Klassifisering av dyrearter

Mathias Eira Karlsen

31 august 2024

Bilde generert ved hjelp av ChatGPT og DALL-E. Prompt: “make an image of an animal that is a mix of many mammals”

1 INTRODUKSJON

Oppgaven var å lage en enkel funksjon for å klassifisere dyrearter inni klasser basert på noen utvalgte egenskaper. Med oppgaven fulgte det med to datasett, et treningssett hvor klassen er kjent, og et testsett hvor dette er ukjent. Det er 16 egenskaper, og 7 klasser som dyreartene skal klassifiseres inni.

2 TEORI

Dyrearter innenfor samme klasse har mange like egenskaper. For å klassifisere dyreartene så kan man se hvor nært egenskapene er til de forskjellige klassene. Man kan finne ut hvordan en klasse ser ut som ved å ta gjennomsnittet av alle artene i en klasse i treningssettet. Så regner man forskjellen mellom dyrearten og klassen ved å ta den absolutte verdien av (A – B) for hver egenskap hvor A er en egenskap til dyrearten og B er gjennomsnittet til en egenskap for klassen. Det blir regnet en forskjells verdi for hver klasse. Den klassen som har minst forskjeller fra dyrearten er den som blir valgt.

3 METODE

Metoden ble delt i to funksjoner. get_animal_averages() og classify_animal(). get_animal_averages() tar ingen argumenter. Den genererer og returnerer en liste som inneholder 7 lister med gjennomsnittsegenskaper, en for hver klasse. Treningssettet forandres ikke, så for å unngå å bruke tid på å generere den samme listen hver gang funksjonen blir brukt, så brukes det en @cache decorator fra Python sin functools modul. Den lagrer resultatet fra den første gangen funksjonen blir brukt, og returnerer det samme resultatet uten å kjøre koden på nytt. classify_animal() tar 16 egenskaper og returnerer et tall i området 1-7 som tilsvarer 7 dyreklasser. Funksjonen henter klassegjennomsnittene med get_animal_averages() og brukes de for å finne hvilken klasse som er nærmest egenskapene.

4 RESULTAT

Når programmet brukes på trenings dataen, så klassifiserer den 71/76 dyrearter i riktig klasse. Tabellen under viser mer detaljert hva programmet fikk riktig og feil. Den fikk 2 av 37 mammal feil og 3 av 6 inverebrate feil.

	Mammal	Bird	Reptile	Fish	Amphibian	Bug	Invertebrate	**Total**
Correct	35	16	2	9	1	5	3	**71**
Wrong	2	0	0	0	0	0	3	**5**

Den andre tabellen viser hvordan programmet klassifiserte dyrene som den fikk feil.

Species	Dolphin	Scorpion	Seal	Slug	Worm
Guess	4 (Fish)	6 (Bug)	4 (Fish)	3 (Reptile)	3 (Reptile)
Correct answer	1 (Mammal)	7 (invertebrate)	1 (Mammal)	7 (invertebrate)	7 (invertebrate)

5 DISKUSJON

Resultatene var bedre enn forventet. Jeg forventet at med en så simpel algoritme så ville den kanskje få omtrent halvparten riktig, men med over 90% riktig så slo den mine forventninger. Årsaken til at resultatet var bedre enn forventet kan være at dyreartene innenfor en klasse var nærmere hverandre enn jeg forventet, eller at gjennomsnittene til klassene var mer forskjellige en jeg trodde. Det at programmet tok feil for delfiner og sel var forventet. Gjennomsnittspattedyret hadde 3.5 føtter, så det at de hadde 0 føtter gjorde det usannsynlig at programmet klarte å klassifisere de riktig.

Testsettet som fulgte med oppgaven hadde mange pattedyr, fugler og fisk, men lite dyrearter innenfor de andre klassene. Så det er stor sannsynlighet at de dyrene ikke er representativ for klassen. Programmet vil også ikke virke bra for dyr som er for langt unna gjennomsnittet, for eksempel akvatiske pattedyr som ikke har føtter. Føtter er den egenskapen som lager mest problemer siden den ikke bare er 0 eller 1.

En bedre metode for å klassifisere dyrene kunne sett på noen egenskaper som viktigere for en klasse en andre, for eksempel så lager alle pattedyr melk, mens ingen dyr i treningssettet lager melk. Alle fugler har fjær mens ingen andre klasser har det, osv. Man kan også forbedre den metoden hvis man har bedre klasser å klassifisere dyrene i. For eksempel er bug en underklasse til invertebrate.

6 KONKLUSJON

Oppgaven var å lage en enkel funksjon for å klassifisere dyrearter basert på egenskaper til dyrearten. Metoden som ble brukt var basert på ideen at dyrearter innenfor samme klasse har mange like egenskaper. Ved å ta gjennomsnittet av artene i en klasse så får man et gjennomsnitts dyr for den klassen, denne blir brukt for å se hvor forskjellig en dyreart er fra en klasse. For å klassifisere en dyreart så regner man forskjellen til hver klasse, og velger klassen med minst forskjeller. Denne metoden gjorde det bra på treningsdataen så ble brukt for å regne gjennomsnittsverdiene, med 71/76 riktig. Metoden klarer ikke å klassifisere dyrearter riktig, hvis de er for forskjellig fra gjennomsnittet.