iStock | BrianAJackson

OSA II

TEKOÄLYN KANSSA ASUNTOKAUPOILLA

Ari Laitala

Jutun ensimmäisessä osassa päädyttiin tilanteeseen, jossa ylioppinut neuroverkko on ryhtynyt mallintamaan myös ilmiöön liittyvää satunnaisvaihtelua. Nyt käsitellään keinoja ylioppimisen ehkäisemiseksi ja verrataan neuroverkon suorituskykyä regressioanalyysin voimaan.

KUTEN ARVATA SAATTAA, on satunnaisuuden mallintaminen vaikeaa, ellei peräti mahdotonta ☺ Parhaimmillaankin satunnaisuutta mallintava malli pätee todellisuudessa vain satunnaisesti. Todelliseksi ongelmaksi ylioppiminen muodostuu kuitenkin vasta silloin, jos mallin laatija ei tunnista ylioppinutta mallia ja uskoo mallin kuvaavan reaalimaailman ilmiötä sellaisenaan.

RULETTIA

Havainnollistetaan asiaa rulettipelistä otetulla esimerkillä. Pelissä voidaan "lyödä vetoa" mm. siitä, minkä väriseen lokerikkoon pallo seuraavaksi jää. Vaihtoehtoina ovat musta (M, 18 kpl), punainen (P, 18 kpl) ja vihreä (V, 1 kpl). Oletetaan, että lähtödatana (opetusdatana) on yhdentoista alkeistapahtuman sarja: MMPMMPMMPMM. Sataprosenttisella varmuudella neuroverkko päättelee alta aikayksikön, että seuraava alkeistapahtuma on P. Näin päättelisimme myös me ihmiset, mikäli ainoa informaatio ilmiöstä olisi em. kirjainsarja. Yksi ylioppimisen syistä on siis suppea lähtödata, mutta vaikka aineisto olisi laajakin, on kutakuinkin täydellisen mallin muodostaminen mahdollista, kun laskentakapasiteettia vain on riittävästi.

ALIOPPIMINENKIN LUONNISTUU

Ylioppimisen ongelmaa voidaan ehkäistä monin yksinkertaisinkin tavoin, kuten pysäytetyllä laskennalla sekä k-kansion menetelmällä. Pysäytetyn laskennan idea on yksinkertainen. Sen sijaan, että neuroverkon annetaan jatkaa mallintamista, laskenta säädetään pysähtymään johonkin tiettyyn kynnysarvoon. Tällaisena kynnysarvona voidaan käyttää esim. epookkia eli sitä, kuinka monta kertaa opetusaineisto lasketaan läpi. Näennäisellä helppoudella on kuitenkin kääntöpuolensa. Se on riski siitä, että neuroverkko alioppii (underfitting). Alioppimisesta on kysymys, kun laskeminen katkaistaan liian varhain eikä kaikkea opetusaineistossa olevaa informaatiota saada hyödynnettyä mallin muodostuksessa. Käytännössä alioppiminen näkyy lähinnä alhaisena selitysasteena.

Yli-ja alioppimisen ehkäisyyn laskennan pituutta säätämällä ei näytä olevan tarjolla vuorenvarmoja yksinkertaisia keinoja. Yksi usein tarjoiltu on kuitenkin jakaa opetusdata kolmeen eri osioon, varsinaiseen opetusdataan, validointidataan ja testidataan. Data voidaan jakaa osajoukkoihin satunnaisotannalla esim. siten, että 60 % datasta muodostaa varsinaisen opetusdatan, ja loput 40 % jaetaan tasan validointi- ja testidataksi.

Eri lähteistä on tarjolla hiukan erilaisia ohjeistuksia mutta johtoajatus näyttää olevan se, että varsinainen mallin muodostus tehdään opetusdatalla siten, että validointidatan perusteella laskennan hyperparametrit säädetään sopiviksi. Opetusdatalla siis lasketaan esim. eri määrä epookkeja siten, että selitysaste nousee mahdollisimman korkeaksi validointidatassa. Lopuksi mallilla lasketaan testidata, jonka selitysaste sitten indikoi sitä, mihin uudella datajoukolla on mahdollista päästä ilman yli- ja alioppimista.

