Tehtävä 1: Koneoppiminen

Lue oheinen koneoppimista käsittelevä teksti ja vastaa kysymyksiin tekstin ja omien tietojesi pohjalta.

Koneoppiminen on tekoälyn alue, jonka menetelmien avulla voidaan lähestyä tietojärjestelmien toteuttamiseen liittyviä tehtäviä ilman, että järjestelmään ohjelmoidaan suoraan sääntöjä tehtävän ratkaisemiseen. Koneoppimisessa pyritään tuottamaan **opetusdatan** perusteella **malli**, jota voidaan hyödyntää tehtävän suorittamiseen. Esimerkiksi kuvantunnistuksessa opetusdata voi koostua joukosta kuvia joissa tunnistettava kohde esiintyy ja joukosta jossa kohdetta ei ole. Kuvantunnistukseen kehitetyn koneoppimismenetelmän avulla tällaisesta datasta voidaan tuottaa malli, jonka avulla joukot voidaan pyrkiä erottelemaan toisistaan.

Koneoppimismenetelmiä voidaan jakaa niiden käyttämän opetusdatan perusteella. Yhden keskeisen luokan muodostavat **ohjatut** menetelmät, joiden opetusdata koostuu **syötteistä** (input) ja niitä vastaavista **tulosteista** (output). Ohjatun koneoppimisen tavoitteena on muodostaa opetusdatan perusteella malli, joka tuottaa uusille syötteille oikeita tulosteita. Esimerkiksi roskapostin suodattamista voidaan lähestyä käyttäen ohjattua koneoppimista siten, että syöte koostuu sähköpostiviesteistä ja tuloste tiedosta, mitkä viesteistä ovat roskapostia. Tällaisesta datasta opittu malli voi sisältää tietoa esimerkiksi siitä, kuinka usein kukin sana esiintyy roskaposteissa verrattuna muihin sähköposteihin. Ohjattuja menetelmiä voidaan ryhmitellä edelleen niiden tulosteen mahdollisen arvojen perusteella. **Luokittelu** on ohjattua koneoppimista, jossa arvot ennustavat äärellistä määrää ennalta määriteltyjä kategorioita (luokkia). Esimerkiksi roskapostin suodattaminen on luokittelutehtävä, jossa tulosteen arvot edustavat luokkia `roska` ja `ei-roska`. **Regressio** taas on ohjattua koneoppimista, jossa tulosteen arvot ovat jatkuvia. Esimerkiksi lämpötilan ennustaminen on regressiotehtävä: lämpötilan mahdolliset arvot ovat jatkuvalla asteikolla.

Ohjaamattomassa koneoppimisessa opetusdata ei sisällä tulosteita, ja tavoitteena on syötedatan rakenteen mallintaminen esimerkiksi sen perusteella, mitkä syötteistä ovat keskenään samankaltaisia. **Ryvästäminen** on ohjattua luokittelua vastaava ohjaamaton oppimistehtävä, jossa tavoitteena on samankaltaisten syötteiden ryhmittely yhteen. Uutisteksteistä koostuvaa dataa ryvästämällä voi olla mahdollista löytää esimerkiksi ryhmiä, jotka käsittelevät politiikkaa, urheilua, ja säätä. Toisaalta, koska ohjaamattomassa oppimisessa toivottua tulosta ei ole määritelty ennalta, ryvästäminen voi tuottaa myös odottamattomia ryhmiä: ryvästysmenetelmä voi esimerkiksi ryhmitellä uutisia aihepiirin sijasta sen mukaan, ovatko uutistekstit pitkiä vai lyhyitä.

Koska koneoppimismenetelmät eivät pysty suoraan käsittelemään kaikkia mahdollisia syötteen muotoja, datan esitystapa koneoppimista varten on yksi keskeinen kysymys näiden menetelmien soveltamisessa. Sen sijaan, että koneoppimismenetelmälle annettaisiin suoraan syötteeksi esimerkiksi suomen kielellä kirjoitettuja tekstidokumentteja, syötedata esitetään tyypillisesti **piirteiden** muodossa: kukin piirre edustaa jotain syötteen ominaisuutta, ja piirteen arvo sitä, missä määrin tämä ominaisuus kuvaa syötettä. Esimerkiksi tekstin luokittelua varten on mahdollista määritellä jokaiselle kielen sanalle piirre, ja asettaa näiden arvot sen perusteella, kuinka monta kertaa kukin sana esiintyy kussakin dokumentissa. Sen sijaan, että piirteet määriteltäisiin tällä tavalla etukäteen, voidaan myös syötteen esitystapa oppia datasta **piirreoppimisen** avulla.

