|  |  |
| --- | --- |
| 2a.1) | Hvorfor ønsker vi å dele dataene inn i trening-, validering- og test-sett? |
| Svar |  |

|  |  |
| --- | --- |
| 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 |  |

|  |  |
| --- | --- |
| 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 |  |

|  |  |
| --- | --- |
| 2a.4) | Kommenter kort på hva du ser i plottene og utskriften (maks 5 setninger). |
| Svar |  |

|  |  |
| --- | --- |
| 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 |  |

|  |  |
| --- | --- |
| 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 |  |

|  |  |
| --- | --- |
| 2b.2) | Hvordan kan du tolke verdien av ? |
| Svar |  |

|  |  |
| --- | --- |
| 2b.3) | Hva angir feilraten til modellen? Hvilket datasett er feilraten regnet ut fra? Er du fornøyd med verdien til feilraten? |
| Svar |  |

|  |  |
| --- | --- |
| 2b.4) | Diskuter kort hvordan koeffisientene () og feilraten endrer seg når forseelse\_diff tas ut av modellen (maks 3 setninger). |
| Svar |  |

|  |  |
| --- | --- |
| 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 |  |

|  |  |
| --- | --- |
| 2b.6) | Hvis du skal finne en så god som mulig klassifikasjonsmodell med logistisk regresjon, vil du velge modellen med eller uten forseelse\_diff som kovariat? Begrunn kort svaret (maks 3 setninger). |
| Svar |  |

|  |  |
| --- | --- |
| 2c.1) | Påstand: kNN kan bare brukes når vi har maksimalt to forklaringsvariabler. Fleip eller fakta? |
| Svar |  |

|  |  |
| --- | --- |
| 2c.2) | Hvilken verdi av vil du velge? |
| Svar |  |

|  |  |
| --- | --- |
| 2d.1) | Gjør logistisk regresjon eller -nærmeste-nabo-klassifikasjon det best på fotballkampdataene? |
| Svar |  |

|  |  |
| --- | --- |
| 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 |  |

|  |  |
| --- | --- |
| 3a.1) | Hvilke 3 siffer har vi i datasettet? Hvor mange bilder har vi totalt i datasettet? |
| Svar | Datasettet inneholder totalt 6000 bilder og består av sifrene 3, 8 og 9. |

|  |  |
| --- | --- |
| 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 | Ligner på siferet ‘9’ |

|  |  |
| --- | --- |
| 3b.1) | Tegn sentroidene av de 3 klyngene fra -gjennomsnitt modellen. Tilpass koden over for å plotte. 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 | Grupperingen av de håndskrevne tallene i klynger ved bruk av -gjennomsnitt modellen ser ut til å bli både relevant og nyttig. Sentroidene som representerer hver klynge viser ganske tydelig og gjenkjennelig sifrene ‘8’, ‘9’ og ‘3’, selvsagt med en viss uskarphet. Dette indikerer at modellen har klart å fange opp de underliggende mønstrene og variasjonene til hvert av sifrene. |

|  |  |
| --- | --- |
| 3b.3) | Vi har valgt for dette eksempelet fordi vi vil finne klynger som representerer de 3 sifrene. Men generelt er vilkårlig. Kom opp med et forslag for hvordan man (generelt, ikke nødvendigvis her) best kan velge . Beskriv i egne ord med maks 3 setninger. |
| Svar | For å finne beste *K*-verdi benyttes gjerne «The Elbow Method» eller «The Silhouette Method», hvorav disse er begge egne tester/utregninger som kan og bør gjøres i Python, eller lignende språk, i forkant av selve klyngeanalysen. «The Silhuette Method» vil være mer presis, men den er også mer krevende/tregere og foretrekkes derfor ikke for større datasett. |

|  |  |
| --- | --- |
| 3b.4) | Kjør analysen igjen med og . Synes du de nye grupperingene er relevante? |
| Svar | Nei, det er veldig åpenbart utifra sentroidene at disse nye grupperingene ikke er relevante. Med K = 3 ser vi et tydelig skille mellom de tre sifrene, hvorav med lavere eller høyre K-verdier så prøver algoritmen å kombinere eller skille ut dataene i datasettet til færre eller flere sifre selv om datasettet bare inneholder tre forskjellige.  A group of numbers in black squares  Description automatically generatedA number in a square  Description automatically generated with medium confidence |

|  |  |
| --- | --- |
| 3c.1) | Vurder dendrogrammet nedenfor. Synes du at den hierarkiske grupperingsalgoritmen har laget gode/meningfulle grupper av bildene? (Maks 3 setninger). |
| Svar | Nei. Grupperingsalgoritmen har blandet mye mellom hvilke bokstaver som tilhører sammen og ikke. Den treffer ofte godt på tallet ‘8’, men har lett for å blande spesielt sifrene ‘3’ og ‘9’. |

|  |  |
| --- | --- |
| 3c.2) | I koden under har vi brukt gjennomsnittskobling (method = 'average'). Hvordan fungerer gjennomsnittskobling? (Maks 3 setninger). |
| Svar | "Average" metoden i **scipy.cluster.hierarchy** for dendrogrammer fungerer ved å først beregne gjennomsnittlig avstand mellom alle par av datapunkter i separate klynger. Under klyngeprosessen kombineres de to klyngene med minst gjennomsnittlig avstand til en enkelt klynge. Denne prosessen repeteres til alle datapunkter er gruppert i én klynge, og resultatet visualiseres som et dendrogram. |

|  |  |
| --- | --- |
| 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 dendrogram ved å tilpasse koden nedenfor. Ser det bedre/verre ut? (Maks 3 setninger). |
| Svar | Endte med å bruke metode 'complete' da denne så merkbart mer konsistent og riktig ut enn noen av de andre metodene. Nå er klyngene mye mer representative for hva som faktisk finnes i datasettet og er for det meste plassert sammen med samme siffer som seg selv. Grupperingsalgoritmen blander fortsatt såesielt sifferet ‘3’ med ‘8’ og ‘9’, dog er likevel nå mer presis enn før. |

|  |  |
| --- | --- |
| 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 | *k*-nærmeste-nabo-klassifikasjon (kNN)  https://www.math.ntnu.no/emner/IST100x/ISTx1003/Klassifikasjon.html  Klyngeanalyse er en form for uovervåket læring. Ved denne typen læring jobber algoritmen selvstendig med å gruppere/klynge dataen i et datasett ved å bruke ulikhetene mellom dataene til å identifisere skillet mellom dem, utifra K-verdien som er satt, elementene som deler  Klyngeanalyse representerer en type uovervåket læringsteknikk. I denne tilnærmingen jobber algoritmen selvstendig med å gruppere data i et datasett basert på likheter og ulikheter innenfor dataene. Målet er å identifisere naturlige grupperinger eller klynger, der elementer innenfor en klynge deler visse karakteristikker, mens de er distinkt forskjellige fra elementer i andre klynger. |