Toinen menetelmä ylioppimisen ehkäisemiseksi on k-kansion menetelmä, jossa ajatus on se, että aineisto jaetaan esim. 5 joukkoon. Näistä neljää joukkoa käytetään mallin muodostamiseen ja yhtä testidatana. Sitten laskentaa jatketaan siten, että seuraava osajoukko jätetään testidataksi ja loput muodostavat opetusdatan. Näin jatketaan, kunnes jokainen osajoukko on ollut kertaalleen testidatana.

PYSÄYTETTYÄ LASKENTAA

Kaivamme esiin kirjoituksen ykkösosassa esitellyn pääkaupunkiseudun omakotiaineiston. Tällä aineistolla kokeiltiin 10 000, 25 000, 50 000 ja 100 000 epookin laskemista (Huom. kuvion 2 tekstissä puhuttiin virheellisesti 100 000 epookista, vaikka laskentaikkunassa nähdään lukema 50 000). Kirjoituksen ensimmäisessä osassa pääteltiin, että 25 000 epookin laskeminen näyttää johtavan vakavaan ylioppimiseen selitysasteen noustessa yli 93 prosentin. Lasketaan hintamalli seuraavaksi 25, 250 ja 2 500 epookilla. Vastaavat selitysasteet ovat 36, 72 ja 84 prosenttia. Selitysasteet eivät kuitenkaan anna vihiä siitä, missä vaiheessa alioppiminen muuttuu ylioppimiseksi. Kokemusperäisesti voidaan arvioida, että tällä aineistolla 36 % lienee kuitenkin liian matala selitysaste.

Pysäytetystä laskennasta on olemassa hiukan sofistikoituneempi versio nimeltään Percentage Split. Siinä aineisto jaetaan esim. 2/3 ja 1/3 suuruisiin satunnaisotantoihin. Tässä variaatiosssa 2/3 otoksesta käytetään mallin muodostamiseen. Loppuosalle lasketaan saadulla mallilla sitten hintaennusteet. Menettelyn hyöty on siinä, että hintaennusteista on olemassa myös todelliset arvot, jolloin päästään vertaamaan laskettuja ja todellisia arvoja. Näin saadaan jo varsin hyvä käsitys mallin suorituskyvystä. Eli jos mallinnus viedään liian pitkälle, alkaa selitysaste varsinaisessa testidatassa laskea. Tämä laskentatapa on siis varsin lähellä k-kansion menetelmää, k:n saadessa arvon 1.

Lasketaan mallit jälleen 25, 250, 2 500 ja 25 000 epookilla. Saadaan seuraava tulokset.

TAULUKKO 1.

Epookit	Selitysaste	Keskivirhe
25	18,96 %	105 409
250	50,04 %	79 677
2 500	36,04 %	111 830
25 000	33,13 %	145 445

Näyttää siltä, että paras selitysaste saavutetaan lähellä 250 epookkia. Selvitetään asiaa hiukan tarkemmin yrityksen ja erehdyksen kautta. Tehdään pikainen kokeilu 300 ja 200 epookilla. Ensin mainittu huonontaa selitysastetta, mutta jälkimmäinen nostaa sitä muutamalla kymmenyksellä ja 150 kohdalla se alkaa jälleen tippua. Päätellään siis, että tässä tapauksessa ali- ja ylioppimisen raja on n. 200 epookin kohdalla, jolloin saadaan aineiston 2/3 ja 1/3 jaolla selitysasteeksi noin 50,6 prosenttia.

