



GÖTEBORGS UNIVERSITET

STUDENT

0002-MHH

TENTAMEN

TIG113 Tentamen

Kurskod	--
Bedömningsform	DO
Starttid	19.12.2023 08:30
Sluttid	19.12.2023 12:30
Bedömningsfrist	--
PDF skapad	17.09.2024 15:31
Skapad av	Lyudmila Vafaeva

i Information

OMTENTAMEN

Big data och framtidens beslutsfattande, 15hp

DAG: 19/12 -2023 TID: 08.30 – 12.30

Ansvarig: Faramarz Agahi

Förfrågningar: 031-786 28 22

Resultat: Anslås senast den 8/1 -2024

Betygsgränser: Godkänd 23p. Väl godkänd 33p.
Maximal poängsumma är 40p.

Hjälpmedel: Inga

Skrivningsfrågor består av 8 frågor om sammanlagt 40 poäng. Varje fråga ger maximalt 5 poäng. Du får maximalt poäng på varje fråga om ditt svar får mig att tro att du kan allt om vad uppgiften handlar om. Detta innebär att du behöver svara relativt omfattande så att jag har chansen att göra denna typ av bedömning.

Lycka till!

1 Brus

Diskutera kopplingen mellan följande begrepp inom ramen för dataanalys och bedömning: Bias, Systembrus, Nivåbrus, Mönsterbrus och Situationsbrus. Analysera hur dessa begrepp är sammankopplade och påverkar varandra när du bedömer och utvärderar data. (5p)

Skriv in ditt svar här

Vid bedömningar sker både Bias och brus, speciellt när det gäller människors bedömningar. Vid oönskade fall, sådant som när subjektivitet inte är något man vill ha, så gör bias och brus att precisionen och korrektheten i bedömningar försämras. Bias är en tendens att konsekvent dra sig åt en viss föruppfattade mening (såsom confirmation bias). Vilket gör att du kanske är "off-target". Brus handlar om variabilitet som har inverkan på precisionen hos bedömningar. Brus går att mäta även om man inte vet vad den "korrekta" bedömningen är för att man kan mäta hur mycket bedömningarna varierar från tillfälle till tillfälle, hur stor spridning de har. Systembrus är helheten av hur mycket brus det finns i ett "system", typ en organisation eller grupp etc.

Man brukar dela upp systembruset i 2 delar och de är nivåbrus och mönsterbrus, man beskriver det med en relation som ser ut såhär: $\text{Systembrus} = \text{Nivåbrus}^2 + \text{Mönsterbrus}^2$. Nivåbrus är den skillnaden i variabilitet som finns mellan olika individers snitt i omdöme. Så 2 domare kanske är olika hårda i sina bestraffningar i snitt vilket skapar en variabilitet mellan dem.

Mönsterbrus är variabilitet i en individs egna bedömningar vid samma utgångspunkt. Så 1 domare kanske är hårdare i sin bestraffning mot kvinnor för att de har blivit bedragen av en kvinna eller nått. En man som har begått samma brott kanske då får ett lättare straff jämfört med kvinnan, då visare domaren på variabilitet i sina egna bedömningar. Sedan brukar man dela upp Mönsterbrus i 2 delar och det är Stabilt mönsterbrus vilket i princip är det som beskrevs längre upp, samt Situationsbrus som är de faktorer som har inverkan från tillfälle till tillfälle. Så samma bedömning av en domare kan se olika ut beroende på sådana saker som väder, om deras lag vann en match dagen innan eller om de inte ätit lunch ännu.

Så tillsammans leder alla (bias och brusen olika uppdelningar) dessa beståndsdelar till mindre korrekta, precisa och rättvisa bedömningar.

Ord: 323

Besvarad.

