Grupputvärdering

1. Vem du har arbetat i grupp med?

   Ali Hamza, Kamila Nigmatullina, Kicki Nocoj Bici, Leonardo Sjöberg, Matthew Motallebipour, Umut Arslan.

2. Hur har ni i gruppen arbetat tillsammans?
   Diskuterat och stämt av vad/hur vi ska samla data och modellens syfte. Varje medlem samla ungefär 44 observationer var.

3. Vad var bra i grupparbetet och vad kan utvecklas?
   Kamera är på när man pratar.
   Behövs mer grupparbete kanske?

4. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp?
   Min styrka är att alltid bekräfta med varandra så alla hamna på rätt spår, det är trist att fortsätta vidare när någon inte hänger med. Bättre att stanna en stund och räcker en hand.
   Utvecklingsmöjligheter att arbeta i grupp är att inlärningsförmågan blir mycket bättre, för jag känner jag lär mig bättre av att lyssna, prata, läser och skriver i grupparbetet.

5. Finns det något du hade gjort annorlunda? Vad i sådana fall?
   Inget jag kan tänka mig just nu.

TEORI 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.

SVAR:

QQ-plot ä ren grafisk metod att jämföra två sannolikhetsfördelningar genom att plotta deras kvantiler mot varandra. Plotten visar ”äkta” data och ”förväntade” data, som visa hur väl de två matchar varandra genom att placera punkter på en graf. Om de ”äkta” data punkterna ligger längs en rak linjer betyder det att datan passar den förväntade fördelningen. Ju nämre desto närmre blir y = x och desto bättre matchning.

2. 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:

Hej Karin,

det stämmer att i Maskininlärning att fokuset ofta är på att göra prediktioner. Inom Maskininlärningen används algoritmer för att förutspår framtida händelser baserat på den inmatade historiska data. Exempelvis kan vi använda maskininlärning till att förutspå väderprognoser, aktiekurser och/eller huspriser.

Statistisk regressionsanalys däremot har både prediktion och statistisk inferens och har två huvudsakliga mål.

1. Att utföra prediktioner genom regressionsmodeller för att predikera en variabel utifrån en eller flera förklarande variabler.
2. Att använda statistisk inferens för att förstå sambandet mellan variabler och dra slutsats från populationen. Här testa vi hypoteser om parametrar, till exempel veckans materbjudande påverkar försäljningen, och beräkna konfidensintervall för just den parametern.

Några fler exempel:

För prediktioner, kan det användas för att förutsäga nästa veckas matförsäljning baserat på tidigare försäljningar och taktiska marknadsföring (ex. reklamutgift, rabatter, erbjudande).

För statistisk inferens, så testa vi om marknadsförings förslaget har någon signifikant effekt på försäljningen och beräkna därefter ett konfidensintervall för relevant parameter.

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

SVAR:

Konfidensintervall – där det sanna, okända värdet förmodligen finns.

Prediktionsintervall – mer komplex än Konfidensintervall. Förutsäger där framtida mätningar kan förmodas hamna.

Med andra ord, konfidensintervall talar om var det förväntade värdet troligen förmodas hamna medan prediktionsintervallet talar om var framtida observerat värde förmodas hamna.

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....β_p: Förändringen i Y för varje enhetsändring i x_p, medan andra variabler hålls konstanta. Regressionskoefficienter

X1, x2... x_p: 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,

vi kan börja med att förklara BIC (Bayesian Information Criterion) som är ett informationskriterier verktyg och mått som används för att jämföra olika statistiska modeller. Logiken bakom BIC är att straffar modeller som är för komplexa och prioritera hellre en modell som ger god anpassning till data samtidigt som modellen har färre parametrar. Detta för att förhindra överanpassning som är ett vanligt problem i statistisk modellering.

Svaret på din fråga om träning, validering och test set behövs är; Ja, de olika data sets behövs för att testa på om modellen fungera på verkliga data. Eftersom BIC ge indikation på modellens komplexitet och anpassning, därvid är BIC oberoende av hur vi delar upp våra data sets. Genom att använda BIC ersätter vi inte träning, validering och test set, men vi kan använda BIC som ett komplement.

6. Förklara algoritmen nedan för ”Best subset selection”

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using using the prediction error on a validation set, C_p (AIC), BIC, or adjusted R^2. Or use the cross-validation method.

SVAR:

Algoritmen är för ”Best subset selection” och förklaras som följande;

1. M_0 som är nullmodellen innehåller ingen information om prediktorer, prediktorer är de variabler som vi tror påverkar resultatet. M_0 kommer därmed bara gissar medelvärdet på alla observationer.

2. Alla möjliga modeller kommer prövas och lägger in k prediktorer i varje modell och sedan utvärderas varje modell. T.ex. om vi har 5 prediktorer så testar vi modellerna med 1 prediktor, sen 2 prediktor…k prediktorer och så vidare. Modellernas utvärdering är baserat på hur bra den passar data, i laymans term hur nära de verkliga resultaten den kommer till.

3. Algoritmen kommer slutligen välja den bästa modellen bland alla modeller med k prediktorer. Modellen som väljs är den som presterar bäst enligt valda kriterium, det finns olika kriteriumsmetoder såsom BIC, AIC, R^2 eller genom cross-validation metoden.

Målet är att hitta den bästa delmängden av prediktorvariabler att inkludera i en modell baserat på ett specifikt kriterium, såsom att minimera prediktionsfel eller maximera modellens anpassning.

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 oavsett hur bra eller komplex en statistisk modell är så kommer det endast speglas som en förenkling av den verkliga världen. Det betyder att ingen modell egentligen kan fånga upp alla faktorer som finns i verkliga världen. Detta är så eftersom modeller är byggt på approximationer och antaganden och är därmed aldrig eller sällan exakta.

Trots att alla modeller är en förenkling av den verkliga världen och ej kan spegla fullständig resultat, så finns det däremot vissa användbara vid rätta omständigheter. Eftersom vi är medvetna om modellernas begränsningar och är noga med att använda dem med försiktighet, vet vi att, genom att använda korrekt modell till relevant fenomen kan vi ändå få värdefull insikt för att hjälpa oss lösa praktiska problem.

Därmed även om dem är felaktiga så är dem användbara verktyg och metoder för att förstå och hantera komplexa problem.