KUMPI VOITTAA

Tässä vaiheessa kiinnostutaan myös siitä, millaisiin suorituksiin regressioanalyysillä päästää tällä aineistolla. Valitaan regressioanalyysin algoritmiksi stepwise eli hintamalliin lisätään yksi muuttuja kerrallaan, aina siinä järjestyksessä, joka korreloi parhaiten selitettävän muuttujan eli kokonaishinnan kanssa. Mallista voidaan myös tiputtaa jo sinne kerran lisätty muuttuja, mikäli sen selitysvoima (korrelaatio) heikkenee liiaksi malliin lisättävien uusien muuttujien johdosta. Muodostetaan malli samoin kuin neuroverkolla. Valitaan mallin muodostukseen satunnaisesti 2/3 osa aineistosta. Regression tuloksena saadaan malli, jonka selitysaste on 63,5 %.

Ensivaikutelma siis on että regressiopohjainen malli päihittää selvästi neuroverkkoon perustuvan kilpailijansa. Tehdään pieni lisätesti. Otetaan etuovi.com palvelusta listan ensimmäinen pääkaupunkiseudulta myynnissä oleva omakotitalo-kohde. Kohteen tiedot on lisätty alla olevaan taulukkoon, jossa näkyvät myös äsken lasketun regressiomallin kertoimet. Kohde sijaitsee Vantaalla.

TAULUKKO 2.

Regressiomalli				Neuroverkkomalli
Muuttuja	**kerroin**	**määrä**	**=**	**ANN**
intercept	184 280	1	184 280	
Area	1 580	101	159 580	
D_condgood	76 417	1	76 417	
Age	–2 002	6	–12 012	
D_sauna	–45 444	1	–45 444	
D_Kauniainen	238 087	0	0	
D_Helsinki	82 626	0	0	
D_Espoo	58 861	0	0	
D_condpoor	–74 563	0	0	
			362 821	300910

Kuten yllä olevista luvuista nähdään, ennusteissa on melko suuri ero. Regressioanalyysin tulos osuu lähes tarkalleen pyyntihintaan, joka on 365 000 euroa. Hintapyyntö ei kuitenkaan ole minkäänlainen todiste kohteen markkina-arvosta.

Regressiomalli kärsii myös voimakkaasta heteroskedastisuudesta siten, että yli 400 000 euron hinnoilla malli alkaa voimakkaasti aliarvostaa kohteita. Tämä ongelma voidaan yleensä korjata, mutta siihen ei tässä artikkelissa enää ryhdytä.

Regressioanalyysi näyttäisi muutaman indikatiivisen testin perusteella päihittävän neuroverkkokilpailijansa. Tulos on näin pienellä ja lineaarisesti mallintuvalla aineistolla odotettu. Tilanne yleensä kuitenkin muuttuu, kun data ja sen sisällä oleva kompleksisuus kasvaa.

Tässä kohtaa tarkkaavainen lukija on varmasti jäänyt pohtimaan, miksi yllä olevassa taulukossa ei ole esitetty neuroverkon laskemia kertoimia regressiomallin muuttujille. Syy on yksinkertainen. Tällaisia muuttujien kertoimia (varjohintoja) ei neuroverkkomalleissa ole olemassa. Kysymyksessä on ns. mustan laatikon ongelma, jonka olemukseen ei tässä uppouduta.

Tämän kaksiosaisen kirjoitussarjan tarkoituksena on ollut herättää kiinnostusta tekoälyä ja neuroverkkolaskentaa kohtaan. Olisi kovin suotavaa, että neuroverkkolaskennasta tulisi pian peruskauraa myös alan opiskelijoille. Ennen kaikkea oleellista meidän alallamme olisi löytää mahdollisimman mutkaton lähestymiskulma käytännön sovelluksiin. Ja toki asioiden teoreettisemmallakin reunalla olisi paljon voitettavaa. Esim. **Anna Tulkin** diplomityöstä ”Itseorganisoivan kartan soveltuvuus asuntojen hintojen arviointiin” on jo 21 vuotta. Työssä sovelletaan ns. Kohosen algoritmia, joka lienee laajimmalle levinnyt suomalainen tieteellinen innovaatio. Voitaneen siis sanoa, että kyllähän tässä jo perinteetkin velvoittavat.

AIHEESTA LISÄÄ:

Ari Laitalan videoesitys ”Tekoäly mukaan energiatehokkuustyöhön”, http://q-r.to/banoP8