Løsningsforslag til prosjektoppgaven i ISTx1003

Mette Langaas IMF/NTNU

20 December, 2020

Oppgave 1: Regresjon

Q1.1

a) Skriv ned ligningen for den estimerte regresjonsmodellen. Forklar de ulike elementene.

Estimert regresjonsmodell:

$$\hat{y}_i = 1.0699 + 0.0199 x_{1i}$$

der x_{1i} er høyde for person i og $\hat{y}_i$ er predikert verdi for person i. De to tallene er estimatene for skjæringspunkt (også kalt konstantledd) $\hat{\beta}_0 = 1.0699$ og for stigningstall $\hat{\beta}_1 = 0.0199$.

Gale svar: ikke vite forskjell på kovariaten x og estimerte koeffisienter $\hat{\beta}$. Det skal være en hatt på responsen y siden det er en prediksjon, men det skal ikke være en hatt på kovariaten x. Det er ikke med feilledd e_i ved prediksjon.

Galt: vår x heter kovariat (ikke kovarians), eller prediktor (ikke predikat) eller forklaringsvariabel.

Galt: $Y = \hat{\beta}_0 + \hat{\beta}_1 x_i + e$: ser du alle feilene? Hvis det er Y til venstre har vi de sanne koeffisientene β_0 og β_1, og husk indeks i på feilleddet e. Det riktige er $Y_i = \beta_0 + \beta_1 x_{1i} + e_i$ - hvis det er den sanne modellen vi snakker om og så er det riktig med $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ - eller eventuelt med indeks i $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$ når vi snakker om prediksjonsmodellen.

b) Hvordan vil du tolke den estimerte verdien til skjæringpunktet (Intercept) $\hat{\beta}_0$?

Her er $\hat{\beta}_0 = 1.0699$, som er verdien til antall røde blodceller (ganger 10^{12} pr liter) for en person som er 0 cm høy. Det er ikke meningen at det skal tolkes, men det er en konsekvens av linjen vi har tilpasset for verdiene av høyden vi har observert. Modellen er bare gyldig i det intervallet vi har høydeobservasjoner.

Q1.2

a) Vi ser at for 'Hoeyde' er 'coef' lik 0.0199. Hva er formelen som er brukt for å regne ut denne verdien? Hvordan vil du forklare dette tallet til en medstudent som ikke har hørt om enkel lineær regresjon?

La x_{1i} være høyde til utøver nummer i og y_i være antall blodceller til utøver i. Videre er $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ gjennomsnittlig antall blodceller for utøverne og $\bar{x_1} = \frac{1}{n}\sum_{i=1}^{n} x_{1i}$ gjennomsnittlig høyde for utøverne. Da er formelen for estimert stigningstall $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$.

Estimert verdi for dette tallet er 0.0199. Hvis vi ser på to utøvere A og B, og utøveren B er 1 cm høyere enn utøver A. Da vil vi i gjennomsnitt vi vente at utøver B har 0.0199×10^{12} flere blodceller per liter enn utøver A.

Galt svar: Det er mange som ikke har forklart dette noe godt, bare sagt noe ala "det forklarer forholdet mellom høyde og antall blodceller". Det er for vagt.

b) For 'Hoeyde' er det også gitt de to tallene 0.013 og 0.027 under kolonnene "[0.025 0.975]". Hva er disse to tallene og hvordan tolker du tallene?

De to tallene er nedre og øvre grense i et 95% konfidensintervall for stigningstallet. Vi har stor tiltro til at det sanne (ukjente) underliggende stigningstallet vil ligge i dette intervallet. Intervallet hensyn til hvor sikker vi er på verdien vi har estimert for stigningstallet, og blir smalere hvis vi har mange observasjoner.

Galt: Det er ikke et prediksjonsintervall og vi kan ikke si at vi er sikre på at en ny observasjon vil havne i intervallet.

Galt: Det er ikke slik at $\hat{\beta}$ kan være utenfor intervallet - det der jo det vi bruker som midtpunkt for intervallet. Det er da heller ikke slik at "det er 95% sannsynlighet for at den estimerte koeffisienten ligger i intervallet" - vi ser jo at den estimerte koeffisienten alltid er i intervallet! Intervallet er for den sanne ukjente koeffisienten!

