Problem 20: Obserwacja ptaków

Punkty: 55

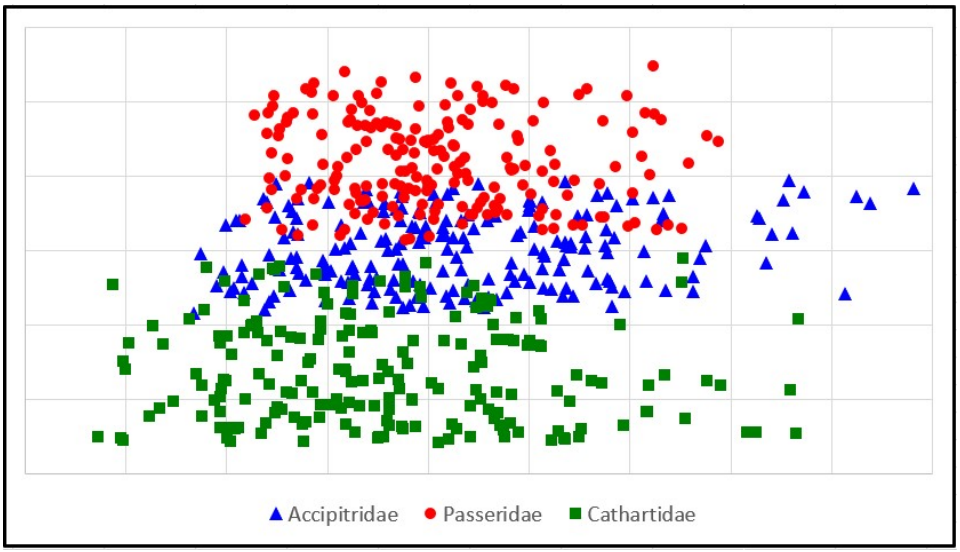
Autor: Joe Worsham, Colorado Springs, Kolorado, Stany Zjednoczone

Wprowadzenie do problemu

Uczenie maszynowe (ML - Machine Learning) to wymyślny rodzaj algorytmu sztucznej inteligencji, który wykorzystuje wzorce w uprzednio zarejestrowanych danych do budowy przewidywań dotyczących danych, z jakimi się jeszcze nie spotkał. Algorytmy ML mogą być bardzo złożone i wykorzystywane do rozwiązywania bardzo trudnych problemów; na przykład, sieci neuronowe symulują zachowanie poszczególnych komórek w prawdziwym, żywym mózgu. Niektóre z nich są jednak proste, zwyczajne i bardzo skuteczne w rozwiązywaniu problemów dotyczących wzorców. W omawianym problemie macie stworzyć własny system uczenia maszynowego!

Współpracujecie ze służbami lokalnego parku, aby posortować wykonywane przez nich ostatnio zdjęcia ptaków. Chcą używać tych zdjęć do śledzenia populacji pewnych gatunków, ale zdjęć jest bardzo dużo i potrzebna będzie właściwa metoda organizacyjna. Postanowili skorzystać z metod stosowanych w taksonomii i poukładać zdjęcia według rodziny ptaka widocznego na fotografii. Przy takiej liczbie fotografii potrzebują automatycznej metody do poukładania zdjęć.

Wasza współpracowniczka proponuje, by użyć pomiarów ptaków uzyskanych ze zdjęć do przewidywania rodziny taksonomicznej danego osobnika. Znajduje zbiór informacji na temat długiej listy gatunków ptaków i pokazuje, że między pewnymi pomiarami a rodziną ptaka istnieje widoczna korelacja:



Powyższy wykres to algorytm T-distributed Stochastic Neighbor Embedding (t-SNE), który zamienia pewną liczbę pomiarów (w tym przypadku cztery) na dwuwymiarowe współrzędne. Każdy punkt jest oznaczony kolorem wskazującym na rodzinę taksonomiczną ptaka. Zgodnie ze spostrzeżeniami współpracowniczki większość kolorowych punktów jest pogrupowana w pobliżu siebie, co nadaje wagi jej hipotezie o możliwości znalezienia wzorca. Proponuje, by ustalić, jak „daleko" znajduje się nieznany ptak ze zdjęcia od tych pomiarów i użyć tej wartości do postawienia hipotezy dotyczącej rodziny tego ptaka.

Opis problemu

Pomysł waszej współpracowniczki jest nazywany algorytmem k-Nearest Neighbor (kNN). Pozwala on przewidywać rodzinę nieznanego ptaka na podstawie dostępnych danych taksonomicznych i pomiarów badanego osobnika. Algorytm kNN działa obliczając odległość między nieznanym punktem danych i każdym znanym punktem danych. Następnie *k* znanych punktów danych, znajdujących się najbliżej nieznanej danej, używa się do „zagłosowania" w celu otrzymania ostatecznej decyzji.