2 Algoritmer

Beskriva de farhågor som tas upp i kurslitteraturen angående användningen av algoritmer i beslutsfattande. Hur kan algoritmer potentiellt förstärka fördomar, och vad är implikationen för demokratin, som framhållits av Cathy O'Neil? Vidare, baserat på forskningsresultat som nämns i Brus-boken, under vilka omständigheter kan algoritmer överträffa mänskliga bedömare i beslutsfattande, och vilka faktorer bidrar till denna framgång? (5p)

Skriv in ditt svar här

Cathy O'Neil tar upp mycket farhågor i sin bok och några av dem är

O'Neil beskriver hur Algoritmer kan gör bedömningar som leder till framtida bedömningar av den sorten blir mer och mer fördomsfulla. Ett exempel på detta skulle vara att en algoritm kanske analyserar ett område och ser att det sker flera arresteringar där vilket kan vara ett tecken på högre brottslighet i det området. Algoritmen anser då att flera poliser bör anställas i det området vilket i sig leder till att fler arresteringar sker i området eftersom att flera poliser kan göra flera arresteringar, så här skapas en ond spiral som leder till att fördomar kring detta område är att det är ett problemområde med mycket brottslighet. Ett annat exempel är hur ranking i amerikanska universitet leder till att de universitet som redan är dåligt rankade har det svårt att höja sin ranking. På grund av att ett ranking system som är väldigt brett vedertaget av många så får ett sådant system väldigt mycket "makt". En negativ spiral som sker här skulle vara att skolor som har lite sämre ranking i systemet får mindre ansökningar och mindre intresse från professorer. Detta leder till att de får mindre elever och mindre pengar samt att deras utbildning kanske tappar kvalite när det inte finns intresse från professorerna. Då leder det till att de får ännu sämre ranking osv. Ranking har kanske här inte något att göra med hur bra utbildning man får på universiteten eller hur bra eleverna mår på det universitetet utan det cirkulerar bara kring hur många studenter som söker sig dit.

Algoritmer kan påverka demokratin genom att till exempel vara mer biased till vissa kandidater i ett val. I en undersökning testade några att tilldela försöksdeltagare olika sökmotorer som hade viss tendens att visa mer av en viss kandidat i ett val vilket ledde till att de använde den sökmotorn var 20% mer benägna att rösta för den kandidaten. Ett annat sätt som en algoritm kan verka på demokratin är att kanske får fler att rösta eller mindre att rösta, till exempel visade facebook att de kunde ställa om sina algoritmer så att de visade mer politiskt innehåll för användare vilket ökade antalet som röstade med 3%.

Människor är ganska dåliga på att hantera data och göra bedömningar som är bias och brusfria, detta är dock algoritmer bättre på. Generellt kan man säga att algoritmer är bättre på det men det betyder inte att de är super bra, man kan säga att människor är ganska dåliga och algoritmer är lite bättre, i allafall när det kommer till prediktiva analyser. Traditionellt har man pratat om "broken leg exception" som en fallgrop för algoritmer och det betyder att algoritmer kanske inte alltid kan avgöra och beräkna vissa saker som att avgöra om någon ska gå på bio ikväll. Algoritmen kan utgå från historisk data och avgöra sannolikheten men om ny data kommer in som säger att personen har brutit benet samma dag så kan en människa avgöra att det inte är så troligt att den personen kommer gå till bion den kvällen. Algoritmen kanske inte tar detta i hänsyn och ger en dålig uppskattning. Nya tekniker och AI har lett till att mer data kan hantera och flera sådana saker kan tas hänsyn till så att algoritmerna är bättre på att hantera sådant. Så generellt kan man säga att om det finns mycket och bra data kring bedömningar så är algoritmer bättre än människor.

Ord: 575

Besvarad.

3 Minska bruset

"Om vårt syfte är att minska bruset står valet mellan två sätt att reglera människors beteenden: regler respektive normer." Diskutera detta påstående genom att definiera regler och normer, ge praktiska exempel på var och en och diskutera vilken strategi som skulle vara mer lämplig för olika typer av organisationer. (5p)

Skriv in ditt svar här

Regler är något som ska följas och om det inte följs så är det inte korrekt. Vilket inte öppnar upp för tolkningsfrågor eller perspektiv. Man kan se det som IF THEN regler.

Normer är saker som bör följas och om de inte följs så är det bara något som är utanför normen och kanske inte rekommenderat. Man kan se detta som riktlinjer för att hur saker ska hanteras. Detta öppnar upp för lite mer kreativitet och frihet.