Galt: Det er ikke slik at "det er 95% sannsynlig at estimert regresjonskoeffisient ligger mellom øvre og nedre grense i dette intervallet hvis man gjentar forsøket" - det er en andel i det lange løp og det gjelder ikke estimert regresjonskoeffisient, men virkelig sann ukjent regresjonskoeffisient når det lages nye grenser i det nye datasettet.

c) Videre står det for 'Hoeyde' at 'P>|t|' er 0.000. Hvilken hypotese har man testet her? Hva er konklusjonen fra hypotesetesten hvis vi bruker signifikansnivå 0.05? Hvordan henger dette sammen tallene 0.013 og 0.027 fra forrige punkt?

H_0 : Stigningstallet er 0, det er ingen (lineær) sammenheng mellom høyde og antall blodceller for utøverne, og H_1: det er en (lineær) sammenheng.

$H_0 : \beta_1 = 0$ mot $H_1 : \beta_1 \neq 0$.

Vi forkaster nullhypotesen at det ikke er en lineær sammenheng og konkluderer med at det er sammenheng mellom høyde og antall røde blodceller. Dette passer sammen med at 95% konfidensintervall ikke dekker 0.

Galt: noen skriver at hypotesetesten er for $\hat{\beta}_1$ - det er den ikke - den er for den ukjente parameteren β_1.

Galt: 'P>|t|' er p-verdien, og den er derfor ikke "sannsynligheten for at p-verdien er større enn absoluttverdien til t-observatoren".

Galt: vi ser ikke etter om p-verdien er inne i konfidensintervall intervallet - vi ser etter om parameterverdien for nullhypotesen er inne i intervallet - og er det det så vil p-verdien være større enn 5% for et 95% konfidensintevall og større enn 10% for et 90% konfidensintervall osv.

Tilleggsinformasjon: P-verdien for testen er dessverre i Python oppgitt om 0. Det betyr at den er liten og mindre enn presisjonen på det Python skriver ut, og regner vi ut den selv ved å bruke at testobservatoren er t-fordelt med 103 frihetsgrader får vi tallet $6.6 \cdot 10^{-8}$. (Dette visste vi ville være lite.).

Q1.3

a) Hvilke modellantagelser gjør vi i en enkel lineær regresjon?

- Observasjonsparene $(x_{11}, y_1), (x_{12}, y_2), \ldots, (x_{1n}, y_n)$ er uavhengige av hverandre. Men y_1 skal være avhengig av x_{11} selvfølgelig, det er observasjonene til de n utøverne som er uavhengige.
- Det er en lineær sammenheng mellom forklaringsvariabel x_1 og respons y.
- Feilledd er normalfordelt med forventningsverdi lik 0 og konstant varians σ^2 (samme for hver verdi forklaringsvariablen kan anta).

b) Hva er en predikert verdi og hva er et residual (formler)?

En predikert verdi er $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$, vårt beste gjett på hva y_i er gitt at vi vet hva x_{1i} er og bruker den tilpassede regresjonsmodellen.

Residualet er forskjellen mellom observert og predikert verdi: $\hat{e}_i = y_i - \hat{y}_i$.

Galt: Det heter ikke predikat, men predikert verdi.
Galt: For predikert verdi er det hatt på både y og de to βene - vi har estimert dem og bruker dette til å predikere.
Merk: Det er forskjell på feilleddet, e_i, og residualet, $\hat{e}_i$, og det er bare residualet vi kan regne ut.

c) Hvordan kan vi bruke predikert verdi og residual til å sjekke modellantagelsene?