Toisin kuin ihmisten oppimisprosessissa, opitun tiedon unohtaminen ei ole haaste koneoppimisessa: koska kone voi opetella opetusdatan ulkoa, on helppoa kirjoittaa ohjelma, joka palauttaa aina oikean tulosteen kullekin opetusdatan syötteelle. Tällainen ohjelma ei kuitenkaan ole hyö-

dyllinen, ellei se pysty tuottamaan toivottuja tulosteita myös sellaisille syötteille, jotka eivät esiinny opetusdatassa. Esimerkiksi roskapostin suodatin, joka tunnistaa roskaksi ainoastaan sellaiset viestit, jotka on jo aikaisemmin luokiteltu roskapostiksi erehtyisi jokaisen uuden roskapostiviestin kohdalla, joka poikkeaa edes yhdellä merkillä opetusdatan viesteistä. Koneoppimisen keskeisenä tavoitteena onkin sellaisten mallien tuottaminen, jotka ennustavat toivottuja tuloksia opetusdatan lisäksi uusille syötteille, eli pystyvät **yleistämään** opetusdatan esimerkeistä. Mallin arviointiin käytetään **testidataa**, joka on erillinen opetusdatasta ja jota ei ole hyödynnetty menetelmän kehittämisessä. Jotta tuotettu malli olisi hyödyllinen käytännön tehtävissä ja sen arviointi antaisi realistisen tuloksen sen tarkkuudesta, on tärkeää että niin opetusdata kuin testidatakin edustavat sitä dataa johon valmista mallia tullaan soveltamaan.

Koneoppiminen ei aina tuota toivottuja tuloksia, ja sen menetelmien soveltamisessa on useita haasteita. Monet näistä haasteista liittyvät dataan. Jos opetusdatan määrä on riittämätön, sen perusteella voi olla mahdotonta oppia mallia, joka tuottaa toivottuja tuloksia: esimerkiksi tekstin luokittelumallin kehittämiseen tarvitaan usein satoja tai tuhansia esimerkkejä kustakin luokasta. Jos data taas ei edusta tehtävää johon menetelmää kehitetään, malliin voi syntyä vääristymiä ja se voi epäonnistua käytännössä. Jos esimerkiksi puheentunnistusmallin kehittämisessä käytetään vain selkeästi äänitettyä yleiskieltä, malli voi epäonnistua hälyisessä ympäristössä ja murteiden tunnistamisessa.

Koneoppimistehtäviin on mahdollista soveltaa monia eri menetelmiä, ja niiden toteuttamiseen on olemassa työkaluja monille eri ohjelmointikielille. Useat koneoppimismenetelmät perustuvat tilastollisiin menetelmiin. Esimerkiksi yksinkertaisen roskapostin luokittelumenetelmän voi muodostaa laskemalla kuinka usein kukin sana esiintyy kunkin luokan (`roska`/`ei-roska`) viesteissä opetusdatassa ja laskemalla sanakohtaisista todennäköisyyksistä arvion sille, että uusi viesti on roskaa. Monet viime vuosina kehitetyt koneoppimismenetelmät perustuvat **neuroverkkoihin**, jotka mallintavat toisiinsa kytkettyjä yksinkertaisia laskenta-yksikköjä. Näistä mm. kuvantunnistamisessa hyviä tuloksia saavuttaneet **syväoppimismenetelmät** mallintavat suuria, jopa miljoonista laskenta-yksiköistä koostuvia neuroverkkoja. Nämä menetelmät voivat oppia erittäin monimutkaisia kuvauksia syötteistä tulosteisiin, mutta niiden kouluttamiseen voidaan toisaalta tarvita erittäin suuri määrä dataa.

