Af: Sune Hannibal Holm, lektor, ph.d., Sektion for Forbrug, Bioetik og Regulering, Institut for Fødevare- og Ressourceøkonomi, Københavns Universitet - suneh@ifro.ku.dk

DSAM Årsmøde 2024

Etiske værdier i kunstig intelligens

Algoritmerne kommer også til almen praksis. Hvilke etiske overvejelser skal vi gøre os, inden vi benytter dem til patientbehandling? Det var emnet for et af oplæggene ved DSAM's årsmøde 2024 om kunstig intelligens.

Der bliver i disse år udviklet en lang række maskinlæringsalgoritmer (ML-algoritmer), som kan komme med ganske nøjagtige forudsigelser om patienter. Der er således algoritmer, som kan detektere øjensygdomme og forskellige former for cancer og lungesygdomme. Kort sagt, ML-algoritmer, der kan komme med nøjagtige og pålidelige diagnoser, prognoser og behandlingsforslag, banker på døren til almen praksis.

Det vel nok mest almindelige argument for at anvende ML-algoritmer er, at de i kraft af deres nøjagtighed kan sikre bedre kvalitet for patienter og mere effektiv anvendelse af knappe ressourcer. En algoritme, der sætter almen praksis i stand til hurtigt og nøjagtigt at screene diabetikere for øjensygdomme, kan sikre, at en knap ressource som øjenlæger anvendes på dem, der faktisk har behov for den.

Nøjagtighed og effektivitet er imidlertid ikke hele historien om anvendelsen af ML-algoritmer. På trods af deres matematiske karakter og kvantitative natur er algoritmer ikke objektive og værdineutrale. Konstruktionen af en algoritme involverer en lang række beslutninger, som afspejler afvejning af etiske værdier: hvad er godt, og hvad er dårligt, hvad er retfærdigt, og hvad er uretfærdigt. Ansvarlig anvendelse af algoritmer kræver, at der er åbenhed om disse værdiafvejninger.

I det følgende vil jeg kort præsentere en række af de elementer i udviklingen af en ML-algoritme, der vil reflektere antagelser om, hvad der har værdi, uanset om man eksplicit har taget stilling til dem.

Formålet med algoritmen

En helt fundamental etisk overvejelse vedrører formålet med at investere i og anvende en algoritme. Hvad er det algoritmen skal hjælpe med at opnå? Det er afgørende, at formålet bliver beskrevet præcist, da det er afgørende for, hvilke mulige alternativer der kan være til at anvende en algoritme. Ydermere vil en beslutning om at investere i udvikling og implementering af en algoritme i realiteten antyde, at det antages at være en bedre løsning på et problem end tilgængelige alternativer.

Træningsdatasæt – afvejning af værdier

Som det ofte bliver fremført, er kvaliteten af en ML-algoritmes forudsigelser i høj grad afhængig af de data, der anvendes til at træne den. Det er træningsdatasættet, der giver algoritmen en repræsentation af den verden, den skal bruges i – verden ifølge data. Med andre ord er valget af træ-

> Konstruktionen af en algoritme involverer en lang række beslutninger, som afspejler afvejning af etiske værdier.

På trods af deres matematiske karakter og kvantitative natur er algoritmer ikke objektive og værdineutrale. Foto: Mogens Folkmann Andersen.

ningsdata et valg om, hvilken repræsentation af virkeligheden algoritmen skal lære af. Men hvornår er et træningsdatasæt godt nok? Det kan vi kun afgøre ved at afveje forskellige værdier.

Et eksempel på en værdiafvejning i forhold til træningsdata er følgende: En gruppe ingeniører er ved at udvikle en algoritme, der kan analysere billeder med det formål at identificere hudkræft. De har et budget til dataanskaffelse og skal beslutte, hvordan de skal konstruere deres træningsdatasæt. Hvordan skal de anvende deres midler? Bør de sikre, at der er lige repræsentation af forskellige hudtyper eller proportionel repræsentation i overensstemmelse med de pågældende hudtypers prævalens i befolkningen? Eller noget helt tredje? Og hvilke slags overvejelser ville understøtte den ene eller den anden beslutning? Måske vil den overordnede nøjagtighed af algoritmen være højere ved at tillade en stor ubalance i forhold til hudtyper. Men til gengæld vil de færre fejl primært ramme borgere med en hudtype, der er lavt repræsenteret i datasættet.

Uanset hvordan man vælger at gøre det, giver træningsdata ikke en værdineutral repræsentation af virkeligheden, som er uafhængig af menneskelige værdier og interesser. Når det kommer til at bestemme, hvad der skal inkluderes i et træningsdatasæt, vil værdiladede overvejelser altid spille ind.

