$

   Tehtävän opetusdatan eläimistä x_1, x_2, x_6 ja x_7 ovat lintuja, ja x_3, x_4, x_5, x_8, x_9 eivät ole. Näin ollen entropia on aluksi

   $$I(\frac{4}{9}, \frac{5}{9}) = \frac{4}{9} \log_2 \frac{9}{4} + \frac{5}{9} \log_2 \frac{9}{5} = 0.991$$

   Seuraavaksi käydään läpi kaikki muuttujat, ja lasketaan mikä niistä antaa eniten informaatiota. Informaation lisäys saadaan ottamalla todennäköisyyksillä painotettu keskiarvo kaikkien solmusta lähtevien haarojen entropioista.

   Ensimmäinen muuttuja on "lentääkö?". Tarkastettaessa opetusvektoria huomataan, että linnuista x_1 ja x_2 lentävät, mutta x_6 ja x_7 eivät lennä. Muista eläimistä x_4 ja x_5 lentävät. Näinollen entropiaksi saadaan:

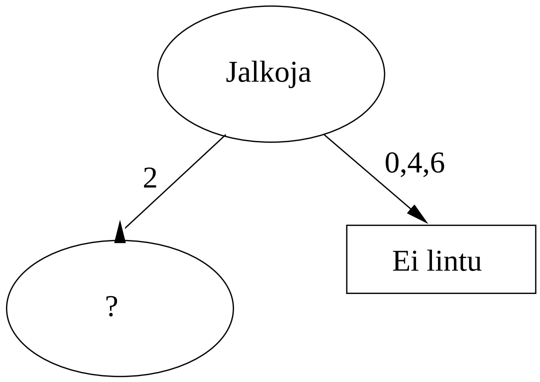
   $$I(\text{Lentääkö?}) = \frac{4}{9} I(\frac{2}{4}, \frac{2}{4}) + \frac{5}{9} I(\frac{2}{5}, \frac{3}{5}) = 0.984$$

   Toisin sanottuna, solmu "lentääkö?" ei auta juuri ollenkaan päätöksenteossa.

   Sitävastoin eläimen jalkojen lukumäärä on paljon parempi vaihtoehto, sillä ainoastaan kaksijalkaisten eläinten joukossa on lintuja:

   $$I(\text{Jalkoja}) = \frac{5}{9} I(\frac{4}{5}, \frac{1}{5}) + \frac{4}{9} I(0, 1) = 0.401$$

   Laskettaessa muiden solmujen entropiat huomataan, että "jalkoja" on paras vaihtoehto ensimmäiseksi solmuksi.



Seuraavaksi etsitään parasta päätömuuttujaa kysymerkillä varustettuun solmuun.

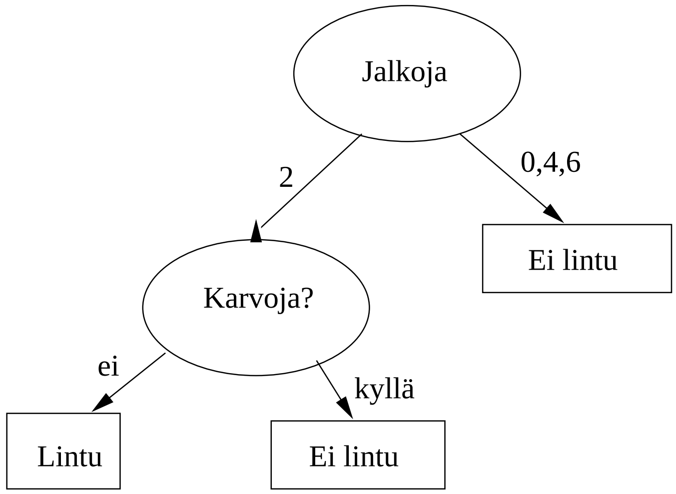
Muuttujalla "lentääkö¿' jäljellä olevat esimerkit jakaantuvat siten, että x_1, x_2 ja x_5 lentävät, x_6 ja x_7 eivät lennä. Tälle

$$I(\text{Lentääkö?}) = \frac{3}{5}I(\frac{2}{3}, 13) + \frac{2}{5}I(1,0) = 0.551$$

Tälläkään kertaa "lentääkö?" ei ole paras, sillä valitsemalla solmu "karvaa?" voidaan linnut erottaa muista eläimistä kokonaan, ja

$$I(Karvaa?) = \frac{1}{5}I(0,1) + \frac{4}{5}I(1,0) = 0$$

Näin olemme saaneet valmiiksi päätöspuun, jolla pystytään luokittelemaan oikein koko opetusvektori.



Opettamalla syntyneiden päätöspuiden ongelmana on se, että ne ovat ainoastaan niin hyviä kuin opetusdata on. Mikäli opetusdata ei ole tarpeeksi kattavaa, voi puu toimia aivan väärin todellisia päätöksiä tehtäessä. Mikäli dataa ja muuttujia on liikaa, voi opetusalgoritmi löytää sieltä yhteyksiä, jotka ovat oikeasti vain tilastollisia sattumia.

Ylläolevan puun ongelmana on liian vähäinen esimerkkidata. Koska ihmisellä on kaksi jalkaa, siirrytään "jalkoja"-solmusta solmuun "karvaa?".

Siellä huomataan, että ihmisellä ei ole karvoja, joten tunnistusalgoritmi luokittelee ihmisen linnuksi.