Statistikämnet är hotat – tack och lov

Data finns numera överallt och är en enorm drivkraft för modern industri. I denna blomstringstid för dataanalys finns paradoxalt nog en oro för statistikämnets död. Vi utmanas av närliggande ämnesområden som maskininlärning. Det är en välbehövlig katalysator till förnyelse av vårt ämne.

Jag har under åren 2011–2019 lett avdelningen för statistik och maskininlärning vid Linköpings universitet med ambitionen att integrera statistik med signalbehandling och datavetenskap, speciellt maskininlärning. I det arbetet har jag tillsammans med kollegor utvecklat och undervisat på ett stort antal kurser i maskininlärning för statistiker och ingenjörer, och forskat i gränslandet mellan statistik, maskininlärning och artificiell intelligens. Jag upplever därför inget hot från angränsande ämnen, utan ser tvärtom stora möjligheter till samarbeten och framför allt en vitalisering av statistikämnet.

Jag ska här föreslå sex områden där jag upplever att vårt ämne bör förändras om vi ska fortsätta att spela en central roll i det nya informationssamhället. Jag är väl medveten om att andra universitet, speciellt de som inte har en närhet till en teknisk högskola, kommer behöva välja en delvis annan väg än den vi valde i Linköping. Men jag vill hävda att mina sex förslag är applicerbara på alla utbildningar i statistik och matematisk statistik; de handlar om modernisering av undervisning i statistisk dataanalys och inferens. Som utgångspunkt för mitt inlägg ska jag först beskriva vad jag själv lägger i begrepp som artificiell intelligens, maskininlärning och data science.

Vad är artificiell intelligens och maskininlärning?

Artificiell intelligens (AI) handlar om att ge maskiner eller mjukvara någon form av människoliknande kognitiv förmåga. Mer konkret handlar AI om att göra det möjligt för en maskin att lära känna sin miljö och fatta beslut som maximerar sannolikheten att uppnå uppsatta mål. Moderna AI-system använder data för att lära upp dessa system med hjälp av maskininlärning. Under senare år har det skett stora framsteg inom den enklare formen av artificiell intelligens där maskiner utför mycket specialiserade uppgifter som att känna igen bilder eller förstå talat språk. Begreppen ”känna igen” och ”förstå” ska inte övertolkas här, dagens AI kan inte sägas förstå språk i någon djupare semantisk mening, men har lärt sig mönster i språket genom att tränas på enorma mängder text och ljud. Genombrott i s.k. generell AI där maskiner, likt människor, kan generalisera kunskaper till helt nya domäner har visat sig avsevärt svårare. En del AI-forskare föredrar därför att tala om utökad intelligens (eng. augmented intelligence) där maskiner (t ex mobiltelefoner) utökar en människas intelligens.

Maskininlärning (ML) är vetenskapen om metoder och algoritmer som gör det möjligt för en maskin eller mjukvara att använda data för att lära sig om världen och fatta optimala beslut under osäkerhet. Det klassiska exemplet är en robot som lär sig om omgivningen från ett antal sensorer och kontinuerligt fattar beslut för att uppnå ett mål. Ämnet överlappar med reglerteknik och signalbehandling. Både namnet maskininlärning och detta klassiska exempel målar dock upp en alltför snäv bild av ämnesområdet. Den stora majoriteten av tillämpningar av maskininlärning involverar inte fysisk hårdvara, utan handlar snarare om olika former av mjukvara med automatiserade prediktioner och beslut. Det kan t ex handla om s.k. appar som rekommenderar filmer åt filmkonsumenten, ger läkaren diagnosförslag baserat på patientmätningar, hjälper jordbrukaren med råd om gödsling från sensorer på åkern, eller utför automatiserad försäljning av aktier. ML har även börjat användas inom samhällsvetenskap och (digital) humaniora.

Under senare år har även termen *Data Science* (DS) populariserats och det finns en stor arbetsmarknad för s.k. data scientists. DS verkar betyda olika saker för olika personer. Det finns en stor spridning i examina på personer som titulerar sig som data scientists, allt från datavetare utan några större statistiska ämneskunskaper till personer med master- eller t.o.m. doktorsexamen i statistik/matematisk statistik. Det finns också alternativa titlar, som data analyst och data engineer, på olika delar av spektrumet från enkel datainsamling till avancerad dataanalys, som

Det är en balansakt att välja vilka delar av dagens kursutbud som ska bort och var betoningen i undervisningen ska läggas.

tolkas lite olika beroende vem man frågar. Kvalificerade statistiker som använder titeln data scientists eller liknande har i allmänhet större kunskaper i datahantering, programmering och algoritmer än den genomsnittlige statistikern.

