Journal for General Philosophy of Science (2021) 52:171-176
https://doi.org/10.1007/s10838-020-09537-z

REPORT

A gépi tanulás és a tudományos magyarázat jövője

Florian J. Boge[1] - **Michael Poznic**[2]

Elfogadva: Online közzététel: 2020. december 24.

1 Bevezetés

2020. február 17-én és 18-án Rafaela Hillerbrand és Paul Grünke (mindketten Karlsruhei Műszaki Egyetem) a "The Impact of Computer Simulations and Machine Learning on the Epistemic Status of LHC Data" című kutatási projekt keretében szervezték meg a "Machine Learning: Magyarázat nélküli előrejelzés?" címmel a Karlsruhei Műszaki Intézetben (KIT).

A projekt a DFG/FWF által finanszírozott, interdiszciplináris "The Epis- temology of the LHC" elnevezésű kutatási egység része; ez a filozófusok, fizikusok, történészek és szociológusok egyedülálló együttműködése, amelyet a közelmúltban további három évre megújítottak.[1]

A workshop célja az volt, hogy összehozza a tudományfilozófusokat és a különböző területekről érkező, a gépi tanulási (ML) technikákat tanulmányozó és alkalmazó tudósokat, hogy interdiszciplináris keretet teremtsen a tudomány változó arculatának megvitatására az ML folyamatosan növekvő alkalmazásának fényében.

Mivel az ML egy bizonyos nézőpontból nem más, mint digitális számítógépek által végrehajtott (statisztikai) optimalizálás, feltételezhetjük, hogy fokozott használata a tudomány hagyományos magyarázó céljától való paradigmatikus elfordulást példázza a puszta szabálytalan felismerés és előrejelzés felé. Továbbá az is nyitott kérdés, hogy hogyan magyarázható az ML hasznosságának meghaladó mértéke, amint azt az elmúlt évek különböző összehasonlító tanulmányai is tanúsítják.

Tekintettel az ML e nehéz episztemológiai státuszára, elgondolkodhatunk használatának társadalmi következményein, valamint tudományos módszerként való történelmi és szisztematikus elhelyezésén.

Ennek megfelelően az előadások tartalmát (i) a gyakorlati szakemberek szempontjai, (ii) az ML magyarázatai, (iii) az ML magyarázatai, (iv) a társadalmi következmények, valamint (v) a globális és történelmi perspektívák szerint csoportosítva tárgyaljuk.

[1] Lásd még: http://dailynous.com/2020/01/20/2-6-million-funding-epistemology-large-hadron-collider/.

Florian J. Boge fjboge@uni-wuppertal.de

Michael Poznic
michael.poznic@kit.edu

[1] Interdiszciplináris Tudományos és Technológiai Tanulmányok Központja (IZWT), Bergische Universität Wuppertal, Gaußstr. 20, 42119 Wuppertal, Németország

[2] Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of

Technology (KIT), P.O. Box 3640, 76021 Karlsruhe, Németország

2 Gyakorlati szakemberek nézőpontjai

A workshopra meghívást kapott többek között Jan Cermak, Uwe Ehret és Erwin Zehe, a KIT három környezetkutatója is, hogy megosszák egymással az ML tudományos gyakorlatban való alkalmazásával kapcsolatos nézeteiket. Előadásuk három részre tagolódott, az első egy optimista, a második egy kíváncsi, szkeptikus, a harmadik pedig egy óvatos, vagy egyenesen pesszimista nézőpontot mutatott be.

Az elképzelés, miszerint az ML fokozott használata a magyarázattól a sikeres előrejelzés felé való fordulást példázza, egyik részben sem került megkérdőjelezésre. Ennek fontosságát azonban Cermak előadásában lekicsinyelték, aki a megbízható eredmények elérésében az ML előnyeire összpontosított, így optimista kilátásokat kínált. Cermak egyik fő állítása az volt, hogy az ML algoritmusok bizonyos esetekben valóban felülmúlják a legjobb fizikai modelleket a környezettudományban, és úgy viselkednek, mint egy jól képzett "szimatoló kutya".

