TEORIFRÅGOR

GRUPPUTVÄRDERING

1. Vem du har arbetat i grupp med?

Kamila Nigmatullina, Kicki Nocoj Bici, Quantri Tran, Matthew Motallebipour, Umut Arslan och Leonardo Sjöberg.

2. Hur har ni i gruppen arbetat tillsammans?

Varje medlem har fått samla ihop cirka 44 obeservationer. Vi har diskuterat och pratat och kommit med enklare insamling av fakta och vi har samarbetat samt varit tillgängliga för varandra.

3. Vad var bra i grupparbetet och vad kan utvecklas?

Det som var bra är att alla gruppmedlemmar varit aktiva och gjort sin del.

Det som hade kunnat utvecklas är att man involverar sig själv lite mer genom att alla diskuterar tillsammans istället för att dela upp oss osv.

4. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp?

Min styrka är att jag lyssnar och verkligen försöker göra min del i tid så att jag inte påverkar en hel grupp och att jag visar en trygghet gentemot mina kamrater.

Utvecklingsmöjligheterna är att jag försöker ta till mig och lära mig från andra och förstå hur kamrater tänker för att bygga en bättre förståelse till uppgifter och mål.

5. Finns det något du hade gjort annorlunda? Vad i sådana fall?

Lagt ner mer tid åt att sköta mina kurser och skapat mer tid för kurserna.

TEORIFRÅGOR

1. Kolla på följande video: https://www.youtube.com/watch?v=X9_ISJ0YpGw&t=290s , beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.

SVAR:

En QQ-plot är en graf som används för att jämföra två sannolikhetsfördelningar. Man ritar ut punkter på grafen där varje punkt representerar en kvantil från den ena fördelningen och motsvarande kvantil från den andra fördelningen.

Om punkterna ligger längs en rak linje, betyder det att de två fördelningarna är lika eller passar bra ihop. Ju närmare punkterna ligger den raka linjen y = x, desto bättre stämmer de överens. Det hjälper att se om din data följer en viss fördelning, till exempel en normalfördelning.

Din kollega Karin frågar dig följande:

”Jag har hört att i Maskininlärning så är fokus på

prediktioner medan man i statistisk regressionsanalys kan göra såväl prediktioner som

statistisk inferens. Vad menas med det, kan du ge några exempel?”

Vad svarar du Karin?

SVAR:

Tjena Karin!

Du har rätt i att maskininlärning ofta fokuserar på att göra prediktioner. Maskininlärningsalgoritmer används för att förutsäga framtida händelser baserat på historisk data. Exempelvis kan vi använda maskininlärning för att förutsäga väder, aktiekurser eller huspriser. Statistisk regressionsanalys å andra sidan har två huvudsakliga mål: prediktion och statistisk inferens.

Prediktion

- **Exempel:** "Vi kan förutse nästa års elförbrukning i en stad baserat på tidigare års förbrukningsdata och väderprognoser."

- **Förklaring:** Här används historiska data och framtida väderprognoser för att förutsäga hur mycket el en stad kommer att använda nästa år. Fokus ligger på att göra en förutsägelse om framtida utfall.

Statistisk inferens

- **Exempel:** "Vi kan undersöka om ett nytt läkemedel minskar kolesterolnivåerna jämfört med ett placebo och testa hypotesen om att läkemedlet har en signifikant effekt."
- **Förklaring:** Här används data från en klinisk studie för att avgöra om det nya läkemedlet har en statistiskt signifikant effekt på kolesterolnivåer jämfört med placebo. Målet är att dra slutsatser om läkemedlets effekt i den bredare populationen.

3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?

SVAR:

Enklaste sättet att förklara skillnaden mellan konfidensintervall och prektionsintervall är att skapa en typ av skriftlig tankekarta. Först och främst förklarar jag vad dessa två intervaller spelar för roll och sedan presenterar jag skillnaden mellan dem för predikterade värden.

