Masinõppe projektiks valmistumine: kontroll küsimustik

Oktoober 2020

Juhendmaterjalist – masinõppe projekti kontroll küsimustik

Käesoleva juhendi eesmärk on aidata otsustada, kas masinõppe projektiga võiks edasi liikuda ning kas erinevate nõuetega on arvestatud. Juhend sisaldab suuniseid masinõppeprojektides kasutatavate andmestike hindamiseks ja ettevalmistamiseks ning projekti läbiviimiseks. Selleks, et teha kindlaks, kas projekt vastab miinimumtingimustele peaksid kõik vastused olema „jah“. Kui mõnele tingimusele vastasite „ei“, siis see ei tähenda, et projektiga ei peaks jätkama, kuid soovitus oleks konsulteerida mõne eksperdiga.

Juhend on mõeldud kõigile projektis osalejatele, nii äripoole esindajatele kui ka tehnilise taustaga tiimi liikmetele. Juhend võib lisaks olla abiks hindamaks hilisemaid andmevajadusi masinõppe projektis kasutamiseks. Käesolevas juhendis väljatoodud loetelu teemasid ei ole ammendav. Kui leiate, et midagi on puudu, siis andke teada!

Ott Velsberg

Andmete juht, MKM

Terminoloogia

Alljärgnevas osas on väljatoodud olulisemad masinõppe ja andmeteaduse valdkonna terminid, mida on käesoleva juhendi lugemiseks vaja:

- Andmetunnus (*Feature*) – edaspidi ka lihtsalt tunnus on sisendatribuut, mida kasutatakse ennustamiseks. Näiteks rämpsposti tuvastamise ülesandes on üheks tunnustest see, kas e-kiri sisaldab saaja nime.
- Juhendatud õpe (*supervised learning*) - masinõppe alamliik, milles sisendi-väljundi paaride põhjal luuakse mudel, mis uue sisendi korral arvutab väljundi. Juhendatud õppe näiteks on e-kirjade klassifitseerimine rämpspostiks või mitte-rämpspostiks. Käesolevas juhendis keskendume peaasjalikult juhendatud masinõppel, kui pole mainitud teisiti.
- Klassifitseerimine – juhendatud õppe meetod, mille eesmärk on ennustada, millisesse etteantud klassi objekt kuulub.
- Masinõpe - arvutiteaduse haru, mis annab arvutitele õppimisvõime ilma neid otseselt selleks programmeerimata.
- Masinõppemudel – edaspidi ka lihtsalt mudel on matemaatiliste sammude (tehete, võrdluste) loetelu, mille abil saab ennustada tunnuste väärtuste põhjal märgendi väärtust.
- Märgend (*target, response, label*) - tunnus, mille sihtväärtust soovitakse ennustada. Näiteks rämpsposti tuvastamise ülesandes võiksime „on rämpspost“ kutsuda märgendiks kuna selle väärtus selgub hiljem kui teised tunnused.

- Regressioon – juhendatud õppe meetod, mille eesmärk on ennustada mõnda pidevat väärtust, näiteks õhutemperatuuri. Väljund on arvuline.
- Sisend – tunnus, mille väärtuste põhjal soovite ennustada.
- Treeningandmestik – andmed (sisend-väljundite paarid) masinõppemudeli treenimiseks.
- Valideerimisandmestik - andmed masinõppe mudeli valideerimiseks.
- Väljund – tunnus, mille väärtust soovitakse ennustada.

Kontroll küsimustik

☐ **1.** Andmete ettevalmistel on kõik toimingud (teisendused, treening- ja testandmeteks jagamine, puuduvate tunnuste täitmine jne) dokumenteeritud.

☐ **2.** Algsed (toor)andmed on võimalikult muutumatul kujul salvestatud ja kättesaadavad. Oluline kuna võimaldab vigade korral alustada nn puhtalt lehelt.

☐ **3.** Kas andmed on tõesed/usaldusväärsed?

Mudeli treenimisel on oluline, et andmed, sealhulgas märgendid, oleksid õiged. Vastasel juhul ei saa mudel anda häid tulemusi. Kui teate, et andmekvaliteediga on probleeme, siis püüdke probleem tuvastada ja vead parandada, näiteks täita puuduvad väärtused asendusväärtusega. Teatud juhtudel võivad andmevead omada suurt prognoosiväärtust, sellistel juhtudel ärge asendage vigased andmeid ega kustutage neid, pigem looge eraldi väli.

☐ **4.** Kas andmed on hõlpsasti juurdepääsetavad ja projektis kasutatavad?