W omawianym problemie wasz algorytm powinien obliczać odległość od nieznanego punktu danych do każdego znanego punktu danych. Po obliczeniu wszystkich tych odległości należy policzyć, ile razy każda rodzina ptaków pojawia się w **K** najbliższych punktów. To właśnie jest proces nazwany powyżej „głosowaniem". Rodzina otrzymująca najwięcej głosów zostaje wybrana jako rodzina nieznanego osobnika.

Zasadniczo, wartość **K** w tym algorytmie musi być liczbą całkowitą; jeśli możliwe są tylko dwie odpowiedzi, jest to zwykle nieparzysta liczba całkowita, aby podczas głosowania nie padł remis. Tutaj mamy trzy możliwe odpowiedzi, zatem remisy będą dopuszczalne. Aby uwzględnić tę przesłankę, zacznijmy od początkowej wartości **K** = 5 dla wszystkich nieznanych ptaków. Jeśli padnie remis, należy zwiększyć **K** o jeden tyle razy, by przechylić szalę na korzyść jednego z wyników; następnie przy kolejnym nieznanym ptaku ponownie zacząć od **K** równego 5.

Wzór służący do obliczania odległości między *N*-wymiarowymi punktami jest następujący:

$$d_{p_1,p_2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + \cdots + (N_1 - N_2)^2}$$

Każdy ptak będzie reprezentowany przez (nie licząc rodziny) cztery punkty danych, czyli jego długość, szerokość ciała, rozpiętość skrzydeł i kąt ustawienia skrzydeł w stosunku do ciała.

Przykładowe dane wejściowe

Pierwszy wiersz danych wejściowych waszego programu, **otrzymanego przez standardowe wejście**, będzie zawierać dodatnią liczbę całkowitą oznaczającą liczbę przypadków testowych. Każdy przypadek testowy będzie zawierać następujące wiersze danych wejściowych:

- Wiersz składający się z dwóch dodatnich liczb całkowitych oddzielonych spacjami: **X**, reprezentującą liczbę znanych ptaków oraz **Y**, reprezentującą liczbę nieznanych ptaków.
- **X** wierszy zawierających informacje o znanych ptakach. Każdy wiersz będzie zawierać następujące wartości oddzielone spacjami:
    - Jedno ze słów „Accipitridae", „Passeridae" lub „Cathartidae", będące rodziną taksonomiczną ptaka
    - Liczbę dziesiętną reprezentującą długość ptaka w calach.
    - Liczbę dziesiętną reprezentującą szerokość ciała ptaka w calach.
    - Liczbę dziesiętną reprezentującą rozpiętość skrzydeł ptaka w calach.
    - Liczbę dziesiętną reprezentującą kąt skrzydeł ptaka w stopniach.
- **Y** wierszy zawierających informacje o nieznanych ptakach widocznych na fotografiach. Każdy wiersz będzie zawierać następujące wartości oddzielone spacjami:
    - Liczbę dziesiętną reprezentującą długość ptaka w calach.
    - Liczbę dziesiętną reprezentującą szerokość ciała ptaka w calach.
    - Liczbę dziesiętną reprezentującą rozpiętość skrzydeł ptaka w calach.
    - Liczbę dziesiętną reprezentującą kąt skrzydeł ptaka w stopniach.

```
1
15 3
Accipitridae 12.30 7.03 25.32 88.59
Accipitridae 21.38 7.57 22.18 88.71
Passeridae 16.57 7.05 25.88 89.27
Passeridae 13.34 6.24 21.37 88.95
Passeridae 15.75 6.58 22.16 89.35
Accipitridae 15.16 5.17 22.43 89.04
Cathartidae 18.61 6.68 23.37 88.83
Accipitridae 21.32 8.14 20.09 88.55
Cathartidae 18.35 7.01 20.64 88.14
Cathartidae 13.61 5.33 23.72 90.21
Cathartidae 16.88 6.63 24.59 88.48
Accipitridae 15.63 8.66 23.19 88.51
Passeridae 17.29 7.62 26.46 89.31
Passeridae 20.03 8.68 20.97 89.05
Cathartidae 19.19 7.74 22.31 88.09
19.37 15.35 17.30 15.28
12.76 21.96 14.41 16.84
20.33 15.51 16.29 17.10
```

Przykładowe dane wyjściowe

W każdym przypadku testowym wasz program powinien wyświetlić przewidywaną rodzinę taksonomiczną każdego nieznanego ptaka, po jednej w każdym wierszu.

```
Accipitridae
Cathartidae
Accipitridae
```