I den vidaste definitionen handlar maskininlärning i princip om alla former av datadriven prediktion och beslutsfattande under osäkerhet. Och någonstans här börjar termen bli kontroversiell för oss statistiker. Det har ju vi statistiker hållit på med i minst 100 år! Jag är den förste att hålla med om att mycket inom ML är traditionella statistiska metoder under nya namn. Jag är också rätt less på den överdrivna tilltron till maskininlärning som ett slags trollstav, som kan omvandla brusiga skräpdata till precisa prediktioner. Men jag välkomnar också hur nya ämnen som ML utmanar statistikämnet med nya perspektiv, och tvingar oss åtminstone till självreflektion, men förhoppningsvis också till förändring. För även om statistik och ML arbetar med samma palett av inferens, prediktion och beslut, så är betoningen på olika delområden ofta radikalt olika. ML har mycket att lära sig från oss statistiker vad gäller modellering och inferensteori. Men vi statistiker kan också lära oss en hel del från ML om effektiva beräkningar, storskalig datahantering och om nya högintressanta tillämpningar och probabilistiska modeller för komplexa problem.

»»»

Sex områden där jag upplever att statistikämnet bör förändras

»»» **SEX OMRÅDEN DÄR JAG UPPLEVER ATT STATISTIKÄMNET BÖR FÖRÄNDRAS**

1. Prediktion och beslut

I traditionella statistikkurser hanteras prediktion ofta mycket styvmoderligt, t.ex. som en lite hastigt presenterad punktprediktion med tillhörande prediktionsintervall i regressionsanalys. Prediktion borde istället ha en central roll och genomsyra det mesta på våra kurser.

Beslutsfattande under osäkerhet är ofta den verkliga slutprodukten av en statistisk analys och borde också ha en mycket större roll i våra kurser. Inte bara för att moderna tillämpningar ofta kräver automatiserat beslutsfattande, men även för att beslutsperspektivet ger en tankeram som underlättar modellval och andra beslut under inferensfasen. Inom ML handlar det mesta om prediktion och beslut, och det får stora konsekvenser för modellering och inferens. Hypotestest används t ex i mycket mindre omfattning inom ML. Vi borde fundera på om vi ska fortsätta att ge så mycket utrymme åt hypotestest i våra kurser. Ett problembaserat beslutsperspektiv skulle automatiskt skifta fokus mot effektstorlekar.

Prediktion och beslut ger också konkreta motiverande exempel för studenter. Jag utvecklade för några år sedan en grundkurs i statistik där jag inleder kursen med att demonstrera prediktion av handskrivna siffror utifrån pixelerade bilder. Jag estimerar en prediktionsmodell i R från 50 000 handskrivna siffor och visar att modellen kan användas för att prediktera en sannolikhetsfördelning över siffrorna 0-9 för en ny pixelerad bild. Jag diskuterar automatiserat beslut utifrån modellen. Exemplet visar hela kedjan av data-sannolikhetsmodell-inferens-prediktion-beslut och förmedlar att statistik används för att lösa verkliga problem. Det finns inte en student som undrar ”vad statistik ska vara bra för” efter den demonstrationen. De exakta detaljerna kan naturligtvis inte förklaras vid första kurstillfället, med jag återkopplar regelbundet till exemplet under kursens gång.

2. Flexibla modeller och regularisering

Modeller i ML är ofta mycket mer flexibla än de modeller som vi lär ut på våra statistikkurser. Med flexibla modeller menar jag rikt parametriserade modeller som kan anpassa en stor klass av funktionella samband och fördelningar. Flexibla modeller behövs i ML där problemen ofta kan vara kraftigt olinjära och icke-Gaussiska. Modellerna är i princip alltid överparametriserade men kombineras med regularisering för att undvika överanpassning. Regularisering kan ses som en komplexitetskostnad som tas hänsyn till vid estimationen, t ex via penalized likelihood. De mest kända varianterna är L1- och L2-regularisering, som leder till Lasso respektive ridge estimatorer. Med regularisering blir den faktiska modellkomplexiteten ofta mycket lägre än antalet parametrar. ML-områdets inriktning mot prediktion och beslut minskar behovet av att kunna tolka individuella koefficienter och att beräkna deras signifikans i dessa komplexa modeller. Istället pågår forskning om hur man ska kunna förstå prediktioner och beslut från komplexa ML-system. Vid en olycka måste en självkörande bil kunna förklara *varför* den valde att väja för ett objekt på vägen, dvs vilka sensordata triggade just det beslutet? Även kausalitet har fått en alltmer framträdande roll inom ML under senare år.