Det er mulig å sjekke modellantagelsene ved å bruke de predikerte verdiene, sammen med residualene som vårt beste gjett for feilleddene, og derfor sjekker vi:

- Er det noen trend hvis vi plotter residualer mot predikert verdi? Hvis det er det så er det noe vi ikke har forklart godt med vår regresjonsmodell.
- Vi ser også på bredden av plottet, er den konstant eller kanskje økende? Det er kun hvis bredden er konstant at vi kan tro på at variansen til feilleddene ikke er avhengig av verdien til kovariaten.
- Vi sjekker også om residualene er normalfordelte i et såkalt QQ-plott.

d) Vi får også oppgitt tallet "R-Squared" til å være 0.238 (ofte også skrevet som 23.8%). R^2 har i enkel lineær regresjon en sammenheng med korrelasjonskoeffienten, men det er en annen definisjon som er relatert til sum av kvadrerte residualer. Hvilken formel er det? Forklar alle symboler. Hvordan vil du forklare tallet til en medelev som ikke har hørt om enkel lineær regresjon?

$$R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}.$$

Her er y_i respons for observasjon i og $\hat{y}_i$ predikert verdi for observasjon i. Videre er n antall observasjoner. SST kalles sums of squares of total og SSE sums of squares of error. Det er totalt SST variabilitet i data, og vi forklarer ikke SSE av variabliteten, dermed forklarer vi SST-SSE. Andel får vi ved å dele på SST.

R^2 er andelen forklart variasjon i den tilpassede modellen. Det sier noe om hvor stor variasjon vi ser rundt en rett linje, i forhold til variasjon rundt en linje med stigningstall 0.

Galt: her er det flere som forklarer om SSE og ikke om R^2. SSE er bare en del av R^2.

Q1.4

a) Studer plottet av predikert verdi mot residual. Hvordan skal et slikt plott se ut hvis modellantagelsene er oppfylt? Hvordan vil du evaluere plottet?

I plottet av predikert verdi mot residual skal det ikke være noen trend, og variasjonen skal være rimelig konstant. Er det slik vil vi ha funnet en god modell og variansen til feilleddene er trolig konstant (varierer ikke med verdien til 'Hoeyde'). Plottet er greit nok, ganske tilfeldig og ikke noen klar trend når det kommer til bredden til punktene.

b) Studer QQ-plottet av residualene. Hvordan vil du evaluere plottet?

QQ-plottet ser fint ut, og punktene følger den røde linjen. Det betyr at residualene ser rimelig normalfordelte ut. Dermed tror vi at det er en pekepinne for at feilleddene også er normalfordelte - som modellen antar.

Ekstra (utenfor pensum, bare til informasjon): Det støttes også av de to hypotesetestene med H_0: residualene kommer fra en normalfordeling, mot H_1 : det gjør de ikke. Her beholdes nullhypotesen (på alle nivå under 0.3) for begge testene.

c) Vil du konkludere med at modellen passer godt?

Jeg vil konkludere med at modelltilpasningen er god, men at det ikke er noe å juble for at man forklarer 24% av variablitet i data. Det betyr at vi bør se etter andre forklaringsvariabler.

Q1.5

Oppsummer kort hva du ser i plottene. Fokus skal være om du tror at det er noen sammenheng mellom Blodceller (som respons) og de fire mulige forklaringsvariablene (Hoeyde, Vekt, Kjoenn og Sport). Hvilket Kjoenn har generelt høyest verdi for Blodceller?

Antall blodceller mot

- Hoeyde: har vi diskutert før, og det ser ut som det kan være en lineær trend
- Vekt: her ser det ut som det kan være en trend, men ikke så klar for for Hoeyde (hadde vi tilpasset en enkel lineær regresjon ville R^2 vært lavere enn for Hoeyde)
- Kjønn: har bare to nivå så det må være en lineær trend
- Idrett: her kan vi bruke dummyvariabelkoden, så helt fint at det ikke er en lineær trend - men det ser ut til at basketball kan skille seg ut?

I tillegg ser vi at Hoyde og Vekt er høyt korrelerte, men vi har ikke lært om multikollineariet så det legger vi ikke vekt på. Men noen kan filosofere om at kanskje det er vanskelig å skille effekten av Hoeyde fra effekten av Vekt på antallet blodceller.

Det er kjoenn=0 som har flest blodcelle - det er oppgitt at det er menn. I tillegg er effekten av kjønn spesielt stor for basketball og minst for roing. Men, vi har ikke fått oppgitt hvor mange observasjoner vi har av hver av idrett og kjønn, så disse effektene kan skyldes små utvalg.

Q1.6

a) Skriv ned ligningen for den estimerte regresjonsmodellen. Hvor mange regresjonsparametere er estimert?

