Fisherowska dyskryminacja liniowa – zastosowanie dla nowych danych

Marcin Samojluk, Gabriel Rączkowski

December 12, 2024

Wprowadzenie

Fisherowska dyskryminacja liniowa (LDA) jest techniką statystyczną wykorzystywaną w analizie danych. Jej celem jest maksymalizacja rozróżnienia między klasami w zbiorze danych. Wykorzystywana jest głównie w klasyfikacji.

- LDA znajduje liniową kombinację cech, która najlepiej oddziela klasy.
- Jest szeroko stosowana w wielu dziedzinach, takich jak biologia, rozpoznawanie twarzy czy analiza tekstów.

Historia metody

Fisherowska dyskryminacja liniowa została zaproponowana przez Ronalda A. Fishera w 1936 roku. Początkowo była stosowana w analizie danych botanicznych, aby rozróżnić różne gatunki roślin na podstawie pomiarów ich cech morfologicznych.

Motywacja do stosowania LDA

Wyobraźmy sobie, że mamy dwie klasy (np. Klasa A i Klasa B), które chcemy skutecznie od siebie oddzielić. Każda klasa może być opisana przez wiele różnych cech, takich jak:

- Wzrost, waga,
- Wiek, poziom wykształcenia,
- Wyniki testów itp.

Problem z jedną cechą: Jeśli użyjemy tylko jednej cechy (np. wzrostu), klasy mogą częściowo się nakładać. Nie będziemy w stanie jednoznacznie przypisać danych punktów do jednej z klas.

Przykład: Przykład danych w przestrzeni 1D, gdzie klasy nachodzą na siebie.

Matematyczne podstawy

Fisherowska dyskryminacja liniowa polega na znalezieniu takiej linii (hiperpłaszczyzny), która maksymalizuje stosunek wariancji między klasami do wariancji wewnątrz klas.
Wzór na funkcję celu:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$$

Gdzie:

- S_B - macierz rozrzutu między klasami,
- S_W - macierz rozrzutu wewnątrz klas,
- $\mathbf{w}$ - wektor wag, który określa kierunek linii separującej.

Wyprowadzenie optymalnego $\mathbf{w}$

Celem jest maksymalizacja funkcji celu:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$$

Aby znaleźć optymalny wektor $\mathbf{w}$, liczymy pochodną $\nabla J(\mathbf{w})$ względem $\mathbf{w}$:

$$\nabla J(\mathbf{w}) = 0$$

Rozwijamy krok po kroku w kolejnych slajdach.

Liczenie pochodnej $\nabla J(\mathbf{w})$

Funkcja celu:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$$

Liczymy pochodną z licznika i mianownika:

$$\nabla \left(\mathbf{w}^T S_B \mathbf{w}\right) = 2 S_B \mathbf{w}, \quad \nabla \left(\mathbf{w}^T S_W \mathbf{w}\right) = 2 S_W \mathbf{w}$$

Stosujemy wzór na pochodną ilorazu:

$$\nabla J(\mathbf{w}) = \frac{2 S_B \mathbf{w} \cdot (\mathbf{w}^T S_W \mathbf{w}) - 2 S_W \mathbf{w} \cdot (\mathbf{w}^T S_B \mathbf{w})}{(\mathbf{w}^T S_W \mathbf{w})^2}$$

Szukamy, kiedy $\nabla J(\mathbf{w}) = 0$.

Warunek stacjonarności

Aby znaleźć maksimum funkcji celu, przyrównujemy pochodną do zera:

$$\nabla J(\mathbf{w}) = 0$$

Co oznacza, że:

$$S_B\mathbf{w}(\mathbf{w}^T S_W \mathbf{w}) = S_W\mathbf{w}(\mathbf{w}^T S_B \mathbf{w})$$

Dzielimy obie strony przez $(\mathbf{w}^T S_W \mathbf{w})$:

$$S_B\mathbf{w} = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} S_W \mathbf{w}$$

Definiujemy:

$$\lambda = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$$

Przejście do równania własnego

Otrzymane równanie:

$$S_B \mathbf{w} = \lambda S_W \mathbf{w}$$

Przemnażamy obustronnie przez S_W^{-1} (zakładamy, że S_W jest odwracalna):

$$S_W^{-1} S_B \mathbf{w} = \lambda \mathbf{w}$$

Jest to równanie własne, gdzie:

- $S_W^{-1} S_B$ - macierz,
- λ - wartość własna,
- $\mathbf{w}$ - wektor własny.

