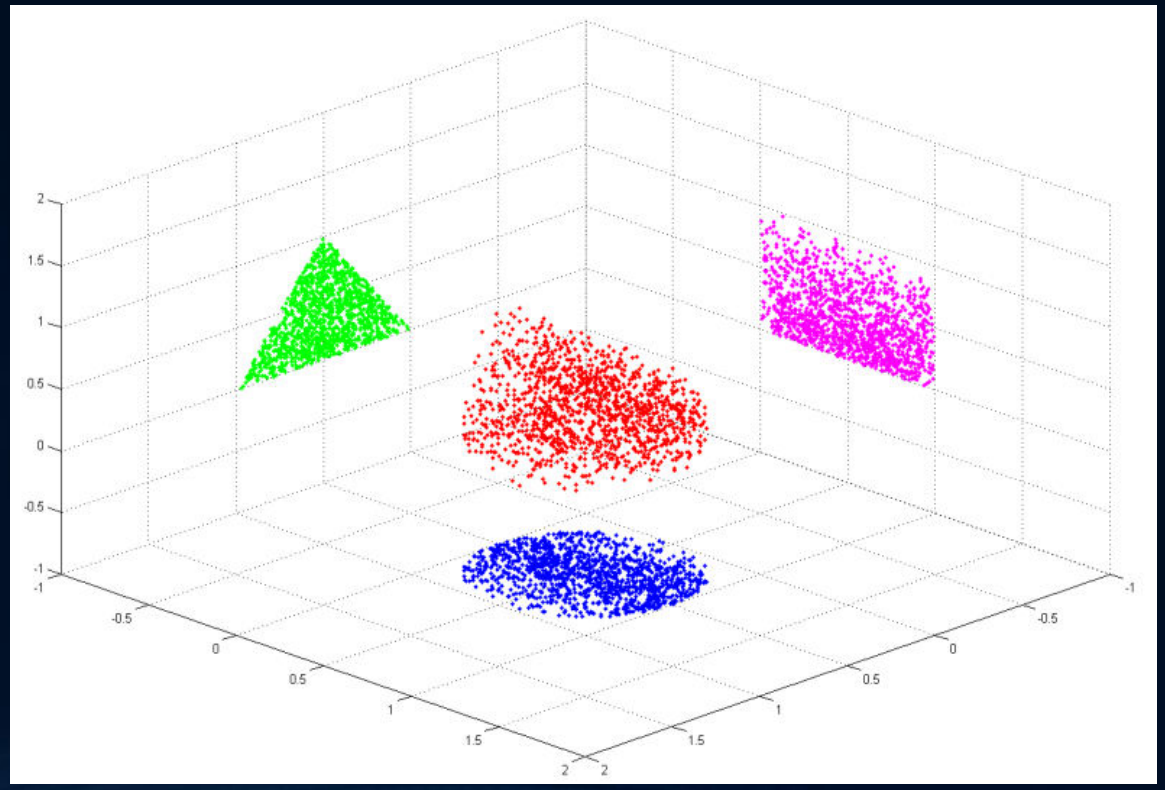
Cel i metody redukcji wymiarowości danych masywnych

