

2a.1)	Hvorfor ønsker vi å dele dataene inn i trening-, validering- og test-sett?
Svar	Å holde testsettet uavhengig av trening- og valideringssettet hjelper med å forhindre at informasjon lekker fra testsettet til treningsprosessen, som kan føre til overoptimistiske resultater.

2a.2)	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?
Svar	Fordeling er 60% treningssett, 20% valideringssett og 20% testsett

2a.3)	La oss si at vi hadde valgt League 1 og 2 som treningssett, Championship som valideringssett, og Premier League som testsett. Hvorfor hadde dette vært dumt?
Svar	Ulikt spiller nivå i ligaene. Dette kan føre til at dataene fra disse ligaene kan være forskjellige når det gjelder spillestil, taktikk, og spillernes ferdigheter. Hvis du trener en modell på lavere nivå data og deretter prøver å bruke den på Premier League-data, kan den gi upålitelige resultater på grunn av disse forskjellene. De underliggende stokastiske prosessene i de forskjellige ligaene er ikke nødvendigvis like, det vil si at man trener på en stokastisk prosess og deretter prøve å predikere en annen stokastisk prosess med denne modellen.

2a.4)	Kommenter kort på hva du ser i plottene og utskriften (maks 5 setninger).
Svar	Scatterplottet viser hvor mye sammenheng det er mellom de forskjellige variablene hvor fargen avgjør om det er seier eller ikke, og hvis vi ser tydelig farge mønster vil det være påvirkning av variabelen mellom seier og tap for hjemmelaget. For eksempel ser vi at høy corner diff samtidig som høy skudd på mål diff har flere hjemmeseiere siden denne delen av scatterplottet er mer orange. Derimot ser vi at det er ingen fargenyanser i plottet corner diff og forseelse diff, og dette betyr at disse ikke påvirker seier i den ene eller andre retningen. Diagonalen viser et enkelt histogram plot for hver av variablene i et histogram for $y = 1$ og $y = 0$ i forskjellig farge. Utskriften viser korrelasjonsmatrisen mellom de 3 variablene.

2a.5)	Hvilke(n) av de tre variablene tror du vil være god(e) til å bruke til å predikere om det blir hjemmeseier? Begrunn svaret kort (maks 3 setninger).
Svar	Korrelasjonsmatrisen viser at det kun er skudd_paa_maal_diff som er positivt korrelert med hjemmelagets seier ($y = 1$). De øvrige variablene viser lite korrelasjon med y . Derimot er det interessant å merke seg at skudd på mål er korrelert med corner.

2b.1)	I en kamp der skudd_paa_maal_diff er 2, corner_diff er -2 og forseelse_diff er 6, hva er ifølge modellen sannsynligheten for at hjemmelaget vinner? Vis utregninger og/eller kode, og oppgi svaret med tre desimaler.
--------------	---

Svar	<pre>#2b.1 skudd_paa_maal_diff = 2 corner_diff = -2 forseelse_diff = 6 intercept = resultat.params["Intercept"] coef_skudd_paa_maal_diff = resultat.params["skudd_paa_maal_diff"] coef_corner_diff = resultat.params["corner_diff"] coef_forseelse_diff = resultat.params["forseelse_diff"] log_odds = intercept + coef_skudd_paa_maal_diff * skudd_paa_maal_diff + coef_corner_diff * corner_diff + coef_forseelse_diff * forseelse_diff import math probability = 1 / (1 + math.exp(-log_odds)) predicted_class = 1 if probability > 0.5 else 0 print("Predicted Probability:", probability) print("Predicted Class:", predicted_class)</pre> <div>Predicted Probability: 0.6097534665532253 Predicted Class: 1</div>
------	---

2b.2)	Hvordan kan du tolke verdien av $e^{\beta_{skudd-paa-maal-diff}}$?
Svar	<p>$e^{\beta_{skudd-paa-maal-diff}}$ er eksponenten til denne koeffisienten. Denne verdien gir en indikasjon på hvor mye multiplikativ effekt en enhets endring i 'skudd_paa_maal_diff' har på oddsen for en positiv utfall (for eksempel 'y = 1').</p> <p>$e^{\beta_{skudd-paa-maal-diff}} = 1$ (Ingen endring i oddsen)</p> <p>$e^{\beta_{skudd-paa-maal-diff}} < 1$ (Økning i oddsen for positivt utfall)</p> <p>$e^{\beta_{skudd-paa-maal-diff}} > 1$ (Minking i oddsen for positivt utfall)</p>

2b.3)	Hva angir feilraten til modellen? Hvilket datasett er feilraten regnet ut fra? Er du fornøyd med verdien til feilraten?
Svar	Feilraten til modellen angir andelen feilklassifiserte observasjoner i forhold til det totale antallet observasjoner i valideringsdatasettet. Den måler hvor godt modellen presterer i å korrekt klassifisere data. Feilraten er regnet ut fra valideringsdatasettet. Å ha en feilrate på 0.285 er solid for å predikere en seier for et hjemmelag.

