Masinõpe ei ennusta tulevikku — see hindab tõenäosusi

2 kuud tagasi - 11.08.2026 Autor: AM

(Sisuturundus)

Masinõppe mudelid ei anna kindlaid vastuseid. See on oluline eristus, mida sageli valesti mõistetakse. Kui algoritm ennustab spordimängu tulemust, ei väida ta, et teab, mis juhtub — ta ütleb, kui tõenäoline on iga võimalik tulemus, lähtudes sellest, mida ajaloolised andmed toetavad. Sport sobib selle mehaanika selgitamiseks hästi: tulemused on mõõdetavad, andmeid on rohkesti ja võistluskontekst on enamikule lugejatele intuitiivselt arusaadav. Küsimus, millele see artikkel vastab, on täpne — kuidas jõuab juhendatud õppe konveier toorstest andmetest kalibreeritud tõenäosusprognoosini?

Tarkvaraarendaja pilguga tootmissüsteemideni

Paulius Pazdrazdys, täispinu programmeerija ja tehisintellekti ekspert, tunneb seda konveieri arhitektuuri äravahetamiseni — ta on ise sarnaseid süsteeme ehitanud. Tema jaoks on oluline just see, et akadeemilistest prototüüpidest räägitakse tihti eraldi reaalsetest rakendustest, kuigi pipeline on põhimõtteliselt sama. Juhendatud õpe treenib mudelit märgistatud ajaloolistel näidetel ja väljastab tõenäosusi, mitte binaarseid otsuseid — just see arhitektuur töötab ka tootmiskeskkonnas.

“Insenerile, kes on ise mudeli treenimisega tegelenud, on selge, et see pole teadusprojekt — see on tarkvara, mis töötleb ajalooandmeid ja tagastab tõenäosushinnanguid. Sama loogika toimib igas valdkonnas, kus ajaloolisi tulemusi saab mõõta.”

Smart Betting Guide'i avaleht on näide sellest, kuidas see konveier tarbijatele suunatud kujul välja näeb — platvorm töötleb mängude ajalooandmeid ja esitab tõenäosushinnanguid, mida saab kõrvutada bukmeekeri kordajatega. Paulius vaatab seda inseneripliiatsiga käes, mitte mängijana.

Juhendatud õppe konveier algusest lõpuni

Kõik algab märgistatud andmetest. Juhendatud õppes tähendab see, et treenimisnäited on paaristataud teadaolevate tulemustega — näiteks mängude tulemused koos kõigi neid mängusid iseloomustavate näitajatega. Mudel ei saa lihtsalt andmeid üle vaadata ja kohe nutikaks saada. Ta itereerib: kohandab oma sisemisi parameetreid korduvalt, et minimeerida kaofunktsiooni, mis mõõdab, kui kaugele jäävad tema prognoosid tegelikest tulemustest.

Treenimise lõppedes tuleb küsida, kas mudel on päriselt midagi õppinud või lihtsalt treeningandmed pähe õppinud. Selleks kasutatakse eraldatud testkogumit — andmeid, mida mudel treenimise ajal pole näinud. Testkogum annab ausat hinnangut, kui hästi mudel üldistub uutele, nägemata juhtumitele. Ilma selle sammuta pole mingit garantiid, et treeningul saadud tulemusnäitajad peavad pärismaailmas paika.

Torandmetest mudelit söötvateks tunnusteks

Algoritmi valik mõjutab tulemust vähem kui arvata võiks. Tegelik töö on tunnuste inseneeria — see on protsess, millega torandmed valitakse, teisendatakse ja kombineeritakse mudeli jaoks kasulikeks sisenditeks. Kehvade tunnustega ei päästa head mudelit miski.

Spordivõistluste puhul tähendab see, et mudelile ei anta lihtsalt kahte meeskonna nime. Sisenditeks on näiteks meeskonna viimaste mängude vorm, ajaloolised omavahelised tulemused, koduväljakuefekt ja põhimängijate vigastused või võistluskeelud — igaüks kodeerib võistluskonteksti erinevat dimensiooni. Iga selline tunnus annab mudelile täiendavat eristusvõimet, eeldusel et see tunnus tõeliselt tulevaste tulemustega korreleerub.