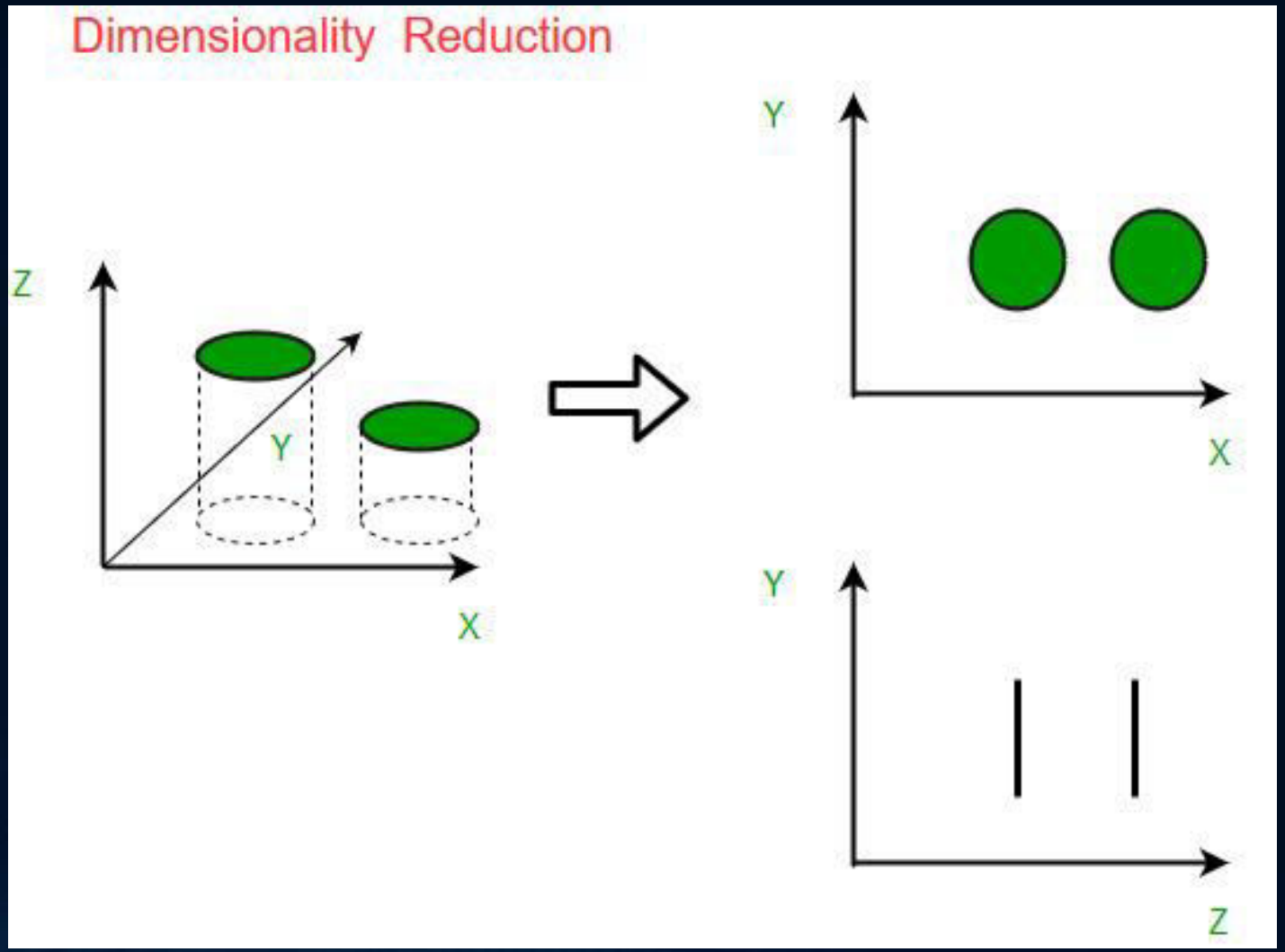
AUTOR: NIKITA BURIK

Redukcja wymiarowości

Redukcją wymiarowości określa się wyznaczanie zbioru cech, mniej licznego od zbioru cech pierwotnych, na podstawie których można z możliwie najmniejszym błędem odtworzyć wartości cech pierwotnych. Jest to równoznaczne ze znalezieniem przestrzeni o niższej wymiarowości, wybranej w taki sposób, aby przerzutowanie do niej danych wiązało się z możliwie małą utratą informacji.



Dlaczego redukcja wymiarów jest ważna w uczeniu maszynowym i modelowaniu predykcyjnym?



Składniki redukcji wymiarów

Istnieją dwa składniki redukcji wymiaru:

- Wybór funkcji: próbujemy znaleźć podzbiór oryginalnego zestawu zmiennych lub funkcji, aby uzyskać mniejszy podzbiór, który można wykorzystać do modelowania problemu. Zazwyczaj obejmuje to trzy sposoby:

    1. Filter
    2. Opakowanie (Wrapper)
    3. Osadzone (Embedded)

- Wyodrębnianie funkcji:

    Powoduje to zmniejszenie danych w przestrzeni o dużych wymiarach do przestrzeni o niższym wymiarze, tj. przestrzeni o mniejszych ilościach wymiarów.