Konfidensintervall

Ett konfidensintervall är ett intervall som med en viss sannolikhet, ofta 95%, innehåller det sanna värdet av en parameter (t.ex. ett medelvärde) baserat på data från ett stickprov. Det används för att uttrycka osäkerheten i en statistisk uppskattning.

Prediktionsintervall

Ett prediktionsintervall är ett intervall som med en viss sannolikhet förutspår var en enskild framtida observation kommer att ligga, baserat på en modell eller tidigare data. Det tar hänsyn till både osäkerheten i den uppskattade modellen och variationen i framtida data.

Skillnaden mellan konfidens –och prediktionsintervall för predikterade värden

Skillnaden mellan konfidensintervall och prediktionsintervall är:

- **Konfidensintervall**: Anger osäkerheten i det genomsnittliga (predikterade) värdet för populationen. Det är snävare eftersom det bara tar hänsyn till osäkerheten i uppskattningen av medelvärdet.
- **Prediktionsintervall**: Anger osäkerheten för en enskild framtida observation. Det är bredare eftersom det inkluderar både osäkerheten i uppskattningen och den naturliga variationen i enskilda data.

4. Den multipla linjära regressionsmodellen kan skrivas som:

$Y = \beta 0 + \beta 1x1 + \beta 1x2 + \ldots + \beta pxp + \varepsilon$.

Hur tolkas beta parametrarna?

SVAR:

β0: Baslinjenivån för Y när alla oberoende variabler är noll. intercepten

B1, B2….βp: Förändringen i Y för varje enhetsändring i xp, medan andra variabler hålls konstanta.

Regressionskoefficienter

X1, x2… xp: dessa är oberoende variabler/prediktorer som förklara variationen i Y.

ε: Slumpmässig variation som inte kan förklaras av de oberoende variablerna. Representerar residuals.

5. Din kollega Hassan frågar dig följande: ”Stämmer det att man i statistisk regressionsmodellering inte behöver använda träning, validering och test set om man nyttjar mått såsom BIC? Vad är logiken bakom detta?” Vad svarar du Hassan?

SVAR:

Hej Hassan,

Ja, det stämmer delvis. BIC hjälper till att välja en modell baserat på hela datamängden genom att balansera modellens passform och komplexitet. Men utan träning, validering och testset riskerar du att inte upptäcka om modellen överanpassar och därmed inte generaliserar bra till nya data.

Logiken bakom BIC (Bayesian Information Criterion) är att balansera modellens passform mot dess komplexitet. BIC straffar mer komplexa modeller (med fler parametrar) för att förhindra överanpassning, vilket gör att den modell som bäst förklarar data med färre parametrar föredras.

6. Förklara algoritmen nedan för ”Best subset selection”

SVAR:

Nullmodell

Startar med en modell utan prediktorer (M_0), som bara gissar medelvärdet för alla observationer.

Modellprövning

Testar alla möjliga modeller med olika antal prediktorer (från 1 upp till k). Utvärderar varje modell baserat på hur bra den passar data, dvs. hur nära den predicerade resultaten ligger de verkliga.

Val av bästa modell

Väljer den modell som presterar bäst enligt ett valfritt kriterium (som BIC, AIC, R^2 eller cross-validation). Målet är att hitta den optimala uppsättningen av prediktorer för att maximera modellens prestanda.

7. Ett citat från statistikern George Box är: “All models are wrong, some are useful.”

Förklara vad som menas med det citatet.

SVAR:

Det betyder att alla statistiska och matematiska modeller är förenklingar av verkligheten och därmed inte kan fånga alla detaljer eller nyanser av den verkliga världen.

Modeller är baserade på antaganden och approximationer som inte är helt korrekta, men de kan fortfarande vara användbara om de ger insikter, förutsägelser eller beslut som är tillräckligt bra för praktiska syften.

Jag tror att man även kan dra ett exempel som att förutsäga vad elkostander blir kommande år med hjälp av data från nuvarande år. Det ger förutsägelser som är tillräckligt bra men mycket kan flrändras inför kommande år och det kan leda till stora eller mindre förändringar för hur elkostanderna kommer se ut.