### 6.6 Skilsmisser og margarine: Korrelationer og årsager

En vigtig viden om ML-modeller er, at de virker ved korrelationer. To hændelser siges uformelt at være korrelerede, hvis de *følges ad*, således at når den ene af dem for eksempel vokser, så gør den anden det også. Nogle gange er der faktisk tale om en kausal sammenhæng, idet man kan finde ud af, at den ene hændelse er en direkte årsag til den anden, som har form af årsagens virkning. Effekten er her direkte, hvis den har form af, at årsagen *trigger* virkningen uden variation. Man kan måske tale om, at relationen nærmest er deterministisk: virkningen er forudbestemt af årsagen, hvorfor årsagen tydeligst må ligge forud for virkningen.

Men når ML-modeller foretager forudsigelser, så er der netop tale om korrelationer og ikke årsag-virkning-sammenhænge: Modellen er trænet på datasæt, hvor en label er tilknyttet hvert datapunkt, så visse variationer i datapunktet vil medføre en forandring i den tilhørende label. Men der er ikke tale om kausal virkning, idet vi (typisk) ikke kan på en række direkte virkninger, men i stedet ser sammenhængende opførsel.

For at illustrere forskellen mellem korrelation og kausalitet, kan man betragte nogle af de såkaldte *spurious correlations* (se Figur 5).

**Figur 5:** En korrelation, der næppe er en kausal sammenhæng – en *tilfældig korrelation* (*spurious correlation*), se https://www.tylervigen.com/spurious-correlations. Se også Ringgaard (2017) for yderligere forklaring og danske eksempler på tilfældige korrelationer.

Selvom to variable fx følges ad over en periode, behøver der langt fra at bestå en kausal sammenhæng imellem dem. Der kan fx være tale om, at begge variable er relaterede til en tredje variabel (*fælles faktor* eller *confounding factor*) eller at det tilsyneladende parallelle forløb er en ren tilfældighed (se Figur 6).

**Figur 6:** Figuren viser forskellige måder, hvorpå to størrelser kan være korrelerede: 1. Tilfældig korrelation, 2. *A* er årsag til *B*, 3. *B* er årsag til *A*, 4. *A* og *B* har en fælles årsag. Der er mange flere muligheder for relationerne, men disse fire alene viser, at man ikke fra korrelation mellem *A* og *B* kan udlede en kausal relation mellem *A* og *B*. Punkterne 2 og 3 alene viser, at eftersom kausalitet har en retning, og korrelationer ikke har, kan vi ikke engang afgøre, hvad der evt. er en årsag, og hvad der er en virkning.

### Sorte kasser for enhver

Vi bruger betegnelsen *black box* til at beskrive dele af et system, som for nogle relevante aktører forbliver lukket, uigennemskueligt og typisk af en størrelse, der overstiger, hvad et individ kan overskue. Denne definition spejler beskrivelser af, hvad der sker, når den individuelle autonome erkendelse bliver umuliggjort af forskellige slags kompleksitet (se fx

Sørensen og Andersen, 2018). Man kan altså skelne imellem forskellige gruppers behov for indsigt (dannelse) i *machine-learning-modellers* virkemåde og epistemiske status, og vi kan i hvert fald identificere tre forskellige grupper: forbrugere, anvendere og skabere af *machine-learning-modeller*.

Fra *forbrugerens perspektiv* vil det for fx forslagalgoritmer fra Netflix eller Facebook være relevant at vide, at anbefalingerne er baseret på, at det enkelte individ placeres i en *referencegruppe* med andre individer i den nærmeste klike baseret på modellens forudsigelser. Man får altså de samme anbefalinger som *andre ligesom en selv* også får, men hvor den nærmere beskrivelse af referencegruppen ikke er til at give, da den typisk er baseret på meget store og højdimentionale datasæt, som typisk forbliver en *black box* for forbrugeren. Resten af modellens opbygning, træning og tekniske funktion forbliver *black-boxet* for forbrugeren.

