

Data-analytiikka – data-analyysimenetelmät

Arvioitu oppijan taitotaso: aloittelija tai keskitaso

Korrelaatioita

Korrelaatio mittaa kahden muuttujan välistä yhteyttä tai samankaltaisuutta. Korrelaatioita on kolmenlaisia: kahden jatkuvan muuttujan, kategorisen ja jatkuvan muuttujan välillä sekä kahden kategorisen muuttujan välillä. Tässä kategorinen muuttuja kuuluu luokkaan, kuten materiaaliin objektintunnistustehtävässä, ja jatkuva muuttuja on numero, jolla voidaan tehdä laskelmia, kuten lämpötila tai sademäärä.

Korrelaatiokertoimen lisäksi assosiaatioarvion laatua tai luotettavuutta voidaan arvioida erilaisilla mittareilla, joista yleisin lienee p-arvo, joka vaihtelee välillä 0-1. Pienemmät arvot viittaavat luotettavimpiin arvioihin, ja p-arvoa, joka on alle 0,05, pidetään tilastollisesti merkitseväenä, mikä tarkoittaa, että tutkimus voidaan todennäköisesti toistaa samantyyppisillä tuloksilla.

Tilastolliset menetelmät

Tilastollisia perusmenetelmiä ovat keskiarvo ja varianssi. Kehittyneempiä tekniikoita ovat entropia ja korkeamman asteen momentit, kuten vinous ja kurtoosi.

Raja-arvot

Raja-arvoanalyysissä etsitään kriittinen arvo, jossa mallin käyttäytyminen muuttuu tai josta voidaan poimia kiinnostavaa tai relevanttia tietoa.

Yksinkertainen esimerkki raja-arvoanalyysistä konenäkösovelluksissa on harmaasävykuvan binäärinen kynnys: pikselit, joiden arvot ylittävät kynnystason, katsotaan etualaksi, ja loput kuvasta hylätään taustana. Jäljelle jääneet (tai poimitut) pikselit – etuala – voidaan sitten analysoida edelleen muilla menetelmillä.

Suodatus ja lajittelu

Suodatus poistaa tietoja joidenkin ehtojen mukaan. Esimerkiksi voidaan etsiä vain tietty määrä suurimpia kohteita tai parhaita vastaavuuksia.

Suodatusta käytetään myös signaalinkäsittelyn ominaisuuksien vähentämiseen tai parantamiseen. Yleisiä sovelluksia ovat kohinanvaimennus ja reunojen tunnistus kuva-analyysissä.



Lajittelu on tietojen järjestämistä joidenkin kriteerien mukaan. Se voi olla nouseva (pienimmästä suurimpaan, A: sta Z: hen) tai laskeva (suurimmasta pienimpään, Z: stä A: han).

Esimerkkejä suodatuksesta ja lajittelusta:

- Etsi 10 yleisintä lyhytaikaisten sairauspoissaolojen syytä (diagnoosia) Kokkolan seudulla vuosina 2017–2018.
- Etsi viikonpäivä, jolloin maaperän kokonaisnettolämpö on pienin maalämpö geotermisesti lämmitetyssä rakennuksessa.
- Etsi lähin kohde konenäkö- tai robotiikkasovelluksesta.
- Etsi objektintunnistustehtävästä vain objekteja, joiden koko on yli 150 kuvapistettä.

Valmis!

Onnittelut! Olet täyttänyt materiaalit!



**Euroopan unionin
osarahoittama**



**Kokkola
Karleby**

centria
University of Applied Sciences

