# Teoretiska frågor

Besvara följande teoretiska 7 frågor:

1. Kolla på följande video: https://www.youtube.com/watch?v=X9\_ISJ0YpGw&t=290s , beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.

En QQ plot är en plot som visar hur mycket datan är normalfördelad.

1. 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?

Jag svarar Karin att maskininlärning är mer fokuserat på själva prediktionen men den ger ingen förklaring till den, medans statistisk regressionsanalys kan man få både prediktionen och de underliggande faktorer som påverkar den. Ett exempel på det kan vara att maskininlärning kan säga vad en person kan tjäna, medans i statistiks regressionsanalys kan säga vad en person kan tjäna men också vilka faktorer som gör att personen kan få den lönen.

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

Ett konfidensintervall säger att till exempel med 95% säkerhet täcker intervallet det fixa med okända värdet och används ofta till att uppskatta genomsnittliga värden. Prediktionsintervall gör detsamma men den lägger till en fel term så intervallet blir mycket bredare. Den gör det för att individuella prediktioner varierar väldigt mycket mer än till exempel ett medelvärde.

1. Den multipla linjära regressionsmodellen kan skrivas som: 𝑌 = 𝛽0 + 𝛽1𝑥1 + 𝛽1𝑥2+ . . . + 𝛽𝑝𝑥𝑝 + 𝜀 . Hur tolkas beta parametrarna?

𝛽0 tolkas som interceptet. 𝛽1𝑥1 + 𝛽1𝑥2+ . . . + 𝛽𝑝𝑥𝑝 tolkas som de separata slope för de olika oberoende variablerna.

1. 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?

Logiken bakom det är att mått så som BIC och Adjusted R^2 kan tas fram med bara tränings datan. Jag svara Hassan, att även fast det går att bara använda tränings data, så är det att föredra att använda träning, validering och testdata då man kan använda fler utvärdering mått och testa modellen på helt osedda data.

1. Förklara algoritmen nedan för ”Best subset selection”

A white paper with black text

Description automatically generated

Best subset selection testar alla möjliga kombinationer av variabler för att hitta den bästa modellen. Först så använder den bara 1 variabel per modell, modellen med lägst RSS eller högst R^2 väljs. Sen använder den alla möjliga kombinationer för 2 variabler per modell och väljer den modellen med lägst RSS eller högst R^2. Och så gör den så för alla möjliga variabel kombination och väljer den bästa modellen för varje variabel mängd. När vi har alla de bästa modellerna så väljs den allra bästa modellen genom att använda prediktions felet på ett validerings set, Cp (AIC), BIC, eller adjusted R2. Eller genom korsvaliderings metod.

1. Ett citat från statistikern George Box är: “All models are wrong, some are useful.” Förklara vad som menas med det citatet.

Det som menas är att ingen modell kommer vara helt rätt, men vissa är tillräckligt bra för att vara användbara.

# Datainsamling

1. Vem du har arbetat i grupp med?

I min grupp var vi, Jag, Robert, Ian, Filip, Magdalena och David

1. Hur har ni i gruppen arbetat tillsammans?

Vi började med att planera vilken data, hur mycket data och hur vi skulle samla in datan. Efter det så samlade all in datan och la in det i en gemensamt excel fil. Därefter så rengjorde och fixade vi så att datan var användbar och såg likadan ut. Efter det så var datan klar för användning.

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

Jag tyckte det mesta gick bra, något som skulle kunna utvecklas, om man ska vara riktigt petig är kanske sätta lite mer exakt regler i början och komma igång. Men jag är väldigt nöjd allmänt.

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

Mina styrkor är att jag har ganska enkelt för att jobba i grupp och kommunicera med alla. Utvecklingsmöjligheter kan vara att kommunikationerna mellan mig och gruppmedlemmar kan bli ytligare lite tydligare så man verkligen förstår vad båda tänker.

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

Kanske att komma igång ännu lite snabbare, men jag är väldigt nöjd med gruppens arbete.

# Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

En utmaning jag hade var att jag inte hade någon tidigare kunskap i R. Sättet jag hanterade det på var att bara lägga ner mycket tid på det och ta det steg för steg, istället för att försöka lära sig allt direkt.

1. Vilket betyg du anser att du skall ha och varför.

Jag anser att jag skall ha VG då jag gjort VG delen och att mitt arbete är gjort på ett fördjupat sätt och med hög säkerhet. Och att jag har redogjort för och kritiskt diskuterat modellval, modellanpassning och modellutvärdering.

1. Något du vill lyfta fram till Antonio?

Jag tycker att denna kunskapskontroll var svårare än den förra, vilket den ska vara. Men den var inte för svår heller. Så bra svårighetsgrad och rolig uppgift!