Vi bör fortsätta att lära våra studenter att börja en analys med enkla modeller, som ofta fungerar tillräckligt bra. Men vi bör i mycket större utsträckning undervisa om mer komplexa modeller och regularisering, och sluta att föra vidare den överdrivna skräcken för nominell komplexitet som är utbredd bland många statistiker.

3. Programmering, simulering och andra verktyg

En data scientist har ofta en kompetens som företag lätt kan se: de har en vana att arbeta med dataanalys i moderna programmeringsspråk och datorverktyg. Det är svårare för ett företag att identifiera och förstå vikten av en *statistikers* kärnkompetens. Det är viktigt att vi även lär ut programmering och andra verktyg mer ordentligt. Datavetare pratar om *computational thinking* som en problemlösningsprocess med datorn som hjälp. Det handlar inte bara om att lära sig grunderna i t.ex. R, eller andra populära språk för datanalys, som Python, utan ett annat arbetssätt, där datorn spelar en central och helt naturlig roll. Våra kurser borde innehålla inslag av mjukvaruutveckling, olika typer av verktyg för rapportskrivning där kod integreras med text och matematik, versionshanteringssystem, felsökning, kodtestning, prestandaoptimering, länkning till andra programmeringsspråk etc. Dessa inslag kan visserligen undervisas av datavetare, men jag tror att det är viktigt att vi statistiker själva integrerar delar av detta i vår undervisning. Det finns nämligen ett intrikat samspel mellan statistiken och dessa färdigheter, och även vi lärare bör vara moderna statistiker. Vi har nyligen startat månatliga möten vid Stockholms universitet där personalen, inkl. doktorander, presenterar olika typer av datorbaserade verktyg och metoder för varandra.

Jag har många gånger imponerats över hur snabbt studenter lär sig dessa färdigheter när de väl får chansen. Datorer och programmering är naturligtvis inget substitut för matematisk analys utan ett komplement som effektiviserar inlärningen, speciellt om det kombineras med simulering. Det ska falla sig naturligt för studenter att gå hem och skriva en liten programkod som simulerar data från modellen som presenterades vid dagens föreläsning. För många studenter är det här ett utmärkt sätt att förstå sannolikhetsmodeller och deras egenskaper.

Programmeringsvana studenter öppnar också upp för datorbaserad examination. På några av mina kurser får studenterna ta med sig kod från datorlaborationer till tentamen i datasal, där de sedan löser problem genom att kombinera matematiska beräkningar på papper med datorberäkningar i ett modernt programmeringsspråk.

4. Beräkningseffektivitet och stora data

Moderna statistiska problem med stora datamängder och komplexa modeller skattas som regel med iterativa optimerings- eller simuleringsalgoritmer. I många tillämpningar sker analysen också i realtid med s.k. strömmande data som anländer sekventiellt med mycket hög frekvens. Beräkningstiden för att estimera en modell på stora, snabba datamängder är därför ofta helt avgörande för den statistiska analysen. Jag har ofta fascinerats över vilka oväntade aspekter som bestämmer om en beräkning kan genomföras inom rimlig tid. Utvecklingen av modeller och inferensmetoder i min egen forskning har ofta drivits av beräkningsaspekter. När man arbetar med stora komplexa datamängder inser man snabbt vikten av ett samspel mellan statistisk modellering och beräkningar. Datavetare och numeriska matematiker tränas hårt i att förstå komplexitet och exekveringstider för olika algoritmer och numeriska metoder. En del av detta bör också ingå som en naturlig del i statistikundervisningen.