Metody redukcji wymiarów

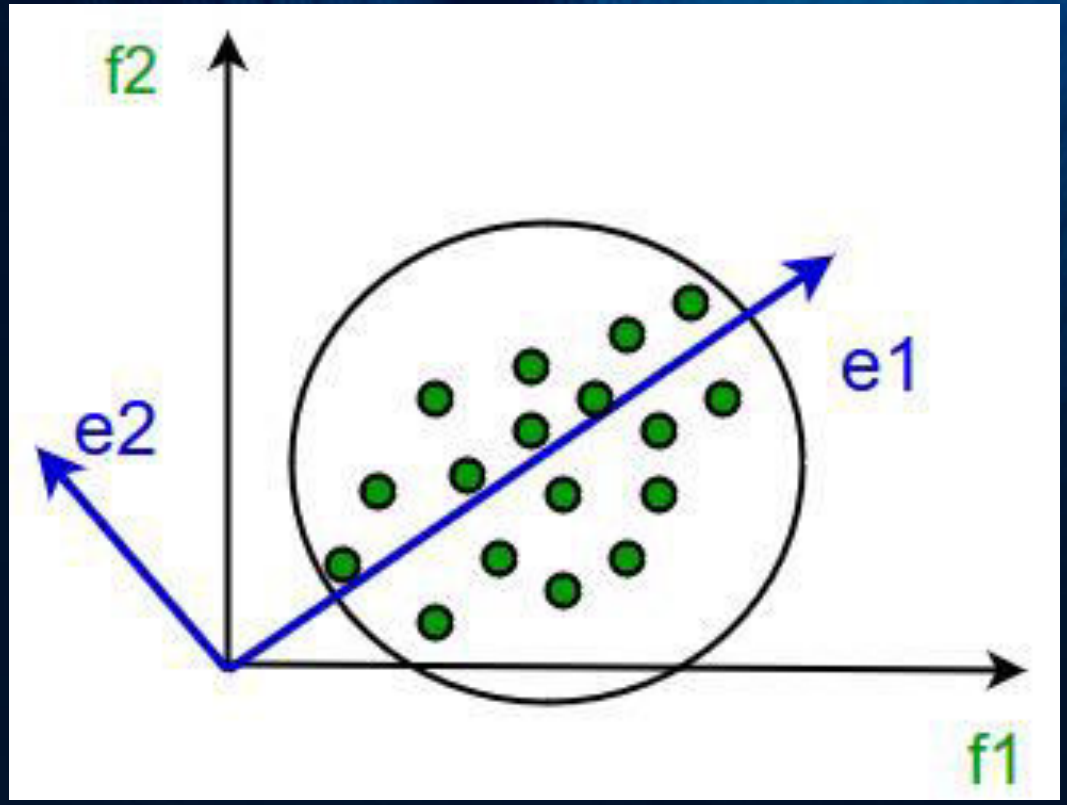
Różne metody stosowane do redukcji wymiarów obejmują:

- Analiza składowych głównych (PCA)
- Liniowa analiza dyskryminacyjna (LDA)
- Ogólne modele analizy dyskryminacyjnej (GDA)

Zmniejszenie wymiarów może być zarówno liniowe, jak i nieliniowe, w zależności od zastosowanej metody.

Analiza składowych głównych (PCA)

Ta metoda została wprowadzona przez Karla Pearsona. Działa pod warunkiem, że podczas gdy dane w przestrzeni wyższego wymiaru są odwzorowywane na dane w przestrzeni o niższych wymiarach, wariancja danych w przestrzeni o niższych wymiarach powinna być maksymalna.



Obejmuje następujące kroki:

- Skonstruuj macierz kowariancji danych.
- Oblicz wektory własne tej macierzy.
- Wektory własne odpowiadające największym wartościom własnym są używane do rekonstrukcji dużej części wariancji oryginalnych danych.

W związku z tym pozostaje nam mniejsza liczba wektorów własnych i może nastąpić pewna utrata danych w tym procesie. Jednak najważniejsze wariancje powinny zostać zachowane przez pozostałe wektory własne.

Zalety redukcji wymiarów

- Pomaga w kompresji danych, a tym samym zmniejsza przestrzeń do przechowywania.
- Skraca czas obliczeń.
- Pomaga także usunąć zbędne funkcje, jeśli takie istnieją.

Wady redukcji wymiarów

- Może to prowadzić do pewnej utraty danych.
- PCA ma tendencję do znajdowania liniowych korelacji między zmiennymi, co czasami jest niepożądane.
- PCA zawodzi w przypadkach, gdy średnia i kowariancja nie są wystarczające do zdefiniowania zbiorów danych.

Liniowa analiza dyskryminacyjna (LDA)

LDA, jak sama nazwa wskazuje, jest jedną z metod analizy dyskryminacyjnej, a więc może być stosowana, gdy chcemy prognozować zmienną jakościową. LDA jest jedną z prostszych i starszych metod dyskryminacyjnych. Jej zaletą jest prostota i brak efektu "czarnej skrzynki". Ta prostota jest też wadą. Szczególnie, jeśli zmienna, którą chcemy prognozować uwikłana jest w skomplikowane zależności.

Na LDA można popatrzeć też z innej strony. LDA, podobnie jak analiza składowych głównych (PCA), szuka kombinacji liniowych predyktorów. Podczas jednak, gdy w PCA miały one najlepiej wyjaśniać wariancję predyktorów, tu mają one najlepiej rozdzielać grupy.
Takie podejście do LDA często nazywane jest analizą dyskryminacyjną Fishera (FDA - *Fisher Discriminant Analysis*), a wynikające z niego kombinacje liniowe nazywa się często dyskryminatorami liniowymi.

Ogólne modele analizy dyskryminacyjnej (GDA)

GDA zajmuje się nieliniową analizą dyskryminacyjną przy użyciu operatora funkcji jądra. Podstawowa teoria jest zbliżona do maszyn wektorów nośnych (SVM), o ile metoda GDA zapewnia mapowanie wektorów wejściowych do wielowymiarowej przestrzeni cech. Podobnie jak w przypadku LDA, celem GDA jest znalezienie projekcji dla obiektów w przestrzeni o niższych wymiarach poprzez maksymalizację stosunku rozproszenia między klasami do rozproszenia wewnątrz klasy.

Dziękuję za uwagę