$\hat{y}_i = 3.66 - 0.71 \cdot \text{Kvinne}_i + 0.22 \cdot \text{Roing}_i + 0.098 \cdot \text{Svoemming}_i + 0.27 \cdot \text{Tennis}_i + 0.17 \cdot \text{Turn}_i + 00126 \cdot \text{Hoeyde}_i - 0.0135 \text{Vekt}_i$

Vi har estimert 8 regresjonsparametere.

Galt: vi teller med at vi har estimert β_0, så det er 8 og ikke 7 regresjonsparametere som er estimert.

Merk: bare kvinne og vekt som er "negative" effekter (gir lavere antall blodceller), og husk at effekten av kvinne er sammenlignet med menn og effekten av idrettene er sammenlignet med Basketball. Det er mann og Basketball som er referansekategoriene for Kjoenn og Sport.

b) Sammenlign den estimerte regresjonskoeffisienten for Hoeyde i denne modellen mot den estimerte regresjonskoeffisienten for Hoeyde i den enkle lineære regresjonen. Har disse to samme fortolkning?

- Enkel lineær regresjon: $\hat{\beta}_{\text{Hoeyde}} = 0.0199$
- Multippel lineær regresjon: $\hat{\beta}_{\text{Hoeyde}} = 0.0126$

Tolkningen er litt annenledes. I MLR så er dette effekten av å være 1 cm høyre hvis vi sammenligner to personer som er av samme kjønn, samme vekt og holder på med samme idrett, mens for enkel lineær regresjon er det generell effekt av høyde - den samme for begge kjønn, ulik vekt og idrett.

c) Hvis vi sammenligner en mann og en kvinne som begge er like høye, veier like mye og begge holder på med samme idrett, hva er gjennomsnittlig forskjell i antall blodceller mellom dem?

Da er i gjennomsnitt det slik at kvinnen har $0.7131 \cdot 10^{12}$ færre blodceller pr liter blod enn mannen.

d) Hva er predikert antall blodceller for en mann som holder på med roing, er 180 høy og veier 75 kg? (Regn for hånd ved å putte inn tall fra resultat.summary().)

$$3.657 + 0 \cdot -0.7131 + 0.2165 + 180 \cdot 0.0126 + 75 \cdot -0.0135 = 5.129$$

10^{12} blodceller pr liter.

Q1.7

a) Forklaringsvariablen 'Sport' er kategorisk og vi har brukt en såkalt dummy-variabelkoding, der 'Basketball' er referansekategorien. Er effekten av de andre sportskategoriene på 'Blodceller' signifikant forskjellig fra effekten for 'Basketball' (på nivå 0.05)?

Ja, for Roing, og Tennis så er det signifikant forskjell fra effekten av Basketball. For de andre (svømming og turn) er det ikke det.

b) Hva er andel forklart variasjon? Ville du forventet at andelen forklart variasjon gikk opp da vi la til flere forklaringsvariabler enn Hoeyde? Hvis vi nå la til en forklaringsvariabel som var IQ til idrettsutøveren, ville da R^2 økt?

Andel forklart variasjon er 71.1%. R^2 vil alltid gå opp eller forbli uforandret hvis vi legger til flere kovariater, så vi forventet en oppgang, men denne oppgangen var relativt stor - fra 24 til 71 %.

Hvis vi legger til IQen til idrettutøveren - som åpenbart ikke kan påvirke antallet røde blodceller utøveren har, så vil ikke R^2 gå ned. Det kan vises at R^22 vil øke eller holde seg uendret hvis man legger til nye forklaringsvariabler til modellen - selv hvis den nye forklaringsvariablen er tilfeldige tall. Grunnen til dette er at det alltid vil være små tilfeldige trender i dataene som fanges opp – regresjonsmodellen finner et tilsynelatende signal i det som bare er tilfeldig støy. Derfor må vi justere R^2 for at vi skal kunne bruke den til å sammenligne to modeller med ulikt antall forklaringsvariabler.

c) Basert på utskrifter og plott. Vil du konkludere med at modelltilpasningen er god?

