Gaussian Mixture Model (GMM)

Klasteryzacja danych za pomocą mieszanki rozkładów Gaussa

AUTORZY:
MARIUSZ FURTEK
WIKTORIA MACHOWSKA

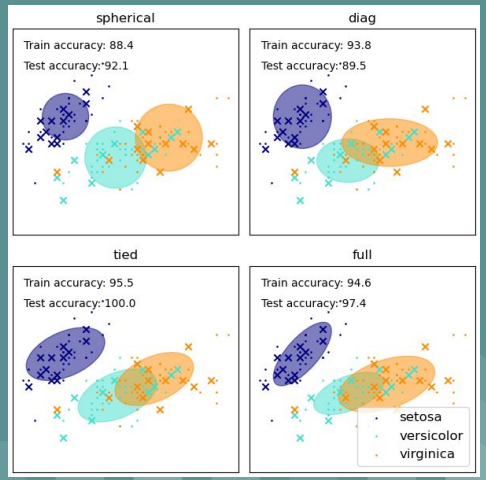
DATA: 02.12.2024 r.

Czym jest klasteryzacja

Klasteryzacja to metoda analizy danych, której celem jest podział obiektów na grupy (klastry) w taki sposób, aby obiekty w jednym klastrze były bardziej podobne do siebie niż do obiektów z innych klastrów. Jest to technika uczenia bez nadzoru, co oznacza, że nie korzysta z wstępnie oznaczonych etykiet danych.

Czym jest Gaussian Mixture Model (GMM)

probabilistyczne podejście do klasteryzacji, które zakłada, że dane mogą być opisane jako mieszanka kilku rozkładów Gaussa. W przeciwieństwie do prostych metod, takich jak K-Means, które przypisują punkt do jednego konkretnego klastra, GMM przydziela prawdopodobieństwo przynależności do każdego klastra.



Teoria GMM

Dane są reprezentowane jako kombinacja K rozkładów Gaussa.

Każdy punkt należy do klastra z określonym prawdopodobieństwem.

μ – średnia (lokalizacja klastra),

Σ – kowariancja (kształt klastra),

π – waga komponentu (prawdopodobieństwo klastra).

Algorytm EM:

E-step: wyliczenie prawdopodobieństw przynależności do klastra.

M-step: dopasowanie parametrów modelu do danych.

Praktyczne zastosowanie GMM

- Segmentacja obrazów w komputerowym rozpoznawaniu wzorców.
- Rozpoznawanie mowy (analiza akustyczna).
- Analiza danych biologicznych (np. analiza genów).
- Wykrywanie oszustw w systemach płatniczych.

Dlaczego GMM jest takie dobre?

Możemy w łatwy sposób wyliczyć niepewność klasyfikacji

Dobrze radzi sobie z danymi o nieregularnych kształtach

Biblioteki użyte w projekcie

matplotlib - biblioteka do tworzenia wykresów i wizualizacji danych

sklearn - biblioteka do uczenia maszynowego w Pythonie, która oferuje szeroki zestaw narzędzi do analizy danych i budowy modeli

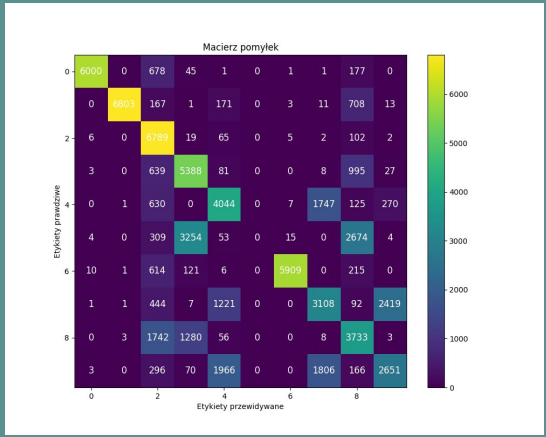
numpy - biblioteka do obliczeń numerycznych w Pythonie. Dostarcza narzędzia do pracy z tablicami (np. macierzami i wektorami) i macierzami wielowymiarowymi

unittest - do konstrukcji testów stworzonego modelu

W projekcie użyto zbioru danych, który zawiera 70 000 obrazków, został on wczytany dzięki następującej linii kodu:

```
mnist = fetch_openml('mnist_784')
```

Wizualizacja wyników



Wizualizację przedstawiono dzięki bibliotece matplotlib. Po przeprowadzeniu klasteryzacji przy użyciu modelu GMM, przypisujemy etykiety uzyskane z modelu do rzeczywistych etykiet cyfr za pomocą funkcji **map_labels**.

Wizualizacja macierzy pomyłek w postaci obrazu pomaga w łatwy sposób zrozumieć, w których przypadkach model się pomylił. Kolory w macierzy wskazują liczbę pomyłek.

Ewaluacja modelu klasteryzacji GMM

Po przypisaniu etykiet za pomocą funkcji **`map_labels`**, możemy ocenić dokładność klasteryzacji, porównując przypisane etykiety z rzeczywistymi etykietami z zestawu danych MNIST. Im bardziej trenuje się model tym lepsze efekty powinien przynosić.

```
D:\PyCharm\PythonProject\.venv1\Scripts\python.exe D:\PyCharm\PythonProject\main.py
Dokładność klasteryzacji: 63.46%

Process finished with exit code 0
```

Testy i wyniki

Wszystkie testy są zaprojektowane w celu zapewnienia, że każda funkcja w projekcie działa zgodnie z oczekiwaniami:

- Ładowanie danych,
- Redukcja wymiarów,
- Trenowanie modeli GMM,
- Mapowanie etykiet,
- Ewaluacja wyników

Przeprowadzone testy upewniają nas, że algorytm działa prawidłowo i zgodnie z założeniami

Podsumowanie

GMM służy do klasteryzacji. Jest modelem probabilistycznym, który pozwala dokładnie analizować złożone struktury danych.

Zaletą jest to, że można modelować dane o różnym kształcie oraz uwzględnia niepewność klasteryzacji.

Jednym z kierunków rozwoju projektu jest możliwość klasteryzacji danych wielowymiarowych.

Dziękujemy za uwagę