Tunnuste inseneeria käib käsikäes ristvalideerimisega. Selle asemel, et lõigata andmed korra kaheks — treening ja test — jaotab ristvalideerimine andmed mitmeks osaks ehk foldiks ning treenib ja hindab mudelit erinevatel kombinatsioonidel. Tulemuseks on stabiilsem ja vähem optimistlik hinnang mudelil tegeliku üldistusvõime kohta.

Kolm mudelifamiliat ja nende loogika

Mudeli valik sõltub andmete struktuurist ja sellest, milliseid seoseid oodatakse. Logistiline regressioon on lihtsaim lähtepunkt: ta rakendab tunnuste lineaarsele kombinatsioonile sigmoidteisendust, mis surub väljundi tõenäosuse vahemikku nullist üheni. Tulemus on otse tõlgendatav ja mudeli käitumine on läbipaistev,

aga lineaarne struktuur tähendab, et keerulisi mittelineaarseid seoseid ta ise üles ei leia.

Gradienttugevdamine lahendab selle teisiti. See ehitab otsuslugude ansamblit järjestikku — iga uus puu treenib parandama eelmiste jäänukvigasid. Tulemus on mudel, mis suudab tabada keerulisemaid mustreid, jäädes samas suhteliselt andmesäästlikuks.

Närvivõrgud lisavad võrrandisse sügavuse. Ühenduste kaalud kohanduvad kihtide vahel, võimaldades mudelil õppida abstraktseid esitusi, mis ühendavad tunnuseid viisil, mida manuaalne inseneeria ei pruugi ette näha. Keerukate mittelineaarsete seoste jaoks on see tugev tööriist — aga selle eest tuleb maksta suurema andmete- ja arvutusvajadusega. Väikeste spordidatasetide korral tekib oht, et võrk õpib treeningandmed pähe, mitte ei üldista.

Miks ka parimad mudelid ei kõrvalda ebakindlust

Iga mudelit varjutab kahe veaallika pinge. Kõrge nihke korral on mudel liiga lihtne — ta ignoreerib andmetes tegelikult esinevaid mustreid. Kõrge dispersiooni korral on mudel liiga keeruline — ta on memoriseerinud treeningandmete müra, mis pärismaailmas ei kordu. See nihke-dispersiooni kompromiss on iga ennustusmudeli keskne projekteerimisküsimus ja seda ei saa lahendada, ainult hallata.

Kuid isegi täiuslikult kalibreeritud mudel tabab lakke, mida ei saa läbida. Redutseerimatu viga ehk aleatoristlik ebakindlus on igale reaalse maailma stohastilisele süsteemile omane müra. Jalgpallimängus võib võtmemängija libastuda, otsustav löök tabada poste, mitte väravat — need ei ole andmetes ennustatavad sündmused. Ükski mudel, olenemata keerukusest, ei suuda seda viga kaotada.

See ei ole nõrkus, vaid aus peegeldustus. Tõenäosuslik väljund ei tähenda, et mudel on ebakindel oma suutlikkuses — see tähendab, et mudel kajastab ausalt, mida andmed toetada suudavad.

Mida tõenäosuspõhine prognoos tegelikult tähendab

Treenimine, tunnuste inseneeria, mudeli valik, ristvalideerimine ja ebakindluse hindamine ei ole eraldiseisvad sammud — need moodustavad ühtse vastuse küsimusele, mida algoritmiline prognoos tegelikult esindab. Väljund ei ole ennustus selle sõna tavatähenduses. See on kalibreeritud hinnang, mis on tuletatud ajaloolistest mustritest, piiratud saadaolevate tunnustega ja tõkestatud redutseerimatu müra poolt, mida ükski konveier kõrvaldada ei suuda.

Seda mõistvad kasutajad suhtuvad ML-põhistesse prognoosidesse erinevalt neist, kes ootavad musta kasti oracle'it. Tõenäosus 0,68 ei tähenda, et tulemus on kindel — see tähendab, et mudel on tuvastanud mustreid, mis ajalooliselt toetavad seda tulemust 68 juhul sajast. Selle vahe mõistmine on see, mis eristab informeeritud kasutajat lihtsast numbrite järgijast.

- Uudised
- Sisuturundus

Pilt