Ja, vi ser ingen klare trender i plottet av predikert verdi mot residual, og QQ-plottet er også ganske lineært. Sjekker vi med normalitetstestene så bekrefter de at nullhypotesen om normalitet ikke forkastes. Ja, modelltilpasningen er god.

Q1.8

a) Hvor mange regresjonsparametere er nå estimert? Hva er signifikante forklaringsvariabler?

Antall estimerte regresjonsparametere er 4: konstantledd, kjønn, høyde og vekt. Både kjønn, høyde og vekt er signifikante. Konstantleddet pleier vi ikke teste.

b) Er modelltilpasningen god?

Vi ser mye av det samme som for Q1.7c - og modelltilpasningen er grei.

c) Sammenlign Adj. R-squared for modellen med og uten 'Sport'. Hvis vi skal avgjøre om 'Sport' skal være med som forklaringsvariabel ved å bruke Adj. R-squared, hva vil du da konkludere med? Begrunn valget ditt.

Justert R^2 er høyere for modellen med sport - 69.0 mot 66.1 for modellen uten sport. Jeg ville gått for modellen med sport, men hvis man vil ha en enkel regel kan man også kutte ut Sport. Det som er viktig her er at man ikke kan bruke R^2 til å velge mellom to modeller som har ulikt antall regresjonsparametere.

Ikke med: ser man på andre konkurrenter til justert R^2 som er AIC og BIC som rapporteres i utskriften så vil AIC foretrekke modellen med Sport og BIC den uten Sport. AIC og BIC skal være så små som mulig.

Galt: Adj R^2 gir ikke andel forklart varians, det er det barer R^2 som gjør. Det er heller ikke mulig å sammenligne R^2 fra enkel lineær regresjon med forskjeller i Adj R^2.

Oppgave 2: Klassifikasjon

Q2.1

a) Hvorfor ønsker vi å dele dataene inn i trening, validering og test-sett?

Vi ønsker å forhindre at vi overtilpasser en fleksibel modell til dataene våre.

Vi vil bruke treningssettet til å tilpasse modellene, og valideringssettet til velge mellom ulike modeller. Testsettet skal vi bruke for å fortelle hvor gode resultater vi har fått.

b) Hva brukes hver av disse delene til i våre analyser?

Treningsettet brukes til å estimere parametere og usikkerhet i disse estimatene for logistisk regresjon. For k-nærmeste nabo er det treningssettet som bestemmer klassifikasjonsregelen for en gitt k.

For logistisk regresjon kan vi bruke p-verdier for signifikans av koeffisienter ved å bare se på treningssettet, men vi kan også bruke antallet korrekt klassifiserte på valideringssettet. For k-nærmeste nabo bruker i valideringssettet til å finne k.

For både logistisk regresjon og k-nærmeste nabo bruker vi testsettet for å si hvor gode regelen vår vil være på fremtidige data.

c) Hvor stor andel av dataene er nå i hver av de tre settene? Ser de tre datasettene ut til å ha lik fordeling for de tre forklaringsvariablene og responsen?

Ser vi på koden så ber vi om at treningssettet er av størrelse $0.8 \cdot 0.75 = 0.6$ (60%), valideringssettet av størrelse $0.8 \cdot 0 - 25 = 0.2$ (20%) og testsettet også 20%.

Vi tenker her ikke på at formelle tester skal gjøres, bare se på utskrift. Her er andelen 0 og 1 i de tre settene ganske like, 234/(234+237) i treningssettet og akkurat 0.5 for validering og test. (Det er ikke så rart siden settene var laget stratifisert på y.) Videre har dobbeltdiff, acedif og upressetdiff alle ganske like gjennomsnitt for de tre settene, men min og max er ganske forskjellige. Vi har delt inn i settene tilfeldig, så vi vil ikke forvente at store skjevheter skal forekomme.

Merk: å bare skrive "de tre datasettene ser ut til å lik fordeling for de tre forklaringsvariblene" er litt vagt, det er fint med å poengtere at vi kan se på gjennomsnitt og standardavvik, eller median og første og tredje kvartil. Det er også viktig å kommentere andelen 0 og 1 responser.

Q2.2

a) Kommenter hva du ser i plottene og utskriften.