Om man kollar på en försäkringshandläggare som ska avgöra om en kund ska få ersättning vid en trafikolycka så kanske en regel skulle vara att om kunden inte har stannat vid en stoppskylt strax innan kraschen så får den ingen ersättning. IF didnt stop THEN no money. Normen skulle kanske då vara att inte ge ersättning men tänk om stoppskylten var skymd eller att den andra bilen som kunden kraschade med körde snabbare än hastighetsbegränsningen, då kanske det finns anledning att ge kunden ersättning, i alla fall en del. Med regler skulle kunden i detta exempel aldrig få ersättning medan riktlinjer skulle kunna öppna upp för möjligheten att kunden inte är 100% orsaken till olyckan vilket kanske kan hjälpa kunden att få ersättning.

Regler är ofta bra om man ska definiera saker med hjälp av olika faktorer, till exempel om en läkare ska ställa en diagnos så kanske den tar mätvärden av olika saker och ser att mätverdena uppnår vissa thresholds vilket då är liksom definitionen av en viss diagnos. Riktlinjer är också användbara för läkare då kanske till exempel en patient uppnår alla thresholds för en viss behandlingen vilket enligt regel säger att den bör genomföras, men på grund utav att patienten kanske är för gammal och det inte är värt att behandla för att då kanske man försämrar patientens sista år istället för att hjälpa dem, här är riktlinjer kanske bättre eftersom man kan göra den värderingen och inte genomföra behandlingen för att man är låst vid en regel.

Ord: 328

Besvarad.

4 Modeller

Farliga modeller, särskilt Weapons of Math Destruction, har tre saker gemensamt. Vilka är dessa? Förklara och ge exempel på ditt svar. (5p).

Skriv in ditt svar här

De 3 saker som de har gemensamt är:

Opacity: Detta handlar om hur transparent en modell är. Vilken data använder den? Vilken logik använder den?

Om man de som påverkas av en modells bedömningar inte kan se vad som har lett fram till dessa bedömningar så kan de inte förstå bedömningarna samt att de kanske inte ens vet om att det är något som gör bedömningarna, "You cannot understand what you cannot see".

Dessutom är det svårt för de som utvecklar och jobbar med algoritmerna att förstå hur de kan påverka och förbättra modellen om de inte kan få insyn i modellen och inte vet hur den fungerar.

Scale: Detta handlar om hur mycket en modell kan inverka på flera olika nivåer och platser i samhället. Om en modell kan påverka livet för nästan varje individ i ett samhälle så har den stor scale. Dessutom kanske den kan påverka varje individ flera oberoende gånger under livet eller dagen.

Damage: Detta handlar om hur vilken typ och hur mycket skada som modellen kan åstadkomma. Detta mäts ungefär i fysiska och psykologiska skador.

Ett exempel som visar problematiken skulle vara en modell som har inverkan på ekonomin i ett online multiplayer game. Spelare kan köpa och sälja saker med varandra och på sätt skapas en ekonomi. En modell som till exempel gör bedömningen att den kan skapa ekonomisk tillväxt genom att börja köpa upp alla varor av en viss typ, detta har stor inverkan på ekonomin utan att spelarna förstår varför, helt plötsligt så kan den varan vara värd mycket mer än vad den har varit förut och ekonomin värdet hos andra varor tappar värde, vissa spelare kanske blir ruinerade. Scale blir här stor inom spelet eftersom att alla i spelet blir påverkade. Damage är kanske inte lika mycket eftersom att spelet kanske inte har så stor påverkan i den riktiga världen och det psykologiska och fysiska välmåendet hos spelaren, dessutom kan spelutvecklarna kanske återställa spelet. Om en sådan modell kanske skulle göra liknande saker i verklig ekonomi så kanske skadan blir enorm och folk blir ruinerade, kan inte bo kvar där de bor, köpa mat etc. Enorm fysisk och psykologisk skada.

Ord: 360

Besvarad.