Ezt a nézetet Ehret részben ellensúlyozta, aki ugyancsak elismerte, hogy az ML-módszerek az adattömörítés igen hatékony eszközei, és a tudományban széleskörűen alkalmazhatóak, de Cermak optimista megközelítésének részleges ellenérveként egy értelmezhető ML-re vonatkozó kívánalmakat is megfogalmazott, például azt, hogy "a helyes okokból legyen igaza".

Végül Zehe, aki a meghívott előadók egyike volt, és aki az átfogóbb szemlélet érdekében két kollégáját is bevonta a konferenciába, azt a (pesszimista) nézetet védte, hogy az éghajlati jelenségek okainak megértéséhez mindig szükség lesz további környezeti modellezésre. Zehe különösen azt a nézetet védte, hogy az ok-okozati összefüggések megértése továbbra is a tudomány globális célja marad, amelyet az ML nem ér el.

3 ML magyarázata

Florian Boge (Wuppertali Egyetem) és Thomas Grote (Tübingeni Egyetem) előadásában azt a kérdést feszegették, hogy lehetséges-e mégiscsak magyarázatokat szerezni sikeres ML-alkalmazásokból.

Boge előadásában a Balmer-képlet és Bohr atommodelljének példáját használta analógiaként arra a fajta szakadékra, amely a tudományos megértés és a felfedezés között a tudományban az ML használata miatt keletkezhet. Boge eredményei lényegében megegyeztek Zehe eredményeivel: Ahogyan a Balmer-képlet által megjósolt szabályszerűségek csak a Bohr-modell hátterében váltak érthetővé, úgy számos ML-alkalmazás további módosításokat igényel majd az előrejelzéseik megértéséhez. Az ML esetében azonban, érvelt Boge, a szakadék sokkal nagyobb lehet. Ezeket a megfontolásokat a részecskefizikából és az informatikából származó esettanulmányok váltották fel az ML fekete doboz jellegéről.

Hasonlóan Grote is vizsgálta az ML-alkalmazások által az orvosi kutatásban okozott nehézségeket. Grote előadásának középpontjában a magyarázhatóság és az értelmezhetőség megkülönböztetése állt, ahol az előbbi fogalom az ML-modellben fekete dobozba zárt különböző részletekre, az utóbbi pedig az ML-eredményekre való igazolás nehézségeire utal.

Ahogy Grote rámutatott, a kétféle igény erősen függ a felhasználó céljaitól: Míg az informatikusok, de az orvoskutatók is érdekeltek a magyarázhatóságban - magának az algoritmusnak a megértése, illetve új orvosi jelenségek felfedezése érdekében -, addig az orvosoknak csak az értelmezhetőséggel kell törődniük. Grote többek között szkeptikusan tárgyalta annak lehetőségét, hogy a multimodális magyarázatok, amelyek a természetes

nyelvű magyarázatok mellett vizualizációt is kínálnak, segíthetnek mindkettő növelésében.

4 Az ML magyarázatai

A legtöbb előadást az ML magyarázatának szentelték. Ezek közül az első Tom Sterkenburg (MCMP München) előadása volt, amely kapcsolatot teremtett az ML-kutatás és a formális episztemológia között. Az előadás kiindulópontja az volt, hogy mivel a formális episztemológia és az ML közös alapokon nyugszik, ha a formális episztemológia segít megérteni a tudomány sikerét, akkor ez átvihető az ML sikerének megértésére is. Sterkenburg fő kapcsolata az indukció problémája és a "nincs ingyen ebéd" tételek között volt, amelyek lényegében azt mondják, hogy egyetlen tanulási algoritmus sem teljesít a legjobban az összes elképzelhető feladatban. Ebből a szempontból Sterkenburg arra a következtetésre jutott, hogy a

Az ML magyarázatokhoz sokat tanulhatunk a filozófusok indukciós megközelítéseiből.