Av tetthetsplottene ser vi at det er noe forskjell på acediff og upressetdiff for respons suksess eller fiasko for spiller 1. For dobbeldiff er fordelingene like de gangene spiller 1 og 2 hadde suksess, og korrelasjonen mellom y og dobbeldiff er lav. Det ser ut som upressetdiff og dobbeldiff sier mye av det samme siden de er ganske høyt korrelerte (0.45).

b) Hvilke av de tre variablene tror du vil være gode til å bruke til å predikere hvem som vant matchen? Begrunn svaret.

Det ser ut som det er informasjon både i acediff og upressetdiff, men ikke så mye i dobbeltdiff.

Kommentar: her var det mange som ikke har kommentert tetthetsplottene i to farger.

Q2.3

a) Hvilke forklaringsvariabler er signifikante i modellen på signifikansnivå 0.05?

Både acediff og upressetdiff er signifikante forklaringsvariabler, mens dobbeldiff ikke er det.
Her har da acediff og upressetdiff p-verdi under 0.05. Når p-verdien er større enn 0.05 så forkastes ikke nullhypotesen at det ikke er noen sammenheng mellom dobbeltdiff og suksess for spiller.

Galt: det er ikke lov å si at H_1 forkastes til fordel for H_0. Det er ikke symmetri her. Hvis p-veriden er stor sier vi at H_0 ikke forkastes - vi har ikke tilstrekkelig grunn til å tvile på H_0.

b) Hvordan kan du tolke verdien av exp(upressetdiff)?

Verdien til exp(upressetdiff) (som står under tabellen) er 0.91. Hvis vi ser på to matcher A og B, der spillerne hadde den samme acediff og dobbeldiff, men man i den match B har upressetdiff+1 mens match A har

upressetdiff. Da vil oddsen for match B (sannsynlighet for å vinne delt på sannsynlighet for å tape) være 0.91 ganger oddsen for match A - dvs at der er mindre sjanse for at spiller 1 å vinne for match B enn A. Gjør man en en ekstra upresset feil endrer oddsen seg - til 0.91 ganger det den var med en feil mindre.

Kommentar: På dette spørsmålet var det svært få som svarte rett.

c) Hva angir feilraten til modellen? Hvilket datasett er feilraten regnet ut fra? Er du fornøyd med verdien til feilraten?

Feilraten til modellen er andelen matcher i valideringssettet som er feilklassifisert. Den regnes ut ved at vi bruker modellen til å predikere resultatet av matchene i valideringssettet og teller antallet feilklassifiseringer. Vi bruker 0.5 som cut-off for sannsynlighet.

Q2.4

a) Diskuter hva du ser.

Acediffs estimerte koeffisient er uendret 0.1895. Mens estimert koeffisient for upresset har endret seg minimalt fra -0.09 til -0.895. Effekten av dobbeltdiff var jo også veldig lav i den første analysen. Feilraten er uendret på 0.215. Vi ser at det er helt fint å ikke har med dobbeldiff, vi har en like god modell som med (iflg feilraten på valideringssettet).

b) Som din beste modell for logistisk regresjon vil du velge modellen med eller uten dobbeldiff som kovariat? Begrunn svaret.

Vi vil alltid ha den enkeste mulige modellen, og modellen uten dobbeltdiff er like god til prediksjon som modellen med - hvis vi kan stole på at valideringssettet et stort nok. Videre er effektene av acediff og upressediff signifikante i modellen med bare disse to, og oddseffekt er 1.2 for acediff og 0.91 for upressetdiff.

Q2.5

Forklar kort hva som er gjort i koden over, og hvilken verdi av k du vil velge.

I koden har vi tilpasset k-nærmeste nabo til treningssettet (bruker acediff og upressediff) for oddetallsverdier av k fra og med 1 til og med 49. Deretter har vi brukt valideringssettet til å predikere om en match blir vunnet av spiller 1 (suksess) eller ikke, og talt opp antallet feilklassifiseringer og regnet ut feilrate. Denne feilraten er så plottet mot k. Vi ser etter minst mulig feilrate, og vil ha en så glatt modell som mulig. Vi ser at lavest verdi for feilraten er for k=31, 33, 43-47. Alle disse verdiene kan godt velges, og jeg velger k=47.