5 Kostnader

Kostnader och fördelar med brusreducering: Förklara vikten av att bedöma om brusdämpande åtgärder är värda sitt pris i detta sammanhang. Ge exempel på kostnader och överväganden som påverkar denna utvärdering. (5p)

Skriv in ditt svar här

Att brusreducera har höga kostnader, det tar tid, det kräver noggrannhet och tydlig struktur. Fördelarna är att man får mer rättvisa bedömningar och bättre stabilitet i organisationer. Först ska man kanske avgöra om brus är något som är bra eller välkommet i de bedömningar som sker, vill man ha lite subjektivitet och variation såsom i smaktester eller recensioner. Om så är fallet så kanske man inte ska försöka brusreducera. Om man ska brusreducera brus som inte är välkommet bör man avgöra om det är värt kostnaden och om det är realistiskt att utföra.

Till exempel skulle jag anse att det är värt att brusreducera i domstolar då det handlar om att avgöra andra människors livsöden. Brus här leder kanske till fängelse eller straffkostnader vid samma fall beroende på olika brusfaktorer (nivå och mönster-brus). Det är 2 ganska olika livsöden som har stor inverkan på livet.

Vid andra tillfällen kanske det är mindre relevant med brusreduktion, till exempel hur en lärare sätter betyg på en elev, om en elev får ett lite orättvist betyg så kanske det inte har så jätte stor inverkan på deras liv (det skulle kunna ha det men inte så ofta).

Hur realistiskt det är att brusreducera är också viktigt att avgöra. Läkare gör väldigt många bedömningar varje dag som vissa av dem kan ha konsekvenser såsom liv och död, så det känns viktigt att reducera brus här men ofta krävs det tid vilket läkarna inte har. De skulle till exempel sätta sig i grupp och diagnostisera med hjälp av något protokoll (typ mediating assesment protocol) där de alla först gör oberoende bedömningar för att sedan diskutera och sedan göra bedömningar igen, här kan informationsflöden regleras både mellan parterna i gruppen men också på ett sekvenserat sätt där mer information ges efter första bedömningen osv. Detta kanske reducerar brus men det tar alldeles för mycket tid och resurser som läkare och annan vårdpersonal inte har idag.

Ord: 315

Besvarad.

6 Datalager

Definiera vad ett datalager (Data Warehouse) är och markera de viktigaste skillnaderna som skiljer det från en vanlig databas. Diskutera egenskaperna hos datalager och förklara varför de är avgörande för specifika analytiska uppgifter. (5p)

Skriv in ditt svar här

Ett datalager är en databas som är gjord för att genomföra analys på stora mängder data genom OLAP(online analytical process). En vanlig databas tar data direkt från källan och uppdateras kontinuerligt genom OLTP(online transactional process) så den har real-time data, där den sparar datan med hög upplösning och hög detalj i tabeller(rader och kolumner), strukturen i en databas är lite mindre rigid och kan vara mer flexibel, man kan ändra den vid vissa behov. Ett datalager tar data från databaser genom en process kallad ETL(Extract Transform Load) där datan extraheras från databaserna ofta genom att den summeras och "städas upp"(typ bortser från data som fattas, är inkorrekt eller outliers), transformeras till rätt format för och struktur som passar datalagret, och laddas in i datalagret i den strukturen som datalagret har. Datalagret får då inte datan direkt från källan utan från databaser och är har då inte real-time data som är kontinuerligt uppdaterad. Datalagret innehåller stora mängder historisk data som inte skrivs över. Det som gör att datalagret är bra för analys är att det har en rigid struktur med ofta summerad data vilket gör att querying går mycket snabbare vid hantering av stora mängder data än i en databas, dessutom är navigering i en sådan struktur mycket enklare. Detta leder till att analys av datan är mycket mer effektivt och lättare.

Ord: 222

Besvarad.

7 stjärnschema

Beskriv syftet med ett stjärnschema och förklara varför det används i samband med datalagring. Definiera vad en dimensionstabell är och dess roll i ett stjärnschema. Hur förbättrar denna schemastruktur effektiviteten i dataanalys? (5p)

Skriv in ditt svar här

Ett star schema är en modell för hur data i en relational database struktureras. Det kallas för "star schema" för att det finns 2 olika typer av tabeller; Fact table och Dimensional table, där ett fact table är omringad av dimensional tables vilket kan likna en stjärna.

Olika dimensional tables innehåller olika entiteter så som "kunder", "anställda", "produkter" osv. Där varje rad står för en till exempel en kund. Den kunden får en unik kod(key) och i andra kolumner så finns attribut som är information om kunden såsom "adress" "email" eller "telefonnr".