Kysymykset

Mitä koneoppimismenetelmiä seuraavissa järjestelmissä käytetään? *Valitse kullekin kysymyksissä 1.1–1.4 kuvatulle järjestelmälle oikeat vaihtoehdot. Valitse vain suoraan kuvauksen pohjalta perusteltavat vaihtoehdot, ja huomaa että oikeita valintoja voi olla useampia.*

Pisteytys: Täsmälleen oikeasta vastauksesta maksimipisteet. Puuttuvista tai ylimääräisistä vaihtoehdoista vähennetään pisteitä. Kunkin kysymyksen minimipistemäärä on 0.

Kysymys 1.1. (2 p.)

Järjestelmän A tarkoitus on tunnistaa tekstin kirjoittaja sen sisällön perusteella. Järjestelmä perustuu tilastolliseen malliin, johon on koostettu laajasta tekstiaineistosta tieto siitä, kuinka usein kukin kirjoittaja käyttää tiettyä sanaa tai sanojen yhdistelmää. Kun järjestelmään syötetään uusi teksti, se arvioi mallin perusteella kuka aineistossa edustettuna olevista henkilöistä on todennäköisimmin kirjoittanut tekstin. Järjestelmässä ja aineistossa on rajoituksena, että kullakin tekstillä on täsmälleen yksi kirjoittaja.

- ☐ Ohjattu koneoppiminen
- ☐ Ohjaamaton koneoppiminen
- ☐ Luokittelu
- ☐ Regressio
- ☐ Ryvästäminen
- ☐ Ei mitään yllä olevista

Kysymys 1.2. (2 p.)

Järjestelmä B on kehitetty osakekaupan tueksi. Järjestelmä sisältää mallin, joka pyrkii ennustamaan kunkin markkinoilla olevan osakkeen tulevan arvon sen perusteella, miten osakkeiden arvot ovat muuttuneet viimeisen vuoden aikana. Malli perustuu matemaattiseen funktioon, jonka parametrit on sovitettu käyttäen osakkeiden arvojen historiallista kehitystä edustavan tietokannan dataa.

- ☐ Ohjattu koneoppiminen
- ☐ Ohjaamaton koneoppiminen
- ☐ Luokittelu
- ☐ Regressio
- ☐ Ryvästäminen
- ☐ Ei mitään yllä olevista

Kysymys 1.3. (2 p.)

Järjestelmän C avulla voidaan hakea sähköpostikokoelmasta kaikki viestit joissa esiintyy käyttäjän määrittelemä sana. Järjestelmä perustuu tietokantaan, johon on koostettu kaikki kokoelmassa esiintyvät sanat ja kullekin sanalle viestit, joissa sana esiintyy. Järjestelmä tukee myös hakumuotoa, jossa tietokannasta haetaan syötteenä saadun sanan lisäksi myös sanoja, joiden järjestelmän kehittäjä on digitaalisen sanakirjan perusteella määritellyt tarkoittavan samaa asiaa.

- ☐ Ohjattu koneoppiminen
- ☐ Ohjaamaton koneoppiminen
- ☐ Luokittelu
- ☐ Regressio
- ☐ Ryvästäminen
- ☐ Ei mitään yllä olevista

Kysymys 1.4. (2 p.)

Järjestelmä D on kehitetty tekstiaineistojen organisointiin. Järjestelmä esittää kunkin sen tietokannassa olevan tekstin siinä esiintyvien sanojen jakaumana, ja muodostaa teksteistä ryhmiä, joiden tekstien jakaumat ovat keskenään samankaltaisia. Järjestelmä nimeää kunkin ryhmän sellaisten sanojen mukaan, jotka esiintyvät sen teksteissä useammin kuin muiden ryhmien teksteissä.