Her kan man godt reflektere over størrelsen på valideringssettet. Det er også mulig å tenke litt på om man burde ha skalert dataene gitt at man bruker euklidsk avstand, eller om de har rimelig likt standardavvik. Jeg har ikke tatt med skalering fordi det ville krevd endel å først regne ut mean og sd for treningssettet og så bruke det på validering og test. Men, noen må gjerne gjøre det hvis de vil - det er mulig det kan bli bedre slik - slik som i kompendiet.

Galt: Det var noen som ikke fortalte hvilket datasett som feilraten ble regnet ut fra, og noen trodde det var på treningssettet.

Riktig: Noen økte k så også høyere verdier ble med, det er helt ok.

Q2.6

a) Vil du foretrekke å bruke logistisk regresjon eller k-nærmeste-nabo-klassifikasjon på tennisdataene?

Jeg foretrekker logistisk regresjon siden det er en enkel metode og den gir minst like gode resultater som kNN her. Hvis vi regner på usikkerhet i estimering av feilratene er det ingen forskjell mellom feilratene på testsettet for de to metodene.

b) Oppsummer hva du har lært at kan være en god metode for å predikere hvem som vinner en tennismatch.

Jeg har lært at den som vinner har hatt flere serveess og færre upressede feil enn motstanderen. oddsratio for acediff er 1.22 og for upressetdiff 0.91.

Gjerne kommentere at logistisk regresjon og kNN med høy k her har funnet ganske lik modell - fra figuren med klassegrenser.

Forvirring: noen var veldig forvirret av at kNN var bedre enn logistisk regresjon på valideringssettet men at det byttet om på testsettet. Vi hadde for små størrelse på valideringssettet og testsettet til å kunne se forskjell på feilraten til logistisk og kNN klassifikasjon - slik at hvis man hadde laget konfidensintervaller for andelen feilklassifiserte (slik vi har lært i fellesdelen med normaltilnærming til binomisk) ville man sett at konfidensintervallene var veldig brede og overlappet hverandre.

Oppgave 3: Klyngeanalyse

Q3.1

a) Hvor mange observasjoner (n) og hvor mange variabler (p) har vi?

Antall observasjoner ser vi i utskriften over, for 'china.jpg' var dette 273280 og $p = 3$ er de tre fargekanalene. Dette kan sees fra utskriften som allerede er implementert.

b) Hvor finner du fargeverdiene til observasjonen med posisjon (x,y)=(10,20) i bildet i den nye tabellen "data_farger"?

```
print(255.0*data_farger[10*640+20,])
```

Formelen vil være: ´data_farger[ant_kolonner * 10 + 20]´ er det samme som bilde[10,20], hvor ant_kolonner er avhengig av bildet som blir lastet inn (640 for 'china.jpg' og ´flower.jpg´0)- så $640 \cdot 10 + 20 = 6420$.

Galt: Mange husker ikke at vi i python teller fra 0 og har sagt $640 * (10 - 9) + 20$. Mange har ikke sjekket at pikselen man finner frem til har samme farge som (10,20) i det originale bildet - det er jo ikke vanskeligere enn å skrive kommandoen over! Det er alltid lurt å sjekke at det man tror er riktig ER riktig. Noen trodde at det var i posisjon 10*20, det er ikke korrekt.

Kommentar: Overraskende mange hadde galt på dette punktet.

Q3.2

Se kode under for hvordan lage et redusert datasett, og plotte rød mot grønn. Du legger til plott av rød mot blå og blå mot grønn. Kommenter hva du ser.

Svaret her er avhengig av hvilket bilde du har valgt, men for 'china.jpg' så er det en klar positiv korrelasjon mellom fargetonen for grønn mot rød, rød mot blå og grønn mot blå. Det betyr at høye verdier for rød ofte er sammen med høye verdier for blå, og ditto for de andre parene. Vi ser også at kodingen er slik at en lav verdi av rød er en mørk pixel og en høy verdi av rød er en lys pixel, dvs at 0=mørk og 1=lys i vårt skalerte bilde.

R vs G: Ingen piksler der G er høy og R er lav, eller G er lav og R er høy.

