

Weekly task 6 - Correlation Analysis / Korrelaatioanalyysi

Taustat

Korrelaatioanalyysi on tilastollinen menetelmä, jonka avulla voidaan selvittää kahden tai useamman muuttujan väliset suhteet. Korrelaatioita on eri tyyppisiä: positiivisia ja negatiivisia, lineaarisia ja ei-lineaarisia sekä yksin- ja moninkertaisia. Korrelaatio voi olla kokonaisvaltaista tai osittaista. Korrelaatioanalyysin ”isoisä” on ranskalainen Francis Galton, joka laati ensimmäisiä tieteellisiä artikkeleita korrelaatio- ja regressioanalyysiä koskien.

Tietotekniikan aihepiirissä korrelaatioanalyysi soveltuu monenlaiseen tutkimukseen, mutta kvantitatiivisen luonteensa vuoksi se vaatii aina suuren määrän kerättyä informaatiota. Siispä se ei sovellu tilanteeseen, jossa halutaan tutkia jotakin harvinaista tai kallista uutta teknologiaa, jolle ei pystytä toteuttamaan riittävästi testitapauksia.

Esimerkiksi artikkelissa *DCCA cross-correlation analysis of 3D wind field signals in indoor and outdoor environments* (Zeng ym, 2016, doi: 10.1109/WCICA.2016.7578747) käytetään korrelaatioanalyysin keinoja, kun suunnitellaan robottia kaasulähteen etsimistä varten.

Tiedon keräys

Korrelaatioanalyysi on määritelmältään kvantitatiivinen tutkimusmenetelmä. Näin ollen menetelmän luonteeseen soveltuu paremmin deduktiivinen näkökulma. Datan keräys voi tapahtua monella tapaa: Se voi tapahtua havainnoimalla olemassa olevia tapahtumia tai luomalla omia testausympäristöjä. Datalle on tyypillistä, että se on numeerisessa muodossa, mutta se ei ole välttämätöntä menetelmän hyödyntämisen kannalta.

Korrelaatioanalyysissä voidaan käyttää erilaisia kertoimia, jotka on nimetty niiden keksijöiden mukaan. Pearsonin korrelaatiokerroin perustuu siihen, miten lähekkäin mitatut tai muuten observoidut arvot ovat toisiaan jollakin asteikolla. Se on siis lineaarinen. Spearmanin korrelaatiokerroin on hieman samantyyppinen kuin Pearsonin, mutta se ei vaadi muuttujilta lineaarista luonnetta: Riittää, että muuttujilla on monotoninen suhde, eli vähintään toinen riippuu toisesta. Kendallin korrelaatiokerroin puolestaan ei ota kantaa siihen, miten lähekkäin arvot ovat, vaan siihen, miten monta muuttujaa ovat yhtäsuuria eri mittauskohteissa. Siispä Kendallin korrelaatiokerroin soveltuu hyvin myös ei-lineaarisille muuttujille.

Toteutus

Fiktiivisessä tutkimuksessamme tutkitaan, millainen korrelaatio on kiintolevyn käyttölämpötilan ja käyttöiän välillä. Näytteeksi tutkimukseen hankitaan 1000 samanlaista kiintolevyä, jolloin saadaan mahdollisimman monta ulkoista muuttujaa eliminoitua.

Tutkimuksessa mitattavat muuttujat ovat testauspaikan lämpötila ja kiintolevyn käyttöikä, kunnes tietty määrä segmenteistä on käyttökelvottomia. Molemmat ovat lineaarisia suureita, joten niitä voidaan mitata selkeällä asteikolla. Tutkimuksessa kontrolloidaan testauspaikan lämpötilaa keinoitekoisesti lämmitysjärjestelmällä, jonka avulla lämpötila voidaan säätää vähintään yhden Celsius-asteen tarkkuudella.

Tutkimus tuottaa tiedon siitä, miten pitkään testattu kiintolevy keskimäärin kestää jatkuvaa kirjoittamista tietyssä lämpötilassa, ja sitä kautta sen, paljonko lämpötilalla on merkitystä kiintolevyä käytettäessä. Tätä tietoa voidaan hyödyntää esimerkiksi silloin, kun määritetään kiintolevyn sopivaa käyttölämpötilaa. Kiintolevyjen valmistajat voivat hyötyä tiedosta esimerkiksi takuuehtoja asettaessa. Myös esimerkiksi konesalien ylläpitäjät hyötyvät tutkimuksen tuloksista, kun he suunnittelevat riittävää jäähdytystä konesaliin.

Mielestämme korrelaatioanalyysi soveltuu tietotekniikan alalla hyvin varsinkin laadunvarmistukseen. Tietotekniikan alalla on yleensä suhteellisen helppoa muodostaa testausympäristöjä, joissa on hyvin kontrolloidut olosuhteet. Reaalimaailmassa toteutetun testaustapahtuman lisäksi on hyvä huomioda simuloinnin mahdollisuus, jolla voidaan myös toteuttaa korrelaatioanalyysijä.

Tutkimusmenetelmän luotettavuus on korkealla tasolla, sillä otanta on suuri ja ylimääräiset muuttujat on saatu eliminoidua ja testaustilanne on hyvin yksinkertainen. Silti on luonnollisesti syytä huomioda, että yksittäiset vialliset kiintolevykappaleet saattavat välttää tutkimuksen otannan.

Tutkimusmenetelmän kelpoisuus täytyy analysoida varauksella, sillä tutkimusta ei voida toteuttaa riittävän laajalla lämpötilaskaalalla. Saattaa olla, että kiintolevyjen optimaalisin käyttölämpötila onkin testatun lämpötila-asteikon ulkopuolella. On myös mahdollista, että asteikon ulkopuolella tapahtuu ennakoimatonta vaihtelua.

Korrelaatioanalyysiin ei liity eettisiä ongelmia ainakaan tässä fiktiivisessä tapauksessa.