Klasteryzacja

Wprowadzenie do sztucznej inteligencji - Wykład 8

Maciek Gębala

9 maja 2025

Notatki

Uczenie nienadzorowane

Uczenie nadzorowane

- Mieliśmy zbiór danych postaci $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ złożony z wektorów cech x_i wraz z etykietami y_i.
- Szukaliśmy modelu $\hat{F}$ który pozwoliłby przybliżyć zależność F między cechami a etykietami na zbiorze uczącym

$$y = \hat{F}(x; \theta) \approx F(x).$$

Uczenie nienadzorowane

- Dane nie zawierają informacji o etykiecie i są postaci $x_1, x_2, \ldots, x_m$ złożone z wektorów cech x_i.
- Dane nie mają etykiet - ich uzyskanie może być trudne i kosztowne.
- Przy uczeniu nienadzorowanym odkrywamy wiedzę zawartą w danych.

Notatki

Uczenie nienadzorowane

- Podstawową metodą uczenia nienadzorowanego jest analiza skupień, nazywana także klasteryzacją.
- Celem jest odkrywanie elementów o wspólnych cechach i łączenie ich w grupy zwane klastrami.
- Jeśli w danych występują cechy, które nie są ważne, można je usunąć przez zastosowanie redukcji wymiarowości (analiza składowych głównych - principle component analisis).

Analiza skupień

- Najprostszym algorytmem klasteryzacji i algorytm k-średnich (k-means).
- Algorytm łączy elementy traktując podobieństwo jako odległość między wektorami cech.
- Algorytm minimalizuje sumę kwadratów odległości między wektorami klastra a jego centrum - wyznaczanego jako średnia wektorów w klastrze.

Notatki

Algorytm Lloyda-Forgyego

- Opracowany w 1965 przez Stuarta Lloyda w Bell Laboratories (ale opublikowany w 1982).
- Podobny algorytm opublikował w 1965 Edward W. Forgy.

- Metoda dobrze się sprawdza jeżeli dane układają się w wyraźne grupy o podobnym rozproszeniu.
- Podział jest zależny od hiperparametru k - liczby klastrów.
- Często przypisanie do klastrów jest różne dla różnych wyborów początkowych.
- Znalezienie optymalnej liczby klastrów i ich kształtów jest problemem NP-trudnym.
- Podział danych metodą k-średnich prowadzi do podziału przestrzeni na komórki według centrów (diagram Voronoi).

Notatki

Algorytm k-średnich

Liczba klastrów?

W jaki sposób dobrać liczbę klastrów aby otrzymać w miarę dobrą klasteryzację?

Inercja

Suma kwadratów odległości elementów zbioru od centrów klastrów (centroidów).

Algorytm k-means dąży do klastrów C_j z centroidami μ_j dla których suma inercji w obrębie klastrów jest najmniejsza

$$\min_{\forall \mu_j \in C_j} \sum_{j=1}^{k} \sum_{x_i \in C_j} ||\mu_j - x_i||^2$$

Jeżeli zwiększanie k wyraźnie zmniejsza inercję, to można wyróżnić w zbiorze kolejną dobrze wydzieloną grupę.

Notatki

Algorytm k-średnich

Celem algorytmu k-średnich jest minimalizacja inercji, ale do jej obliczenia konieczne jest wyznaczenie klastrów i ich centroidów.

Jeżeli mamy już wyznaczone k punktów reprezentujących klastry to algorytm powtarza dwa poniższe kroki

1. Przypisz punkty do klastrów przez przypisanie najbliższego centroidu.
2. Wyznacz nowe centroidy jako środki ciężkości klastra.

aż inercja zmienia się mniej niż zadany próg.

Przykład na tablicy 

Notatki

Wybór początkowych centroidów

- Najprościej jest wylosować k elementów ze zbioru wejściowego - może to prowadzić do zbiegania do lokalnego minimum.
- Można kilkukrotnie powtórzyć procedurę i losowanie punktów początkowych.
- Możemy spróbować ręcznie wyznaczyć przybliżone centroidy i wykorzystać je do inicjalizacji algorytmu (częściowa etykietyzacja zbioru wejściowego).
- Możemy losowy wybór uzupełnić o warunek odpowiedniej odległości między wybranymi punktami.

Notatki

Procedura wyboru początkowego

Jeśli założymy, że klastry są od siebie wyraźnie oddzielone, to w 2006 Arthur & Vassilvitskii zaproponowali skuteczną metodę wyboru początkowych centroidów.

