Samuli Kotiranta
AB8349

Tehtävä 4 – Mallinnus

Varsinainen datan mallinnus tapahtuu koneoppimista hyödyntäen. Matemaattisen mallin rakentamista varten tarvitaan selkeästi rajattu ongelma, jollaiseksi valikoitui poikkeavan vedenkulutuksen seuranta, erityisesti vuototilanteita ajatellen.

Data on aikasarja. Regression käyttäminen on hieman vaarallista, sillä datan sisäiset korrelaatiot voivat aiheuttaa harhaa. Autoregressiomalli voisi sen sijaan toimia. Ehkäpä ongelmaa voisi kuitenkin lähestyä luokitteluongelmana, jossa pyritään seuraamaan yksittäisen talouden kuulumista normaalikulutustaan vastaavaan luokkaan ja varsinainen mielenkiinto kohdistuu yksinomaan tilanteisiin, joissa luokka muuttuu suuremman kulutuksen suuntaan. Luokittelun etuna on, että samalla vaivalla saatua luokitusta voidaan käyttää myös vaikkapa markkinoinnin kohdentamiseen. Ennustearvoa tällaisella luokituksella on vaikea nähdä, ellei sitten kiinnostuta laajemmin tutkimaan kuinka talouden piirteet vaikuttavat vedenkulutuksen kehittymiseen.

Toimivan työkalun tähän tarjoaa Pythonin koneoppimiskirjasto *scikit-learn*, jonka *neighbors* ja *ensemble* -moduuleissa on valmiita työkaluja outlierien ja datassa kehittyvien poikkeamien tunnistamiseen. Miinuspuolena kynnys havaita yksittäisen talouden poikkeava vedenkulutus voi olla hyvinkin korkealla. Jokin tukivektorikone voisi tarjota luokittelutyökaluksi hyvän lähtökohdan.

Seuraavan kysymyksen muodostaa, kuinka data jaetaan mallia varten opetus- ja testijoukkoihin. Teknisesti ottaen tässä pitäisi luokittaa data ensin jollakin tavalla (esim. tavanomainen vedenkulutus vs. poikkeava vedenkulutus) ja tarkistaa datan jakaantuminen eri luokkiin, jotta opetusjoukkoon varmasti saadaan kaikkien luokkien edustajia. Suurella datamäärällä tämä on työlästä. Helpommalla päästään jos käytetään *k*-kertaista ristiinvalidointia vaikkapa 90/10 -suhteella, jolloin toistot keskiarvoistavat huonoiksi valikoituneiden opetusjoukkojen tuottamaa heikompaa algoritmin ennustuskykyä. Validoinnissa *k* voisi nyt olla vaikka 10, koska poikkeavat arvot ovat harvinaisia.

Tuloksen arviointi onnistuu kätevästi ja hyvin automatisoitavasti sekaannusmatriisin avulla. Toisaalta ROC-käyrästä suoritusta on havainnollisempi seurata. Luokitteluaineisto on myös helppoa visualisoida silmäiltäväksi.