2b.4)	Diskuter kort hvordan koeffisientene (β – ene) og feilraten endrer seg når forseelse_diff tas ut av modellen (maks 3 setninger).
Svar	Vi så i forrige oppgave at koeffisienten til forseelse_diff var positiv. Koeffisienten var ikke stor, men den vil fortsatt føre til en økning i oddsen for positivt utfall for hjemmelaget. Når man fjerner denne variabelen vil det påvirke de gjenværende koeffisientene.

2b.5)	Med den nye modellen: I en kamp der skudd_paa_maal_diff = 2, corner_diff = -2 og forseelse_diff = 6, hva er sannsynligheten for at hjemmelaget vinner ifølge den nye modellen? Oppgi svaret med tre desimaler.
Svar	<div>Predicted Probability: 0.5908469115899905 Predicted Class: 1</div>

2b.6)	Hvis du skal finne en så god som mulig klassifikasjonsmodell med logistisk regresjon, vil du velge modellen med eller uten <code>forseelse_diff</code> som kovariat? Begrunn kort svaret (maks 3 setninger).
Svar	Korrelasjonen er lav for seier, som vil si at den har liten betydning i oddsen. Dersom vi henter ut <code>z</code> verdien til <code>forseelse_diff</code> i originalmodellen (<code>resultat.summary()</code>) ser vi at den er på 0.396 og signifikansnivået på 0.95 er derfor ikke oppfylt. Vi velger derfor å velge bort variabelen i modellen.

2c.1)	Påstand: kNN kan bare brukes når vi har maksimalt to forklaringsvariabler. Fleip eller fakta?
Svar	Algoritmen fungerer ved å finne de k nærmeste naboene til et gitt punkt i det flerdimensjonale rommet, uavhengig av antall forklaringsvariabler. Det betyr at påstanden er fleip.

2c.2)	Hvilken verdi av k vil du velge?
Svar	Laveste <code>val_feilrate</code> er: 0.282555282555. De korresponderende knaboene er: [115, 119, 133, 159]. Vi velger derfor knaboene [115, 119, 133, 159], fordi de har lavest feilrate. Dette fører til bedre suksessrate for å predikere seier for hjemmelaget.

2d.1)	Gjør logistisk regresjon eller k -nærmeste-nabo-klassifikasjon det best på fotballkampdataene?
Svar	<pre>Feilrate logistisk regresjon: 0.32843137254901966 Feilrate kNN: 0.3259803921568627</pre> <p>Vi ser at kNN har lavere feilrate og er dermed bedre</p>

2d.2)	Drøft klassegrensene (plottet under) for de to beste modellene (én logistisk regresjon og én kNN). Hva forteller klassegrensene deg om problemet? Skriv maksimalt 3 setninger.
Svar	Logistisk regresjon trekker en helt tydelig linje i dataene, hvor det er hjemme seier på den ene siden og tap på den andre siden som en følge av at vi har valgt 0.5 som sannsynlighetsgrense. For kNN plottet ser vi farge plott av 100 x 100 matrisen og skille linjen er her ikke så jevn, og dette kan være en mulig forklaring på høyere treffrate i predikeringen. Begge metodene predikerer bra og vi ser stort sett samme trend i dataene, og plottet av trenings dataene viser at begge modellene gir et rimelig resultat.

3a.1)	Hvilke 3 siffer har vi i datasettet? Hvor mange bilder har vi totalt i datasettet?
Svar	<p>Type sifre er: 9, 3, 8</p> <p>Total images: 6000</p>

3a.2) Hvilket siffer ligner det 500. bildet i datasettet vårt på? Lag et bilde som viser dette sifferet. (Husk at Python begynner nummereringen med 0, og derfor refereres det 500. bildet til [499])

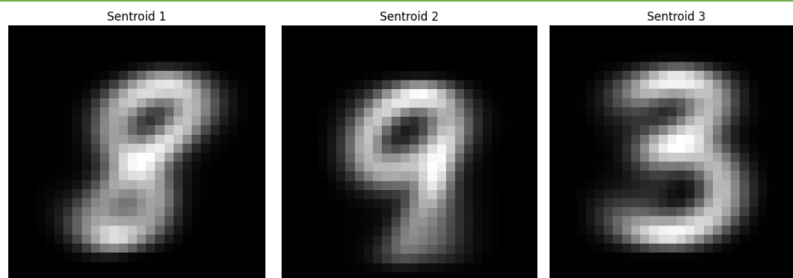
Svar



Dette tallet ligner på tallet 9

3b.1) Tegn sentroidene av de 3 klyngene fra K -gjennomsnitt modellen. Tilpass koden over for å plote. Her kan du ta skjermbilde av sentroidene og lime inn i svararket. Hint: Sentroidene har samme format som dataene (de er 384-dimensjonale), og hvis de er representative vil de se ut som tall.

Svar



3b.2) Synes du at grupperingen i klynger er relevant og nyttig? Forklar. Maks 3 setninger.