➜

Er der en risiko for "defensiv medicin", hvor læger ikke tør gå imod anbefalinger fra AI?

Algoritmisk bias og fairness

Diskussionen vedrørende træningsdata knytter an til problemstillinger vedrørende algoritmisk bias og fairness. Ingen algoritme er perfekt. Der vil forekomme fejlforudsigelser i form af falske positive og falske negative.

I de seneste år er der kommet fokus på, at fordelingen af disse fejl kan ramme meget skævt i forhold til forskellige grupper. En hudkræftalgoritme kan f.eks. være ganske nøjagtig samlet set, men hvis alle dens falske negativer rammer borgere med mørk hudfarve, vil det ofte blive set som et tilfælde af moralsk problematisk bias.

Risiko for ulighed

Så hvad er en fair fordeling af fejl på tværs af grupper? Skal algoritmen have samme sensitivitet og specificitet for mænd og kvinder, mørkhudede og lyshudede? Uanset hvad vil en algoritme som udgangspunkt have forskellig kvalitet for forskellige grupper, og det vil derfor altid være forbundet med værdiafvejninger, hvordan man vælger at konstruere den. Er én type fejl værre eller bedre end en anden? Hvis ja, hvor meget bedre eller værre? Er det acceptabelt at have flere fejl for én gruppe end en anden, hvilket i sidste ende kan betyde, at der vil være en stor ulighed i fordelingen af ressourcer mellem de pågældende grupper?

Uigennemskuelighed er en svaghed

Den sidste vigtige etiske afvejning, jeg vil præsentere, vedrører det forhold, at ML-algoritmers høje nøjagtighed skyldes, at de er meget komplekse. Det betyder, at selv hvis vi mennesker generelt forstår, hvordan et neuralt netværk fungerer, så er det umuligt for os at gennemskue, hvordan et neuralt netværk når frem til en positiv klassifikation af et bestemt individ på baggrund af et bestemt input. Selv hvis vi printede beregningen ud, ville den være totalt uigennemskuelig.

.. og en styrke

Den uigennemskuelige karakter af ML-algoritmer er også deres styrke. Algoritmer, der kan lære af data, er ikke begrænset af menneskets evner til at finde mønstre i store datasæt og skrive computerprogrammer. Omkostningen er, at vi så heller ikke kan gennemskue dem. Der er derfor en vigtig etisk afvejning, som vedrører værdien af at få mere nøjagtige forudsigelser versus muligheden for at kunne gennemskue og forklare, hvordan et input om en specifik patient resulterer i en bestemt klassifikation af patienten. Hvis en algoritme foreslår en bestemt type behandling til patient X, vil det være oplagt at spørge, hvorfor den mener, at lige den behandling er bedst. Som det er nu, er det ikke oplagt, at hverken læge, patient eller nogen andre kan redegøre for det i det konkrete tilfælde. Der er dog en del forskning i "explainable AI," som arbejder på det.

Hvad skal vi overveje?

Selv hvis man har transparens om de skitserede etiske overvejelser, vil jeg mene, at læger, der eventuelt skal anvende dokumenteret nøjagtige og pålidelige algoritmer som beslutningsstøtte, har grund til at begynde at overveje en række spørgsmål.

Hvem bør pålægges ansvar for beslutninger, som er taget med AI-støtte? Hvordan kommunikerer man om AI-støttet beslutningstagning med patienter? Bør patienter have ret til at få en "second opinion" fra AI, hvis en valideret algoritme er tilgængelig? Har de krav på, at lægen ikke anvender AI som støtte? Er der en risiko for, at læger af frygt for at miste autoritet tøver med at anvende algoritmer som beslutningsstøtte? Er der en risiko for "defensiv medicin", hvor læger ikke tør gå imod anbefalinger fra AI?

Måske skal vi starte et andet sted

Der er andre og flere spørgsmål, som er vigtige at overveje i relation til AI i almen praksis og sundhedsvæsenet mere generelt. Måske er det mere presserende for læger at overveje, hvordan de kan anvende sprogmodeller til at udføre papirarbejde, som allerede nu fylder meget? Hvis der på den lange bane er en forventning om, at læger tager forudsigelsesalgoritmer i brug som støtte til at opnå bedre diagnoser, prognoser og behandlinger, vil jeg mene, at de spørgsmål jeg her har præsenteret bør overvejes allerede nu.

Referencer kan fås ved at kontakte forfatteren. //

Algoritmer, der kan lære af data, er ikke begrænset af menneskets evner til at finde mønstre i store datasæt og skrive computerprogrammer.