Pierwszy element jest wybierany dowolnie, a kolejne są wybierane zależnie od ich odległości od centroidów już wybranych zgodnie z prawdopodobieństwem

$$\frac{D(x_i)^2}{\sum_{j=1}^{m} D(x_j)^2},$$

gdzie $D(x_i)$ to odległość elementu x_i od najbliższego centroidu.

Przykład na tablicy 

Notatki

Słabości algorytmu k-średnich

Algorytm działa poprawnie jeśli klastry są w miarę jednorodne i tworzą obszary wypukłe.

Minimalizacja inercji nie sprawdza się gdy klastry mają kształty nieregularne.
Algorytm będzie wtedy próbował tworzyć większą liczbę klastrów o jednorodnej strukturze.

Notatki

Algorytm DBSCAN

Density-Based Spatial Clustering of Applications with Noise

Idea

Klastry są tworzone na podstawie gęstości występowania elementów.

- Ustalamy promień sąsiedztwa (otoczenia).
- Jeżeli otoczenie elementu zawiera pewną minimalną liczbę sąsiadów to element staje się jądrem klastra.
- Wszystkie elementy w otoczeniu jądra klastra należą do tego samego klastra.
- Otoczenie jądra może zawierać inne jądra i w ten sposób formuje się duży klaster.
- Elementy nie mające w pobliżu jąder są uznawane za szum.

Przykład na tablicy 

Notatki

Algorytm DBSCAN

DBSCAN jest kontrolowany przez dwa parametry

- *eps* - promień otoczenia do wyznaczania sąsiedztwa,
- $min_{samples}$ - minimalna liczba sąsiadów do utworzenia jądra.

Większe $min_{samples}$ i mniejsze *eps* oznaczają, że do utworzenia jądra potrzeba większej gęstości elementów.

DBSCAN sam wyznacza liczbę klastrów.

Notatki

Algorytm BIRCH

Balanced Iterative Reducing and Clustering using Hierarchies

- Algorytm zaprojektowany dla dużych baz danych.
- Bazuje na obserwacji, że przestrzeń atrybutów nie jest obsadzona jednorodnie i nie wszystkie punkty są tak samo ważne.
- Algorytm buduje drzewo (hierarchię) pozwalającą na szybkie określenie przynależności do klastra.

Notatki

Podsumowanie: Wady i zalety

- Metoda k-średnich zakłada, że dane układają się w jednorodne klastry.
- Odpowiednia inicjalizacja klastrów pozwala na unikanie lokalnych minimów.
- Dla danych, które nie układają się w jednorodne grupy lepsze wyniki daje metoda DBSCAN.

Notatki

Podsumowanie: Zastosowania

Zastosowania metody analizy skupień to m.in.

- Segmentacja obrazów – podział na obszary podobne do siebie.
- Wykrywanie anomalii – określenie, które zdarzenia są nietypowe (poza klastrami).
- Wykrywanie wspólnot – odkrywanie grup w analizie sieci społecznościowych.
- Wstępne przetwarzanie danych – odnalezione klastry są wejściem do kolejnego etapu obróbki danych.

Notatki

Problem liczby wymiarów

Wraz ze wzrostem wymiarów (liczby elementów wektorów) maleje średnia gęstość elementów w przestrzeni - wzrasta średnia odległość między elementami.

Przykład

Rozpatrzmy parę losowych punktów w d wymiarowej hiperkostce o boku 1. Jaka jest średnia odległość między punktami w zależności od d?

Dla d równego 1 odległość wynosi $1/3$.
Dla d równego 2 odległość wynosi ≈ 0.52.
Dla d równego 3 odległość wynosi ≈ 0.66.
Dla dowolnego d odległość wynosi $\approx \sqrt{d/6 - 7/120}$.

Notatki

Klątwa wielowymiarowości

Klątwa wymiarowości polega na wykładniczym wzroście ilości danych potrzebnych do zbudowania modelu wraz ze wzrostem wymiaru przestrzeni cech.

W przypadku wysokowymiarowych przestrzeni punkty znajdują się daleko od siebie. Jeżeli weźmiemy pod uwagę wystarczająco dużo cech, to każdy badany obiekt ma cechę, która wyróżnia go wśród innych.

Notatki

Klątwa wielowymiarowości

Co to właściwie znaczy?

- Ze względu na rozproszenie punktów, wielowymiarowe zbiory danych mogą być bardzo rzadkie.
- Zatem większość przykładów będzie znajdować się daleko od siebie. Oznacza to również, że element dla którego chcemy wykonać predykcję będzie prawdopodobnie daleko od jakiejkolwiek przykładu ze zbioru treningowego.
- A zatem predykcja będzie znacznie mniej wiarygodna niż w niższych wymiarach, ponieważ będzie oparta na znacznie większych ekstrapolacjach.
- Czyli im większy wymiar wektora, tym trudniej o uogólnienie a łatwiej o przetrenowanie.

