|  |  |
| --- | --- |
| 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 | De tre sifrene vi har i utvalget av datasettet er 3, 8 og 9. Datasettet består av 6000 bilder, mens utvalget består av 10 bilder. |

|  |  |
| --- | --- |
| 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 | Bildet siffer 500 ligner på tallet 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 | Klyngene ser ut til å representere de 3 sifrene som er i utvalget fra datasettet. |

|  |  |
| --- | --- |
| 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 | Artikkelen introduserer to metoder for å finne et forslag for K, «The Elbow Method» og «The Silhouette Method». |

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

|  |  |
| --- | --- |
| 3c.1) | Vurder dendrogrammet nedenfor. Synes du at den hierarkiske grupperingsalgoritmen har laget gode/meningfulle grupper av bildene? (Maks 3 setninger). |
| Svar | Vi syns grupperingsalgoritmen har laget noen gode/meningsfulle grupper av bildene. |

|  |  |
| --- | --- |
| 3c.2) | I koden under har vi brukt gjennomsnittskobling (method = 'average'). Hvordan fungerer gjennomsnittskobling? (Maks 3 setninger). |
| Svar | Med average starter alle MNIST-vektorene som sin egen klynge, og ved hver iterasjon beregner man en ny klynge basert på gjennomsnittsavstanden mellom alle verdiene i vektorene. Det vil si at for hver iterasjon vil to klynger bli til en. |

|  |  |
| --- | --- |
| 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-metoden hadde veldig mange feilkoblinger med enkelt-noder til allerede etablerte klynger og hadde dårlig klyngegruppering når avstanden ble større. Complete-metoden lager et veldig ryddig dendrogram og var veldig god på å sette tre og ni i forskjellige klynger, men åtte ble godt blandet mellom tre og ni. Centroid-metoden lagde et veldig rotete dendrogram og viste ingen bevis på at den klarer å sortere datasettet ordentlig. |

|  |  |
| --- | --- |
| 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 | Dette er en typisk oppgave som egner seg godt for nevrale nettverk. Et nevralt nettverk bruker noder, vektede kanter og bias for å lagre informasjon i en nodegraf. Deretter brukes tilbakepropagering for å trene opp det nevrale nettverket til å gi mer og mer korrekte svar. Da kan man gi et bilde til det nevrale nettverket, og det vil fortelle deg hvilket tall det er. |