Interpretacja równania własnego

- Rozwiązujemy równanie $S_W^{-1} S_B \mathbf{w} = \lambda \mathbf{w}$.
- Wartości własne λ określają zdolność rozróżnienia klas.
- Optymalny wektor $\mathbf{w}$ to ten, który odpowiada największemu λ.

Podsumowanie:

$$\mathbf{w} = \text{wektor własny dla największej } \lambda$$

$$J(\mathbf{w}) \text{ maksymalizowane przy } \lambda_{\text{max}}$$

Rozwiązanie równania

Z równania $\nabla J(\mathbf{w}) = 0$ wynika, że optymalny $\mathbf{w}$ musi spełniać:

$$S_W^{-1} S_B \mathbf{w} = \lambda \mathbf{w}$$

Gdzie λ to wartość własna układu. Rozwiązaniem jest:

- $\mathbf{w}$ - wektor własny macierzy $S_W^{-1} S_B$,
- λ - odpowiadająca wartość własna.

Wybieramy $\mathbf{w}$, które odpowiada największej λ (największemu rozróżnieniu między klasami).

Optymalny wektor **w**

Podsumowanie:

- Macierz $S_W^{-1} S_B$ reprezentuje relację między klasami i zmiennością danych.
- Wektor **w** odpowiada największej wartości własnej macierzy $S_W^{-1} S_B$.
- Ten wektor **w** maksymalizuje rozróżnienie między klasami w projekcji na linię.

Wektory Własne w LDA

W metodzie Linear Discriminant Analysis (LDA), kluczowym zadaniem jest znalezienie odpowiednich kierunków w przestrzeni, które najlepiej oddzielają dane. Te kierunki są reprezentowane przez **wektory własne**.

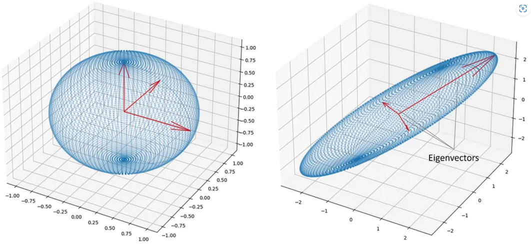
Co to są wektory własne?

- Wektory własne to kierunki, w których dane są najbardziej rozciągnięte, co pomaga je lepiej oddzielić.
- Wartości własne odpowiadają za "intensywność" tego rozciągania w danym kierunku — większa wartość oznacza silniejszą separację.

Dlaczego to ważne?

- Wybieramy te wektory, które mają największe wartości własne, ponieważ pozwalają one uzyskać jak najlepszą separację między klasami.

Obrazek przedstawiający przekształcenie danych



Unity sphere | Transformed unity sphere

Obrazek przedstawia:

- Jak dane przekształcają się w przestrzeni, gdy używamy wektorów własnych.
- Wektory własne pokazują kierunki rozciągania danych, co pomaga w lepszej separacji klas.

Jak działają wektory własne?

W przestrzeni 3D możemy mieć więcej niż jeden wektor własny. W tym przypadku zazwyczaj wybieramy tylko te, które najlepiej rozdzielają klasy.

Jak to działa w praktyce?

- W przestrzeni 3D mamy trzy główne wektory, które wskazują kierunki o największej zmienności danych.
- W LDA, dla dwóch klas, zazwyczaj wystarcza tylko jeden wektor, aby efektywnie oddzielić klasy.
- Dla większej liczby klas możemy wybrać więcej wektorów, ale w podstawowych przypadkach wystarczy jeden.

Obrazek przedstawia:

- Jak dane przekształcają się w przestrzeni, gdy używamy wektorów własnych.
- Przekształcenie pozwala lepiej rozdzielić klasy.

Macierz rozrzutu między klasami S_B

- **Macierz rozrzutu między klasami S_B:** Mierzy, jak różnią się średnie poszczególnych klas w stosunku do globalnej średniej. Celem jest zmaksymalizowanie tej różnicy, aby klasy były jak najbardziej oddzielone.

Wzór na macierz rozrzutu między klasami:

$$S_B = \sum_{i=1}^{k} N_i(\mu_i - \mu)(\mu_i - \mu)^T$$

Gdzie:

- N_i – liczba próbek w klasie i,
- μ_i – średnia klasy i,
- μ – globalna średnia wszystkich próbek.