Notatki

Klątwa wielowymiarowości

Z powodu klątwy wymiarowości nie jest możliwe wykorzystanie odpowiedniej liczby przykładów, która pozwoliłaby na równomierne próbkowanie przestrzeni cech.

Redukcja wymiarowości

Redukcja wymiarów to transformacja danych z przestrzeni wielowymiarowej do przestrzeni o niższym wymiarze, tak aby reprezentacja niskowymiarowa zachowała znaczące właściwości oryginalnych danych.

Metody redukcji wymiarowości dzieli się na metody liniowe i metody nieliniowe.

Alternatywnie: na selekcję cech oraz projekcję cech.

Redukcję wymiarowości można użyć do redukcji szumów, wizualizacji danych, analizy skupień lub jako etap pośredni ułatwiający inne analizy.

Notatki

Metody redukcji wymiarowości

Liniowe

- Analiza składowych głównych – znajdowanie hiperpłaszczyzny leżącej najbliżej obserwacji.
- Linear Discriminat Analysis.
- Rozkład według wartości osobliwych (SVD).

Nieliniowe

- Isomap.
- Autoenkodery.

Notatki

Analiza składowych głównych

Principle Component Analysis

Podstawowa zasada działania PCA: znajdź te podprzestrzenie, które nie wnoszą dużo do danych i odrzuć je.

Zbiór danych $x_1, x_2, \ldots, x_m$ złożony z n-wymiarowych wektorów cech układamy w macierz $X = [x_1 x_2 \ldots x_m]$.

Najczęściej wymiar danych n jest dużo większy od m.

Każda kolumna odpowiada obserwacji n cech. Każdy wiersz odpowiada m realizacjom cechy (zmiennej losowej $x^{(i)}$).

Notatki

Maksymalizacja wariancji

Które cechy najlepiej opisują obserwowany proces?

Takie, które najbardziej różnicują nasze obserwacje, dla których wariancja jest największa.

PCA znajduje kierunki w przestrzeni cech dla których wariancja jest największa – składowe główne.

Metoda PCA operuje na macierzy kowariancji dla obserwacji, z elementami postaci

$$cov(x^{(i)}, x^{(j)}) = E[(x^{(i)} - E[x^{(i)}])(x^{(j)} - E[x^{(j)}])],$$

gdzie $x^{(i)}$ to i-ta cecha w wektorze danych. Alternatywnie

$$cov(x^{(i)}, x^{(j)}) = \frac{1}{2n^2} \sum_{k=1}^{m} \sum_{l=1}^{m} (x_k^{(i)} - x_l^{(i)})(x_k^{(j)} - x_l^{(j)}).$$

Notatki

Maksymalizacja wariancji

Wektory własne macierzy kowariancji odpowiadające największym wartościom własnym odpowiadają kierunkom o największej zmienności.

Normalizacja

Wartości składowych zależą od wartości liczbowych cech i przed procedurą PCA konieczne jest wykonanie normalizacji zmiennych.

Notatki

Analiza składowych głównych

Jak maksymalizacja wariancji usuwa wymiary?

- Rozkład na wartości własne macierzy kowariancji jest wykonywany na macierzy z wartościami momentów centralnych.
- Przekształcenie danych poprzez obrócenie do układu współrzędnych wyznaczonego przez wektory własne macierzy kowariancji jest związane z rozkładem według wartości osobliwych (Singular Value Decomposition).
- Po odrzuceniu najmniejszych wartości osobliwych możliwe jest powrócenie do pierwotnego układu współrzędnych. Odrzucenie pewnych wartości osobliwych powoduje, że znikają współrzędne w odpowiadających im podprzestrzeniach.

Notatki

Analiza składowych głównych

Rozkład według wartości osobliwych jest uogólnieniem rozkładu na wartości własne.

SVD może być wyliczone dla dowolnej macierzy, niekoniecznie kwadratowej.

Notatki

Zastosowanie redukcji wymiarowości

Unikanie klątwy wymiarowości

Jeżeli mamy $2n$ obserwacji, ale każda z nich to wektor o wymiarze n, to redukcja wymiarów pozwala na unikanie przetrenowania.

Kompresja danych

Usuwanie elementów o małej normie. Kompresja stratna obrazów.

Zrozumienie danych

Eliminacja mało ważnych parametrów pozwala wyłapać istotne zależności.

Wykrywanie zdarzeń nietypowych

Notatki

Notatki

Notatki

Notatki