Klasteryzacja w uczeniu nienadzorowanym z wykorzystaniem autokodera wariacyjnego i GMM

Martyna Grygiel

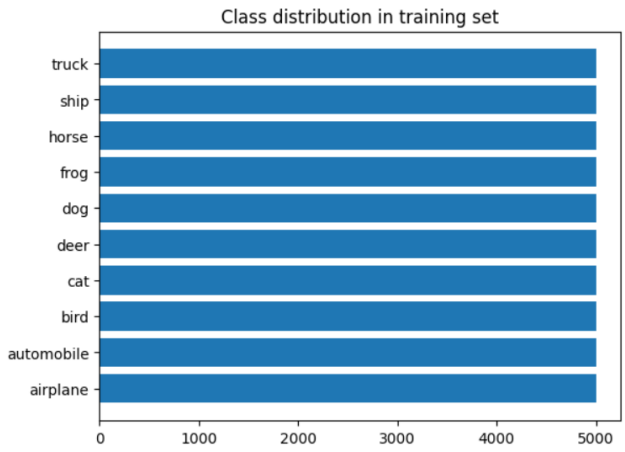
Iga Miller

1. Eksploracyjna analiza danych

Zbiorem wykorzystanym do realizacji zadania jest Cifar10. Jest to zbiór obrazów utworzony przez Canadian Institute for Advanced Research. Składa się z 60 000 kolorowych obrazów 32x32 w formacie RGB, w 10 różnych klasach. Klasy są reprezentowane przez samoloty, samochody, ptaki, koty, jelenie, psy, żaby, konie, statki i ciężarówki. Zbiór posiada 50 000 obrazów treningowych i 10 000 obrazów testowych. Istnieje 6000 zdjęć każdej klasy. Poniżej znajdują się przykładowe obrazy z każdej klasy.

Rysunek 1 Przykładowe obrazy ze zbioru CIFAR10

Przeprowadzona analiza dystrybucji w klasach potwierdziła, że dane są zbalansowane. Histogram przedstawia rozkład obrazów w klasach.



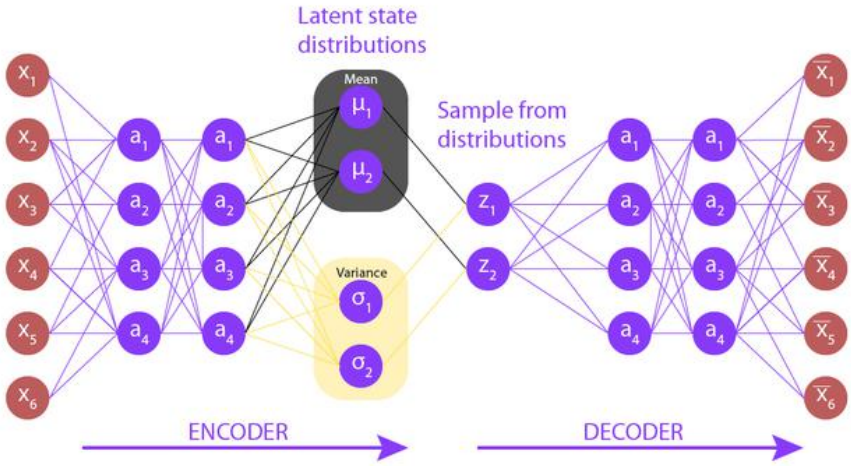
Rysunek 2 Rozkład obrazów w klasach

2. Modele

1. Wariacyjny autokoder

Wariacyjne autokodery (VAE) reprezentują klasę głębokich modeli generatywnych, które znalazły szerokie zastosowanie w uczeniu maszynowym i sztucznej inteligencji. VAE, podobnie jak tradycyjne autoenkodery, składają się z dwóch podstawowych elementów: kodera i dekodera. Koder

odwzorowuje dane wejściowe na ukrytą (ukrytą) reprezentację, podczas gdy dekoder odwzorowuje tę ukrytą reprezentację z powrotem na pierwotną przestrzeń wejściową. Jednak w przeciwieństwie do tradycyjnych autoenkoderów, VAE wprowadzają elementy probabilistyczne do tego procesu.