Hasonló üzenetet hordozott Timo Freiesleben előadása is (szintén MCMP), amely a magyarázható ML esetében az ellentételező magyarázatok használatát mutatta be. Freiesleben egy olyan formális keretet ismertetett az ML-irodalomból, amelyben a legközelebbi lehetséges bemenetet tekintjük, amely helyes előrejelzést eredményezett volna - ahogy Lewis is bevezette a lehetséges világok közötti távolságokat annak értékelésére, hogy milyen körülmények között lett volna "valami így és így".

Freiesleben ekkor azt javasolta, hogy lazítsunk azon a feltételezésen, hogy a kimeneti térnek értelmezhetőnek kell lennie, és összpontosítsunk azokra a feltételekre, amelyek mellett a jóslat beépül. Ez adhat magyarázatot az olyan ellenpéldákra, amikor egy kis mennyiségű zajt adunk hozzá például egy képhez, amelyet aztán a neurális hálózatok teljesen félreosztályoznak.

Szergej Titov (Szentpétervári Egyetem) előadása kapcsolatot vont az ML magyarázó képessége és a statisztikai relevancia magyarázatok között. Titov különösen a (homogén) partíciók szerepét hangsúlyozta a sikeres statisztikai magyarázatokban. Ha meg akarjuk magyarázni, hogy miért van valami, x, aminek A attribútuma van, és miért van B attribútuma is, akkor meg kell találnunk a B attribútum partícióját.

az összes As osztálya, úgy, hogy x-nek a partíció C_i cellájában való elhelyezkedése különbséget jelent a valószínűségben (azaz $P(B|A) \neq P(B|A \wedge C_i)$).

Olyan példák alapján, mint például egy macska képe, amelyet a karakterek alapján felismernek...

terisztikus fülek, orr és száj, de általában nem a pixelek eloszlása alapján, Titov ezután azt javasolta, hogy a statisztikai relevancia magyarázatok jó modelljei lehetnek az ML magyarázatoknak. Ezt a következtetést az ML-közösségben javasolt magyarázhatósági keretrendszerrel való összehasonlítással támasztotta alá.

Végül Maël Pégny (Université de Lorraine) különbséget tett az ML algoritmusok tudományos és pedagógiai magyarázatai között. Az előadás középpontjában az utóbbiak álltak, amelyek elsősorban a laikusok számára relevánsak. Pégny azonban felvetette, hogy ezek több szempontból is érdekesek lehetnek a tudósok számára.

5 Társadalmi következmények

Annette Zimmermann (Princeton Egyetem) meghívott előadása volt az egyetlen, amely kifejezetten annak a ténynek a társadalmi következményeire összpontosított, hogy az ML működését gyakran nehéz megmagyarázni. Zimmermann különösen a gyakran hivatkozott magyarázatokhoz való joggal foglalkozott az ML kontextusában, megkülönböztetve a magyarázatot az *igazolástól*.

Az átláthatatlan ML-technikák döntéshozatalra történő alkalmazása esetén a magyarázat nyilvánvaló hiánya szkepticizmushoz vezethet a kérdéses döntések indokoltságát illetően. Az ő

javasolt megoldás az volt, hogy komolyan fontolóra kell venni, hogy a magyarázat nem feltétlenül szükséges az igazoláshoz.

Ezt egy gondolatkísérlettel is hihetővé tették, amelyet "pontos Kate"-nek neveztek el: Ha a fiktív szereplő, Kate történetesen furcsa viselkedést mutat, amikor egy kávézóban látszólag véletlenszerűen leönti az embereket kávéval, de aztán kiderül, hogy ezzel pontosan kiszúrja a csalókat, akkor még mindig követelhetjük, hogy miért teszi Kate azt, amit tesz, de ettől függetlenül a viselkedését indokoltnak tekinthetjük.