- ☐ Ohjattu koneoppiminen
- ☐ Ohjaamaton koneoppiminen
- ☐ Luokittelu
- ☐ Regressio
- ☐ Ryvästäminen
- ☐ Ei mitään yllä olevista

Mistä koneoppimiseen liittyvistä syistä seuraavat järjestelmät eivät tuota toivottuja tuloksia? *Valitse kullekin kysymyksissä 1.5–1.7 kuvatulle järjestelmälle oikeat vaihtoehdot, määrittele ongelma tai ongelmat tarkemmin omin sanoin, ja kerro miten ne voitaisiin ratkaista. Vastaa kuhunkin kysymykseen lyhyesti enintään 40 sanalla. Valitse vain suoraan kuvauksen pohjalta perusteltavat vaihtoehdot, ja huomaa että oikeita valintoja voi olla useampia.*

Esimerkkivastaus: Kysymys 1.x: Opetusdatan valinta – Sähköpostin luokittelumenetelmän opetusdata koostuu pelkistä roskapostiviesteistä. Opetusdataa tulisi laajentaa siten, että se sisältää myös esimerkkejä viesteistä, jotka eivät ole roskapostia.

Pisteytys: Täsmälleen oikeasta vastauksesta maksimipisteet. Puuttuvista tai ylimääräisistä vaihtoehdoista vähennetään pisteitä. Sanallinen perustelu on pakollinen. Puuttuva sanallinen perustelu tarkoittaa automaattisesti 0 pistettä. Kunkin kysymyksen minimipistemäärä on 0.

Kysymys 1.5. (2 p.)

E kehittää järjestelmää eläinlajien tunnistamiseen valokuvista Python-kieltä ja uusimpia syväoppimisen menetelmiä käyttäen. Järjestelmän kehitykseen koostetaan opetusdata, jossa on yksi valokuva kustakin eläinlajista jonka järjestelmä pyrkii tunnistamaan, sekä vastaavasti koostettu testidata. Järjestelmä koulutetaan luokittelemaan syötteenä oleva kuva siinä esiintyvää eläinlajia vastaavaan luokkaan. Järjestelmä epäonnistuu kuitenkin tuottamaan toivottuja tuloksia riippumatta siitä, miten se toteutetaan ja koulutetaan.

- ☐ Opetusdatan valinta
- ☐ Testidatan valinta
- ☐ Ohjelmointikielen valinta
- ☐ Koneoppimismenetelmän valinta
- ☐ Piirteiden valinta

Kysymys 1.6. (2 p.)

F omistaa laajan valokuva-aineiston, joka on järjestetty aiheittain siten, että kukin kuva esittää täsmälleen yhtä aihetta. F on laajentamassa aineistoa samoista aiheista otetuilla kuvilla, ja toivoo koneoppimispohjaista järjestelmää avuksi uusien kuvien järjestämiseen aiheiden perusteella. Järjestelmän kehittämisen tueksi F koostaa kustakin aiheesta sitä esittävien valokuvien kokoelman, joka lähetetään ohjelmistokehitykseen erikoistuneelle konsulttifirmalle. Tämä tuottaa järjestelmän käyttäen Lua-kielellä toteutetun koneoppimiskirjaston ryvästysmenetelmiä. Kun järjestelmä otetaan käyttöön, todetaan kuitenkin että sen tuottama ryhmittely poikkeaa täysin etukäteen toivotusta.

- ☐ Opetusdatan valinta
- ☐ Testidatan valinta
- ☐ Ohjelmointikielen valinta
- ☐ Koneoppimismenetelmän valinta
- ☐ Piirteiden valinta

Kysymys 1.7. (3 p.)

G kehittää kännykkä-applikaation tueksi kasvontunnistusmenetelmää käyttäen Java-kieltä ja neuroverkkoihin perustuvaa koneoppimiskirjastoa. Datan keräystä varten otetaan studio-olosuhteissa kattava kokoelma kuvia, joissa esiintyy monia eri kasvoja, ja vastaava kokoelma, jossa kasvoja ei esiinny. Aineisto jaetaan satunnaisesti opetus- ja testidataan, ja opetusdatan perusteella koulutetun piirreoppimispohjaisen järjestelmän todetaan luokittelevan testidatan materiaalin korkealla tarkkuudella. Tästä huolimatta kännykkä-applikaation käyttäjäpalautteen mukaan järjestelmän kasvontunnistustoiminto on erittäin epäluotettava.

- ☐ Opetusdatan valinta
- ☐ Testidatan valinta
- ☐ Ohjelmointikielen valinta
- ☐ Koneoppimismenetelmän valinta
- ☐ Piirteiden valinta