Ett fact table är gjort för att spåra till händelser så som "sales", "anställningar" eller "leveranser". I ett fact table så kan det finnas både measures som "datum" eller "vikt" men också de unika koder som refererar till unika entiteter i dimensional tables. Så om man har till exempel ett fact table som heter "leveranser" så kanske en rad innehåller measures "vikt", vikten för leveransen, skrivet i en siffra, sen kanske nästa kolumn är "kund" vilket då betyder att det står en unik kod som tillhör den kunden som leveransen ska till och koden refererar då till kunden med samma kod i "kunder" dimensional table. Detta betyder att i ett fact table så finns informationen för flera leveranser, vilket skulle kunna innehålla flera leveranser till samma kund. Om man inte skulle ha star schema struktur så skulle kundens information(adress, email osv.) behöva finnas på flera rader i tabellen vilket skapar redundans. Med star schema behöver saker som kundens information bara stå på 1 ställen 1 gång för att sedan kunna refereras till genom den unika koden.

Med denna struktur går querying mycket snabbare och navigeringen i en sådana struktur är mycket lättare, om till exempel letar efter alla kunder så kan man lätt navigera till det dimensional table och få den informationen summerad utan redundans, så det är lätt att få översikt och göra operationer på.

Ord: 318

Besvarad.

8 Relation och samband

Kursen diskuterade tre olika kategorier av dataanalysverktyg, nämligen Deskriptiva, Prediktiva och Preskriptiva analysverktyg, samt två grundläggande typer av relation och samband: Kausalitet och Korrelation. Beskriv och ge exempel på vilka dataanalysverktyg som mest sannolikt genererar kausala relationer och/eller som är mer lämpade för att identifiera korrelations samband. (5p)

Skriv in ditt svar här

Om vi ska definiera de olika analys processerna så är:

Deskriptiva: när man utgår från historisk data för att tala om hur något har varit är just nu. What has happened? Deskriptiva datanalysverktyg som har använts i kursen har till exempel varit Tableau eller data warehousing. De utgår från historisk data och säger hur något är i data såsom siffror. Detta är högst korrelations relaterat eftersom att man kan beskriva att till exempel 2 variabler håller samman, man kan se att stolar säljer bra i en region, detta beskriver inte varför det är så i en kausal relation.

Prediktiva: när man utgår från historisk data för att göra förutsägelser om framtiden. What can happen?

Prediktiva dataanalysverktyg som har använts i kursen har till exempel varit WEKA. Här används machine learning och data mining för att med hjälp av historisk data ta fram förutsägelser om framtida händelser såsom sannolikheten att jag kommer spela golf i helgen, med faktorer såsom väder, vind, årstid eller arbetsbörda så kan man avgöra om jag kommer spela eller inte, om väder = sol, årstid = sommar, vind = lite och arbetsbörda = låg så är kanske sannolikheten ganska stor att jag kommer spela. Denna typen av analys kanske kan anses som lite mer kausal eftersom man kan bryta ut nyckel faktorer som leder till hög sannolikhet till spela golf. Väder = sol -> spela golf. Dock byggs dessa analyser upp med hjälp av historisk data som har korreleration relation.

Preskriptiva: är när man är normativ och beskriver med hjälp av historisk data hur något ska vara. How should it be?

Preskriptiva dataanalysverktyg som har använts i kursen är till exempel decisions support systems(DSS eller InstantTea) som ger förslag på hur problem ska hanteras. Dessa kan fungera med hjälp av IF THEN regler som kanske säger hur något ska hanteras beroende på en eller flera faktorer. Detta är också mer av en kausal relation, men även dessa typer av regler skrivs med hjälp av historisk data som någonstans grundar sig i korrelationer mellan olika faktorer.

Så slutsatsen är väl att det kan gå att hitta kausala relationer men att de ofta börjar med korrelerande faktorer som man sedan kan kanske kan härleda och undersöka för att hitta kausala relationer. På nått sätt är det ju så att korrelationer inte nödvändigtvis innebär kausalitet men kausalitet visar alltid upp korrelation.

Ord: 390

Besvarad.