Kui andmetele juurdepääs pole piisavalt hõlbus tuleks enne projektiga alustamist piirangud kõrvaldada. Lisaks tuleks arvestada, et andmed oleksid piisavalt kiirelt loetavad, et ei tekiks pikki viivitusi – see on oluline just suurte andmemahtude korral. Samuti on vajalik veenduda enne projektiga alustamist, et teise asutuse andmeid saaks kindlasti projektis kasutada.

☐ **5.** Kas andmed on kirjeldatud?

Oluline on omada ülevaadet projektis kasutatavetest andmestikest ning neis sisalduvatest andmetest. Andmete kirjeldamisel on oluline, et andmeelemendi kirjeldus kataks järgnevat: andmeelemendi tähis, nimetus, lühikirjeldus, andmeelemendi tüüp, andmeelemendi kasutamise/kogumise algus ning vajadusel mõõtmise kirjeldus ning mõõtühik. Soovitus on juhinduda andmekirjelduse standardist, mis põhineb DCAT ja DDI standardil ning on leitav kratid.ee lehel.

☐ **6.** Kas andmed on läbivalt kokkulepitud kirjaviisiga?

Oluline on tagada andmete süntaktiline õigsus, kus tunnuste väärtused on tehniliselt korrektselt kirjas. Näiteks kirjavigade puudumine, ühtne suure ja väikese algustähe kasutus, ühesugune asutuste nimede kasutus jne.

☐ **7.** Kas kõigil andmeväljadel on läbivalt sama tähendus?

☐ **8.** Kas andmed sisaldavad sihtmärgi ennustamist võimaldavat teavet ehk kas olemasoleva põhjal on võimalik ennustada sihtmärki?

Näiteks liiklusõnnetuse toimumise prognoosimiseks konkreetsel teelõigul võiks vaja minna teavet teelõigu piirkiiruse, tee kuju, teeolude, toimunud liiklusõnnetuste, õhu temperatuuri ja muu kohta. Selline teave nagu sõidukijuhtide nimed ei ole prognoosimiseks vajalikud.

☐ **9.** Kas andmed on ennustamiseks piisava sageduse/kaetusega?

Näiteks kord päevas või kord nädalas mõõdetud õhutemperatuur ei ole liiklusõnnetuste toimumise prognoosimiseks piisavalt sage. Samuti ei saa ennustada koondandmete põhjal, näiteks kuus toimunud liiklusõnnetuste arv.

☐ **10.** Kas puuduvad väärtused moodustavad sisendist väikese osa?

Oluline on arvestada, et puuduvad väärtused võivad olulisel määral mõjutada andmete kasulikkust ja kasutatavust. Tegu on sagedase olukorraga, mille osas liigselt ei pea muretsema. Soovituslik on andmete imputeerimine.

☐ **11.** Kas puuduvate väärtuste põhjused on teada?

Vajalik veenduda, et puuduvate väärtuste jaotus on juhuslik. Kui väärtused puuduvad reeglipäraselt/süstemaatiliselt, siis võib andmete põhjal loodavas mudelis tulla nihe ning mudel võib rakendamisel mitte töötada.

☐ **12.** Kas andmed on märgendatud?

Märgendatud andmed on vajalikud juhendatud õppeks. Näiteks pildituvastuseks on vajalik märgendada pildi või videofailid, et muuta need masinloetavaks. Juhend piltide- ja videote annoteerimiseks on leitav kratid.ee lehel.

Kui sihtmärk on kategooriline on vajalik märgendatud andmeid kõigis võimalikes kategooriates. Näiteks keelatud reklaamteksti tuvastamiseks ei piisa ainult keelatud reklaamtekstist vaid vaja läheb ka lubatud reklaamtekste.

Kui sihtmärk on arvuline on vajalik märgendatud andmeid piisavalt erinevate väärtuste kohta. Näiteks piirikontrollis sõidukite pikkuse hindamiseks piltide põhjal on vaja pilte nii pikkadest, lühikestest kui ka keskmise pikkusega sõidukitest ning nende tegelikke pikkusi.

☐ **13.** Kas andmeid on piisavalt?

Andmete piisavuse hindamiseks pole ühtset kindlat viisi. Hindamisel võib lähtuda järgnevast. Iga tunnuse kohta vähemalt 20 näidet ning arvuliste väärtuste ennustamisel vähemalt 50 näidet. Ennustava mudeli loomise puhul on soovituslik kasutada andmeid vähemalt n+2 printsiibil ehk näiteks liiklusõnnetustes hukkunute prognoosimiseks järgneva 3 aasta jooksul võiks kasutada andmeid eelnevast 5 aastast.

☐ **14.** Kas andmed on nihketa?

Näiteks kui kasutate mudeli treenimiseks osalist andmestikku (näiteks andmekaitsepiirangute tõttu), siis veenduge, et kasutatavas andmestikus säilis esialgne andmete jaotus.