R vs B: Her har vi ingen piksler der R er lav og B er høy, men noen få omvendt (mot orange).

R vs B: Ingen piksler med mindre G enn B.

Kommentar: Her var det veldig variabelt hvor mye man skrev.

Q3.3

a) Hva er sentroidene for ditt bilde?

Svaret er avhengig av bildet. For bildet 'china.jpg' så har vi en sentroide '[0.82423598 0.85495987 0.88452347]]' som er lysegrå, og en '[0.28338183 0.25556003 0.18524481]' som er nesten brun.

b) I hvilken klynge havner fargene svart, hvit, rød, grønn, blå og gul?

Svart, rød, grønn, og blå hører til klyngen "brun", og hvit og gul hører til "lysegrå" for ´china.jpg´.

Galt: Noen har tilpasset k-means bare til de 6 datapunktene bestående av svart, hvit, rød, grønn, blå og gul. Hvis man gjør det havner rød i klynge med hvit og gul. Dette har ingenting med vårt bilde å gjøre og er helt galt. Man skal ikke tilpasse k-means på nytt, men gjøre slik man får beskjed om i teksten, kalle kmeans.predict (ikke kmeans.fit).

c) Diskuter kort hva du ser.

Her er mesteparten av svaret at man har klart kodingen.

´china.jpg´. I kryssplottene ser vi at det er nesten som om man har delt 90 grader på diagonalen med brunt under og grått over for hver av RG, RB og BG. Det er noen få piklser som er på "gal side".

Galt: her er det endel som har kodet galt, ved at de har satt nye verdier for R, G og B - det skal man ikke, man skal bare endre verdien til fargen i pikslet. Dermed blir kryssplottet veldig rart, fordi det er jo bare to mulige verdier for R,G, B - og alle pikslene blir plottet oppå hverandre.

d) Ved å se på det opprinnelige bildet, er det mulig å se hvilke deler av bildet som hører til hvilken klynge? Forklar!

For akkurat dette bildet får vi et skille mellom forgrunn og bakgrunn, hvor den lyse bakgrunnen hører til klyngen "lysegrå", mens forgrunnen er "brun".

Q3.4

Kommenter og forklar hva du observerer.

Svaret er avhengig av hvilket bilde man bruker. For 'china.jpg' så får vi segmentert bilde i bak- og forgrunn, bortsett fra noen lyse felt i forgrunnen som havner i samme klynge som bakgrunnen. Noen har også brukt defaultbildet ´flower.jpg´ der blomst og bakgrunn var de to klare klyngene.

Q3.5

Hva er hovedforskjellene mellom K-gjennomsnitt-klyngeanalyse og hierarkisk klyngeanalyse? Vi ber deg ikke om å finne klynger i bildet ved hjelp av hierarkisk klyngeanalyse. Hva kan være grunnen til at vi ikke gjør det?

I hierarkisk klyngeanalyse starter vi med å regne ut avstand mellom alle observasjoner (piksler)i bildet. Dette setter vi inn i en avstandsmatrise. Det blir en veldig stor matrise og algoritmen vil gå veldig tregt og trenge veldig mye minne. I tillegg er det veldig mange piksler som må slås sammen hierarkisk for å komme frem til to klynger. Hierarkisk klyngeanalyse (den agglomerative varianten vi har lært om) egner seg derfor ikke godt til slike store datasett.

Q3.6

Hvor mange klynger trenger du for at du synes at bildet ser omtrent ut som det opprinnelige bildet? Prøv ut med ulike antall klynger og finn et klyngeantall du synes gir en god tilnærmelse, både med tanke på farger og detaljer. Hvor mange bit blir brukt per piksel i ditt valg av antall klynger over?

Svaret blir selvfølgelig veldig subjektivt, men jeg synes at vi begynner å få en god tilnærming med 8 klynger. Vi bruker da 3 bit per piksel. Hvis man har 16 klynger blir det 4 bit pr piksel og med 32 klynger blir det 5 bit pr piksel.

Galt: Det var ikke meningen at man skulle lese av filstørrelsen for det nye bildet.

Kommentar: Dette klarte alle som hadde klart å kode i Q3.3, og her var det mange gode svar.