Macierz rozrzutu wewnątrz klas S_W

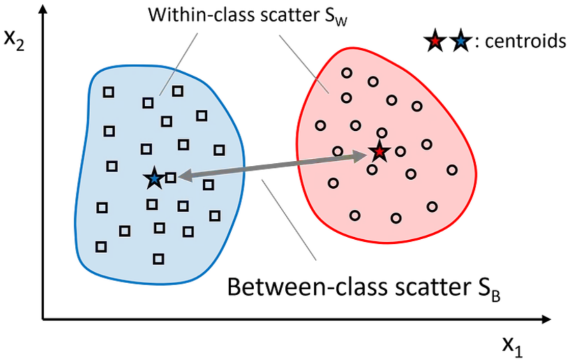
- **Macierz rozrzutu wewnątrz klas S_W:** Mierzy, jak rozproszone są punkty danych w obrębie każdej klasy. Celem jest minimalizacja tej zmienności, aby dane w obrębie każdej klasy były jak najbardziej jednorodne.

Wzór na macierz rozrzutu wewnątrz klas:

$$S_W = \sum_{i=1}^{k} \sum_{x_j \in C_i} (x_j - \mu_i)(x_j - \mu_i)^T$$

Gdzie:

- C_i – zbiór punktów należących do klasy i,
- x_j – pojedyncza próbka w klasie i,
- μ_i – średnia klasy i.



- Intuicyjnie rozproszenie wewnątrz klasy sprawdza, jak zwarta jest każda klasa.
- Rozproszenie między klasami bada, jak daleko od siebie znajdują się różne klasy.

Rozróżnianie między klasami

Fisherowska dyskryminacja liniowa dąży do znalezienia najlepszego wektora **w**, który maksymalizuje rozróżnienie między klasami. Celem jest:

- Projektowanie przestrzeni, w której różnice między klasami są jak najbardziej widoczne.
- Projekcja danych na wektor **w** pozwala na łatwiejsze przypisanie nowych danych do odpowiednich klas.

Przykład zastosowania

Załóżmy, że mamy dane dotyczące roślin i chcemy je sklasyfikować na podstawie cech takich jak długość i szerokość liści.

- Wybieramy dwie cechy (np. długość i szerokość liści).
- Fisherowska dyskryminacja liniowa oblicza najlepszą linię separującą te dwie klasy.

Dzięki tej metodzie możemy łatwo oddzielić klasy roślin na podstawie dwóch prostych cech.

Zastosowanie w rozpoznawaniu twarzy

LDA jest także szeroko stosowane w rozpoznawaniu twarzy, gdzie cechy twarzy (np. odległości między oczami, szerokość nosa) służą do klasyfikacji osób.

- Twarze osób są reprezentowane jako wektory cech.
- LDA znajduje projekcję, która maksymalizuje różnice między twarzami różnych osób.

"Rozpoznawanie twarzy jest jednym z najczęściej stosowanych zastosowań LDA."

Zastosowanie w analizie tekstów

LDA może być również wykorzystywana w analizie tekstów. Na przykład w klasyfikacji e-maili na spam i nie-spam:

- Cechy: obecność słów, długość e-maila, liczba załączników.
- LDA identyfikuje najlepsze cechy, które pozwalają na skuteczną klasyfikację.

„Fisherowska dyskryminacja liniowa jest bardzo efektywna w zadaniach klasyfikacji tekstów.”

Przykład z danymi medycznymi

LDA jest także używane w medycynie, np. w klasyfikacji przypadków chorób:

- Zbieramy dane o pacjentach, np. wyniki badań krwi, ciśnienie.
- LDA pomaga oddzielić pacjentów zdrowych od chorych na podstawie cech medycznych.

Wzór na linię separującą

Jeśli mamy dane 2D, najlepsza linia separująca jest opisana przez wzór:

$$\mathbf{w}^T\mathbf{x} + b = 0$$

Gdzie:

- $\mathbf{w}$ to wektor wag (prostopadły do linii),
- $\mathbf{x}$ to dane (punkt na wykresie),
- b to przesunięcie (odległość od początku układu współrzędnych).

Wnioski

Fisherowska dyskryminacja liniowa jest prostą, ale potężną techniką klasyfikacji. Znajduje szerokie zastosowanie w różnych dziedzinach, takich jak medycyna, analiza twarzy czy analiza tekstów. Choć jest to technika liniowa, może być wystarczająca w wielu praktycznych zastosowaniach.

Bibliografia

- Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". *Annals of Eugenics*.
- Bishop, C. M. (2006). "Pattern Recognition and Machine Learning". *Springer*.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). "An Introduction to Statistical Learning". *Springer*.