Mivel az ML-algoritmusokról ismert, hogy a képzés során olyan dolgokat érnek el, mint például a faji előítélet, Zimmermann arra a következtetésre jutott, hogy a politikai döntéshozatali feladatok és a bírósági ítéletek során bizonyos algoritmusok használatának igazolása kulcsfontosságú lenne, függetlenül azok részletes magyarázhatóságától.

6 Globális és történelmi perspektívák

Andreas Kaminski (Stuttgarti Egyetem) és Johannes Len- hard (Kaiserslauterni Műszaki Egyetem) két meghívott előadása globális perspektívákat kínált a magyarázattal és előrejelzéssel kapcsolatban az ML kontextusában. Mindketten történelmi bizonyítékokkal is érveltek téziseik mellett.

Kaminski a "Types of Explanation, Kinds of Reason" című előadásában abból a kérdésből indult ki, hogy mit jelent az ML modellekben való bizalom. Abból kiindulva, hogy az ember elsősorban az emberekben bízik, és hogy ez a fajta bizalom egy nem emberi entitásban valami sajátos dolog, kétféle bizalmat különböztetett meg: egy normatív vagy etikai és egy episztemikus bizalmat.

Az előadás egyik köztes eredménye az volt, hogy megbízni valamiben, például egy modellben, azt jelenti, hogy meg tudjuk érteni vagy meg tudjuk magyarázni ezt a modellt. Ezzel a magyarázatról szóló tézissel a műhely központi témáját a bizalom szempontjából értelmeztük: Amikor az ML kontextusában magyarázatokat keresünk, akkor elsősorban az említett episztemikus értelemben vett bizalomra vagyunk kíváncsiak.

A legjelentősebb pontok között volt Kaminski vitája Heinz von Foerster triviális és nem-triviális gépek közötti megkülönböztetéséről. A nem triviális gépek nem vezetnek könnyen magyarázathoz, megértéshez és előrejelzéshez. Kaminski szerint a triviális és a nem triviális gépekhez kétféle átláthatatlanság kapcsolódik. A triviális gépekkel kapcsolatos társadalmi-technikai átláthatatlanság vagy fekete doboz-szerűség feloldható, míg a techno-matematikai átláthatatlanság másfajta magyarázatot igényel. Ez utóbbi gyakran az ML-re vonatkozik, és nem teszi lehetővé az egyszerű ok-okozati magyarázatot és megértést.

Lenhard "A matematizálás története és a jóslás új kultúrája" című előadása gazdag történeti beszámolót nyújtott a matematikai technikák gyakorlati célokra történő kifejlesztéséről, például az ágyúgolyók röppályájának előrejelzéséről. Nicolo Tartaglia tizenhatodik századi ballisztikai értekezése, a *La Nova Scientia* szolgáltatta a kiindulópontot, majd Lenhard felvázolta, hogy az ezt követő vita - amelyhez a tizenhetedik században Galilei, a tizennyolcadik században pedig Benjamin Robins és Leonhard Euler is hozzájárult - hogyan vezetett ahhoz az állításhoz, hogy a matematika követhető formákra való korlátozása a racionalitás konstitutív eleme.

Egy másik példa, amelyet Lenhard tárgyalt, egy keverék két komponensének desztilláció során történő szétválasztásának leírásának problémája volt a termodinamikai mérnöki gyakorlatban. Itt Len- hard "egy feltáró-iteratív előrejelzési kultúrát" fedezett fel: Az egyre jobb és jobb előrejelzések elérése érdekében az ember az ideális gáztörvényből

indul ki, majd a molekuláris részleteket is figyelembe vevő van der Waals-egyenlethez, és továbblép a Virial-egyenlethez, amely a molekuláris részleteket is figyelembe veszi.

különösen hasznos formát biztosít. Ezek a matematikai modellek általában nem követhetők, de a szimulációs modellezés lehetővé teszi, hogy előrejelzési célokra használjuk őket.