Svar

Klyngeanalyse kan være nyttig for å identifisere tydelige grupper i dataene, som kan være nyttige for segmentering eller prediksjoner.

3b.3) Vi har valgt $K = 3$ for dette eksempelet fordi vi vil finne klynger som representerer de 3 sifrene. Men generelt er K vilkårlig. Kom opp med et forslag for hvordan man (generelt, ikke nødvendigvis her) best kan velge K . Beskriv i egne ord med maks 3 setninger.

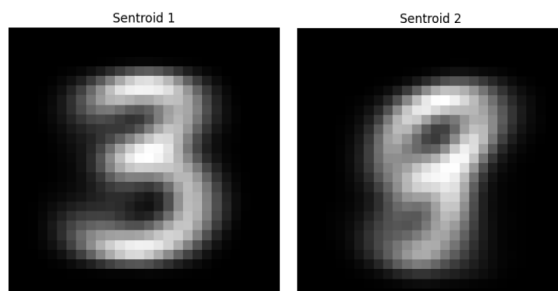
Svar

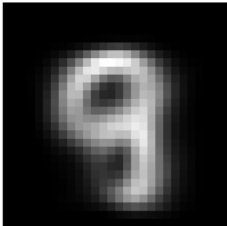
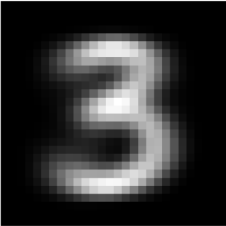
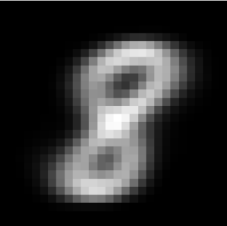
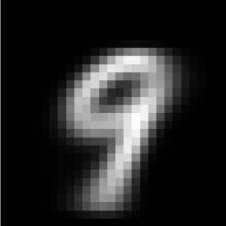
The Elbow Method: Dette innebærer å kjøre K -gjennomsnitt med forskjellige verdier for K og plote summen av kvadrerte avstander (SSE) mot K . Man ser etter et punkt i grafen der SSE begynner å flate ut, som ligner på en albue.
Silhouette Score: Dette måler hvor godt hver datainstans er klynget sammen med andre i samme klynge sammenlignet med nærmeste naboende klynge.

3b.4) Kjør analysen igjen med $K = 2$ og $K = 4$. Synes du de nye grupperingene er relevante?

Svar

$K = 2$



	<p>K = 4</p> <div style="display: flex; justify-content: space-around; align-items: flex-start;"> <div style="text-align: center;"> <p>Sentroid 1</p>  </div> <div style="text-align: center;"> <p>Sentroid 2</p>  </div> <div style="text-align: center;"> <p>Sentroid 3</p>  </div> <div style="text-align: center;"> <p>Sentroid 4</p>  </div> </div> <p>Tallene som tidligere var delt mellom tre klynger, blir nå presset inn i to og fire klynger. Dette fører til en større sammenblanding av tallene og mindre klare skiller mellom gruppene. Vi ser for eksempel at i K = 2 så er tallet 9 og 8 blandet sammen i samme klynge.</p>
--	--

3c.1) Vurder dendrogrammet nedenfor. Synes du at den hierarkiske grupperingsalgoritmen har laget gode/meningfulle grupper av bildene? (Maks 3 setninger).

Svar Min vurdering av dendrogrammet er positiv. Det ser ut som den hierarkiske grupperingsalgoritmen har produsert meningsfulle grupper av bildene. De er tydelig og mulig å forstå for mennesker.

3c.2) I koden under har vi brukt gjennomsnittskobling (`method = 'average'`). Hvordan fungerer gjennomsnittskobling? (Maks 3 setninger).

Svar Gjennomsnittskobling er en metode som brukes i hierarkisk klyngeranalyse for å beregne avstandene mellom klynger. Metoden vurderer gjennomsnittet av avstandene mellom alle par av datapunkter i to forskjellige klynger når man kobler dem sammen.

3c.3) Velg en annen metode enn 'average' til å koble klyngene sammen (vi har lært om dette i undervisningen, her heter de `single`, `complete` og `centriod`) og lag et nytt dendogram ved å tilpasse koden nedenfor. Ser det bedre/verre ut? (Maks 3 setninger).

Svar Single vil koble sammen klynger basert på den korteste avstanden mellom datapunktene i de to klyngene. Dette kan føre til klynger som er langstrakte og har en tendens til å være følsomme for støy og uteliggere. Single vil gjøre det verre, på grunn av at den er mer følsom på uteliggere og støy

3d.1) Hvis vi skulle brukt en metode for å predikere/klassifisere hvilket siffer et håndskrevet tall er, og ikke bare samle dem i klynge, hva ville du brukt?

Svar Vi ville brukt konvensjonelle nevrale nettverk for bildegjenkjenning og klassifisering. Man kan trene et CNN med et stort datasett av håndskrevne siffer og bruke det trente nettverket til å forutsi sifferklassen for nye bilder.