Anlægger man derimod det perspektiv, som tager udgangspunkt i dem, der skal *anvende machine-learning-modeller* fx i beslutningsprocesser eller til rådgivning, hører der yderligere elementer til en dannet forståelse. Foruden at vide, at forudsigelser er baseret på *referencegrupper*, vil anvenderen også skulle vide, at disse referencegrupper er baseret på fortidig data, og at *machine-learning-modeller* derfor har sit eget *induktionsproblem*: Den kan, ligesom en papegøje, kun gentage det, den allerede er trænet på. For anvenderen har dette særlig betydning, når biases i datasættet, som er baseret på fortidige forhold, propagerer ind i fremtiden. Men hvordan datasæt er *kurateret* vil ofte stadig forblive en *black box* for anvenderen, ligesom selve modellens interne funktion også vil være det.

Endelig, og måske mest relevant i denne kontekst, skal vi behandle spørgsmålet om nødvendig indsigt ud fra perspektivet, som tilhører en, der *skaber machine-learning-modeller*. Den overordnede skematiske beskrivelse af en superviseret *machine-learning-model* i Figur 1 dækker over en række elementer, der er af central betydning for udfordringerne ved at forklare *machine-learning-modeller* ud fra et *skaber-perspektiv*. I Figur 7 er denne figur udvidet med yderligere grupper af designvalg, implementationsvalg og hyperparametre, alle indrammet i blåt. Figuerens element om *datasæt* er indrammet i grønt, hvilket angiver, at indsigt i dette område fra et *skaber-perspektiv* vil blive taget op igen i forbindelse med diskussioner om etisk og professionelt ansvar.

Når man bygger og udvikler *machine-learning-modeller*, er der først en række *designvalg*, der skal træffes. Det drejer sig blandt andet om valg af det underliggende netværks arkitektur og antallet af neuroner, og det drejer sig om at fastsætte, hvorvidt neuronerne aktiveres. Disse valg er *pragmatiske* i den forstand, at de ikke behøver at være *absolut optimale*, men blot skal være *gode nok*, til at modellen kan udføre den funktion, den er bygget til, på en tilfredsstillende vis. Men i praksis er de også hyppigt formet af hensyn til *sædvane* og tidligere *erfaring*. På den måde ligner disse valg, som ingeniører træffer, når de skal udvikle nye teknologier, idet dog en omfattende inddragelse af *sædvane* og *taus viden* vil placere konstruktionen af *machine-learning-modeller* som snævert for egentlig ingeniørvidenskab.

Noget tilsvarende gør sig delvist gældende for de valg, som træffes under selve implementationen. Men implementationsvalgene er typisk noget, man lettere kan eksperimentere med at justere for bedre performance. På den måde synes disse valg mere at ligne empirisk udledte dimensioneringer, som foretages af *ingeniører*. En sidste række valg handler om modellens *hyperparametre*, som er parametre, der blandt andet bestemmer træningsprocessen og de tærskler, som forudsigelserne er underlagt. Disse parametre er (eller

er ved at blive det) empiriske, idet man enten kan træne modellen med mange forskellige parametre, ligesom en *scenarie-beregning*, eller man kan forsøge at få en slags meta-modeller til at forudsige passende *hyperparametre* for en given model.

Alle disse valg er med andre ord nogle, som skaberen af modellen skal forholde sig mere eller mindre eksplicit til. Når der er tale om pragmatiske valg kan den nærmere betydning ofte være uden for udviklerens rækkevidde, men ganske tit vil de kunne opsøges, og disse elementer er altså ikke *black-boxet* i samme grad, som tilfældet var fra de andre gruppers perspektiver. Til gengæld vil valg, der er truffet på grundlag af *sædvane og tavs viden* være epistemiske sorte kasser, idet den enkelte udvikler her deler fælles men tavs viden.