Rysunek 3 Architektura VAE

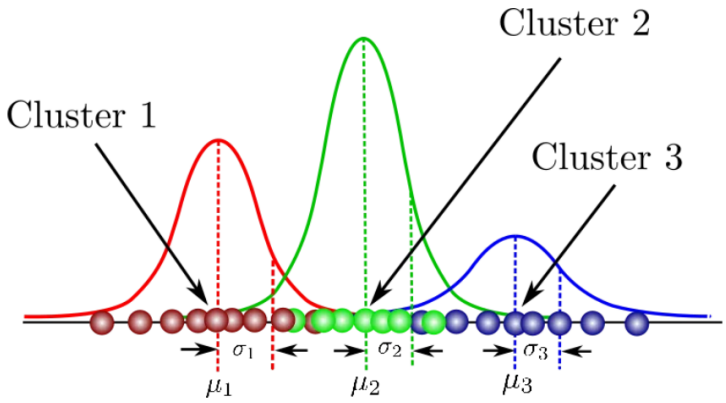
W VAE koder nie odwzorowuje sygnału wejściowego na pojedynczy punkt w przestrzeni utajonej. Zamiast tego wyprowadza parametry rozkładu prawdopodobieństwa. Ten rozkład jest zwykle wybierany jako wielowymiarowy Gauss ze względu na wykonalność obliczeniową. Dlatego zamiast mapowań deterministycznych, VAE uczą się mapowań stochastycznych z przestrzeni wejściowej do ciągłej przestrzeni utajonej. Ten losowy proces składa się z generowania wektora ukrytego z rozkładu a priori i generowania obserwacji z rozkładu warunkowego.

Podczas szkolenia, zamiast po prostu przekazywać średnią rozkładu jako ukrytą reprezentację (co uniemożliwiłoby propagację wsteczną ze względu na nieodłączną losowość), VAE wykorzystują technikę znaną jako „sztuczka reparametryzacyjna". Sztuczka polega na próbkowaniu zmiennej losowej ze standardowego rozkładu normalnego, a następnie skalowaniu i przesuwaniu jej za pomocą wektorów średniej i odchylenia standardowego generowanych przez koder. Pozwala to na przechodzenie gradientów przez koder i dekoder, dzięki czemu możliwe jest kompleksowe szkolenie z wykorzystaniem wstecznej propagacji.

Funkcja celu VAE (funkcja straty) mierzy, jak dobrze VAE może zrekonstruować dane wejściowe i często jest to ujemny logarytm wiarygodności danych wejściowych przy zrekonstruowanym wyjściu. Dywergencja Kullbacka-Leiblera (KL) dąży do tego, aby wyuczony rozkład utajony był zbliżony do wcześniej zdefiniowanego rozkładu (zwykle standardowego rozkładu normalnego). Służy to jako forma regularyzacji, zapobiegająca przeuczeniu i zapewniająca płynniejszą interpolację w przestrzeni ukrytej.

2. GMM

Do zamodelowania priora została wykorzystana Mikstura Gaussów. GMM to model statystyczny zakładający, że wszystkie punkty danych są generowane z mieszaniny skończonej liczby rozkładów Gaussa, z których każdy o nieznanych parametrach. Można go traktować jako probabilistyczny model reprezentujący subpopulacje o normalnym rozkładzie w całej populacji. Każdy składowy rozkład w modelu mieszaniny oddaje jedną z subpopulacji. Aspekt „mieszaniny" modelu wynika z założenia, że każdy punkt danych pochodzi z jednego ze składowych Gaussa, przy czym określony składnik jest wybierany losowo, zgodnie z pewnymi prawdopodobieństwami. Rozkład Gaussa (znany również jako rozkład normalny) jest sparametryzowany za pomocą wektora średniego i macierzy kowariancji, opisujące odpowiednio środek rozkładu oraz kształt i orientację rozkładu.