5. Nya datatyper, brus och anomalier

Våra kurser tenderar att presentera data som en prydlig tabell där raderna är observationer på ett antal kolumnvariabler. Men många data kommer allt oftare i avsevärt konstigare former, t ex som högupplösta bilder, datorgenererade loggfiler eller komplexa sensormätningar, eller som text i elektroniska dokument på webben. System som automatgenererar stora mängder data innehåller ofta mycket svårhanterligt brus, en stor andel saknade observationer och outliers. Vi bör därför lära våra studenter hur man samlar in, tvättar och representerar data inför statistisk analys, och hur dessa aspekter påverkar modellval och estimation. Här finns också en viktig roll för försöksplaneringen. Idag samlar företag in data mer eller mindre oplanerat, ofta helt drivet av vilken typ av mätteknik de råkar ha tillgänglig. Resultat blir alltför ofta att de insamlade datamaterialen är helt fel för problemet man vill studera med ML, eller av så dålig kvalitet att de blir statistiskt värdelösa. Här är statistisk kompetens ovärderlig och vi bör träna våra studenter i att identifiera rätt data för rätt problem, och rätt modell för en given datatyp och datakvalitet. Här kommer återigen beräkningseffektivitet in som ett viktigt kriterium vid val av modell och estimationsalgoritm. Allt hänger ihop.

6. Bayes

Bayesiansk inferens spelar en mycket stor roll inom maskininlärning. Många av de mest populära läroböckerna i maskininlärning har ett bayesianskt perspektiv. Bayes ses som det naturliga sättet att hantera prediktion, beslutfattande och regularisering. Många attraheras också av möjligheten att kombinera data med apriorinformation. En stor anledning till att Bayes är så populär i ML-kretsar är att regularisering med fördel kan formuleras som en aprioriförd elning, t.ex. är L1-regularising ekvivalent med en Laplacefördelning som prior. En aprioriförd elning används ofta också för att lägga in subjektiv information om någon slags mjukhet för ett flexibelt parametriserat funktionssamband, som annars riskerar att överanpassa data. Komplexa modeller för intrikata datatyper i ML-tillämpningar är också ofta lättare att hantera bayesianskt via simulering från posteriorfördelningen med t ex Markov Chain Monte Carlo (MCMC) eller genom s k stochastic variational inference speciellt anpassat för stora datamängder. Bayesiansk inferens bör vara en integrerad del av modern statistikundervisning. Jag har själv undervisat en grundkurs i statistisk inferens utifrån ett bayesianskt perspektiv och det går alldeles utmärkt.

AVSLUTANDE ORD

Strålande tider, härliga tider!

Jag har försökt beskriva min syn på maskininlärningsområdet och hur jag tror att det kan fungera som en katalysator för förändring. I några fall handlar det om rätt små förändringar i befintliga kurser, i andra fall om rejäla omtag, men även att helt nya kurser bör ersätta andra förlegade kurser. Det här arbetet är en balansakt där vi måste vara försiktiga med att inte förlora vårt ämnes matematiska grund. Vi måste noga välja vilka delar av dagens kursutbud som ska bort och var vi ska lägga betoningen i undervisningen. Men det är hög tid att ompröva gamla sanningar nu. Kill your darlings!

Jag vill också påpeka något uppenbart: om vi statistiker ska undervisa i maskininlärning och datordriven dataanalys, så bör vi även *forska* inom dessa områden. Jag har under de senaste 2-3 åren börjat se tecken på en positiv förändring här, där ett åtminstone ett fåtal svenska statistiker och probabilister har börjat publicera sig i de ledande ML-forumen. I statistikfältet finns det också flera tidskrifter i computational statistics av hög kvalitet där även ledande ML-forskare publicerar sig.

ML-området är extremt dynamiskt och har genomgått stora förändringar under de senaste 4-5 åren. Neurala nätverk har gjort en spektakulär återkomst, speciellt inom bilder, text och ljud, och det har utvecklats en stor ingenjörsdriven subkultur inom ML för denna typ av ansats. Det kommer alltså bli viktigt att skilja ut verkliga genombrott från tillfälliga trender, men även att tillåta en viss flexibilitet i kursinnehållet.

Min prediktion är att statistikämnet kommer att genomgå stora positiva förändringar under detta nya decennium, delvis som ett resultat av det upplevda hotet från angränsande ämnen som maskininlärning. Det är strålande tider, härliga tider!

MATTIAS VILLANI,
PROFESSOR I STATSTIK VID
STOCKHOLMS OCH LINKÖPINGS UNIVERSITET

(Artikeln ovan finns inom kort även att läsa på https://statfr.blogspot.com/)