Az egyik fő probléma azonban a szabad paraméterek jelenléte az ilyen egyenletekben. Ezek beállítása a változatok sokaságát eredményezi, ami Lenhard szerint a szimulációs modellezés "sötét oldalát" jelenti. E hatás megszelídítése az elméleti mag és az állítható paraméterek közötti egyensúlyt igényli, amelyet elméleti és gyakorlati megfontolások kombinációjával kell kialakítani. Az ML-módszerek azonban Lenhard szerint ezt az egyensúlyt a végletekig eltolják, mivel két fontos különbséget tartalmaznak a szimulációs modellezéshez képest: Először is, az ML a nagy adatok felé irányul, így háttérbe szorítva az elméleti elemeket, másodszor pedig az ML-ben a mai napig csupán egy "rendszer" létezik, nem pedig az előrejelzés kultúrája.

7 Következtetés

A magyarázat témája az ML kontextusában minden bizonnyal a jövőben is a tudományfilozófusok érdeklődésének középpontjában marad. Tekintettel arra, hogy az ML sikeres tudományos alkalmazása terén számos előrelépés történt, a közeljövőben akár a tudományfilozófia egyik fő témájává is válhat. Amit a workshop minden bizonnyal megállapított, az az, hogy a környező kérdések megközelítéséhez különböző nézőpontokra van szükség. Nemcsak a magyarázat elvesztésével kapcsolatos különböző nézőpontok lehetségesek, amelyekkel akkor szembesülhetünk, amikor az ML sikeres előrejelzéseket ad, hanem magyarázatot kereshetünk éppen ezekre az előrejelzési sikerekre, gyökeret verhetünk a tudománytörténetben, vagy vizsgálhatjuk társadalmi következményeiket.

Összefoglalva, a fő eredmények között tarthatjuk számon a tudománytörténeti kérdésekkel, vagy akár a hagyományos filozófiai problémákkal, például Hume indukciós problémájával való szoros kapcsolatot, ahogy azt Kaminski, Lenhard és Sterkenburg előadásaiban megállapították. Ezzel némileg ellentétben állt Boge állítása az előrejelzés és a magyarázat közötti mély szakadékról, amely az ML tudományos alkalmazásainak sajátja, valamint Grote a tudományos felfedezés érdekében a magyarázhatóság hangsúlyozása. Nyitott kérdés marad, hogy ez a tudomány számára a gyakorlatban jelent-e problémát, ahogy azt Zehe, Cermak és Ehret előadásai is bizonyítják.

Ezzel szemben ott vannak az ML működésének, és különösen kudarcainak magyarázatával kapcsolatos sürgető kérdések, amelyekkel Freiesleben, Sterkenburg, Titov és Pégny előadásai foglalkoznak. Ezek akkor a legsürgetőbbek, ha az ML részt vesz a politikai döntéshozatalban és a politikai vagy orvosi döntések meghozatalában, ahogy azt Zimmermann előadása megállapította.

Itt jegyezzük meg azt is, hogy a *Minds and Machines* című folyóiratban a témáról különszámot fogadtak el publikálásra, amely várhatóan 2021-ben jelenik meg.[2] Tehát az az olvasó, aki mélyebb betekintést szeretne nyerni az ebben a jelentésben felvetett kérdésekbe, érdemes lehet erre figyelni.

Köszönetnyilvánítás FJB a "The Impact of Computer Simulations and Machine Learning on the Epistemic Status of LHC Data" című projektben dolgozott a DFG/FWF által finanszírozott "The Epistemology of the LHC" kutatási egységben, miközben e jelentés társszerzője volt, és ezért a Deutsche Forschungsgemeinschaft támogatását élvezte (FOR 2063 számú támogatás). Hasznot húztunk továbbá Gregor Schie- mann és Helmut Pulte észrevételeiből.

Finanszírozás A projekt DEAL által lehetővé tett és szervezett Open Access finanszírozás.

[2] Lásd https://www.springer.com/journal/11023/updates/18180316.