Rysunek 4 GMM

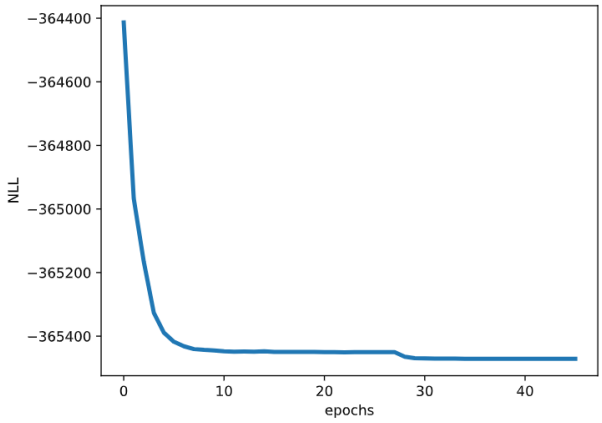
GMM jest sparametryzowany za pomocą zestawu wag mieszania, zestawu średnich i zestawu macierzy kowariancji. Wagi mieszania reprezentują prawdopodobieństwo, że losowo wybrany punkt danych pochodzi z każdego składowego Gaussa. Celem jest jest oszacowanie tych parametrów. Zwykle odbywa się to za pomocą algorytmu maksymalizacji oczekiwań (EM).

3. Eksperymenty

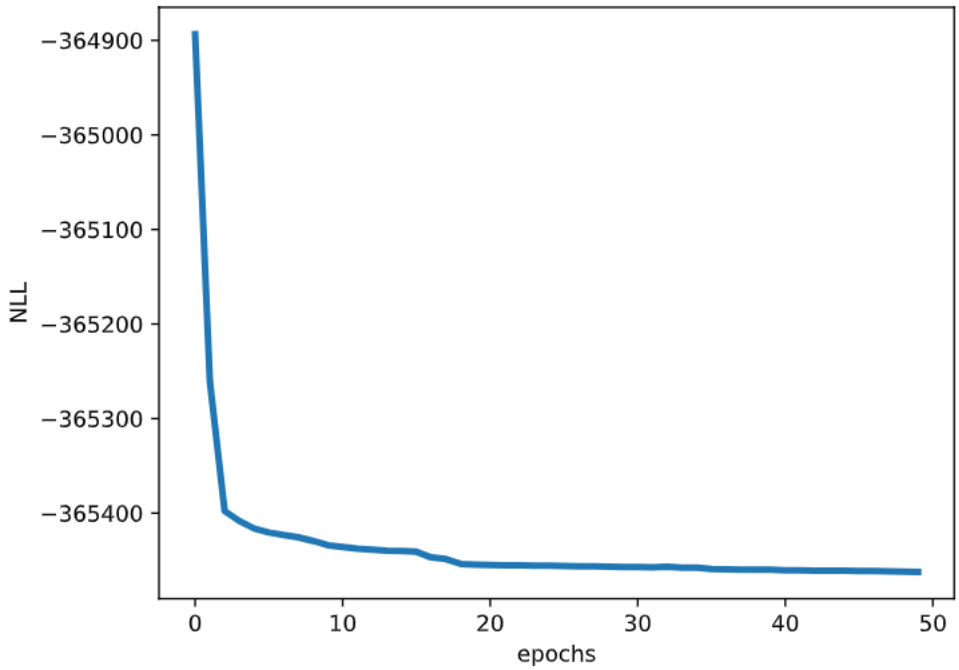
Przeprowadzone eksperymenty obejmowały badanie wpływu na otrzymane wyniki:

- **Priora**– analizowano użycie GMM i rozkładu Gaussa
- **Optymalizatora** – rozpatrywano optymalizator Adam oraz AdaMax
- **Liczby neuronów w warstwie ukrytej** – wybrano do eksperymentów 64, 128 i 256 neuronów
- **Rozmiaru ukrytej reprezentacji** – analizowano zastosowanie rozmiaru 2, 4, 10 i 20

Eksperymenty zostały wykonane dla 50 epok uczenia, ze stałym hiperparametrem patience równym 5 epok. Wykreślono również wykresy dla negatywnego logarytmu prawdopodobieństwa (ang. Negative Log-Likelihood). Poniżej zaprezentowano przykładowe zestawienie przebiegów NLL dla domyślnych ustawień z priorem GMM i rozkładem Gaussa.



Rysunek 5 NLL w zależności od kolejnych epok uczenia na zbiorze walidacyjnym – rozkład Gaussa z domyślnymi ustawieniami jako prior



Rysunek 6 NLL w zależności od kolejnych epok uczenia na zbiorze walidacyjnym – GMM z domyślnymi ustawieniami jako prior

Do porównania jakości klasteryzacji wykorzystano metryki Silhouette Coefficient oraz Davies-Bouldin score. Poniżej w tabeli przedstawiono zestawienie wyników dla przeprowadzonych eksperymentów:

Tabela 1. Wpływ priora na wyniki klasteryzacji

Prior	Silhouette Coefficient	Davies-Bouldin
GMM	- 0,00049	45,11945
Rozkład Gaussa	-0,000515	45,74189

Tabela 2. Wpływ optymalizatora na wyniki klasteryzacji

Optymalizator	Silhouette Coefficient	Davies-Bouldin
Adam	-0,00048	45,60274
AdaMax	-0,00047	45,24829

Tabela 3 Wpływ liczby neuronów na jakość klasteryzacji

Liczba neuronów	Silhouette Coefficient	Davies-Bouldin
64	-0,00044	45,48605
128	-0,00044	45,37883
256	-0,00047	45,50234

Rozmiar ukrytej reprezentacji	Silhouette Coefficient	Davies-Bouldin
2	-0,00046	45,77360
4	-0,00046	45,46548
10	-0,00046	45,43227
20	-0,00049	45,36972

Oprócz tego sprawdzono inny sposób wyliczenia dywergencji KL – z wykorzystaniem funkcji kl_divergence z biblioteki Pytorch. Wyliczenia przeprowadzono dla rozkładu Gaussa. Wyniki jednak nie zmieniły się znacząco w stosunku do reszty: Silhouette Coefficient na poziomie ~-0.0005, a Davies-Bouldin ~45.77.

4. Podsumowanie

Analizując wyniki Silhouette i Davies-Bouldin, nie uzyskano szczególnie dobrych wyników klasteryzacji w żadnym z przypadków. Próba dostrajania hiperparametrów metody nie zakończyła się sukcesem, wynik Silhouette zbliżony do zera w każdym przypadku wskazuje na nakładające się klastry, a wartość ujemna dodatkowo informuje o nieprawidłowościach w przypisaniu do poszczególnych klastrów. Dodatkowo na mieszanie się klastrów wskazuje zaskakująco wysoki Davies-Bouldin. Brak zmian w stosunku do hiperparametrów wskazuje na szukanie przyczyny raczej w architekturze modelu niż w m.in. rozmiarze reprezentacji ukrytej czy ilości neuronów. Problemem nie jest także wyliczenie empiryczne dywergencji KL, bo korzystając z funkcji zaproponowanej przez bibliotekę Pytorch otrzymano zbliżone wyniki. Ze względu na znaczne podobieństwo krzywych NLL nie zamieszczono pozostałych wyników.

W kontekście przyszłych analiz dobrym pomysłem, który wysnuwa się przy analizie powyższego problemu byłoby przeanalizowanie modelu dla łatwiejszego w analizie i interpretacji zbioru, na przykład MNIST.