Liniowe metody klasyfikacji, część I

Justyna Fijałkowska

Uniwersytet Wrocławski

21 kwietnia 2023

Plan prezentacji

1. Regresja liniowa macierzy indykatorów
   - Klasyfikacja
   - Problem maskowania
2. Liniowa analiza dyskryminacyjna
   - Statystyczna teoria decyzji

Klasyfikacja

Przykłady klasyfikacji:

- Czy u pacjentki rozwinie się rak piersi?

Klasyfikacja

Przykłady klasyfikacji:

- Czy u pacjentki rozwinie się rak piersi?
- Czy użytkownik polubi nowy produkt?

Klasyfikacja

Przykłady klasyfikacji:

- Czy u pacjentki rozwinie się rak piersi?
- Czy użytkownik polubi nowy produkt?
- Kto zostanie nowym prezydentem?

Klasyfikacja binarna

Obserwujemy pary (x_i, y_i) dla $i = 1, 2, \ldots, n$, gdzie y_i oznacza klasę i-tej obserwacji, a $x_i \in \mathbb{R}^p$ to p wymiarowe zmienne niezależne.

- Zbiór etykiet: $\{1, \ldots, K\}$,
- Skonstruowana reguła klasyfikacji: $\hat{f}(x)$

Klasyfikacja binarna

Obserwujemy pary (x_i, y_i) dla $i = 1, 2, \ldots, n$, gdzie y_i oznacza klasę i-tej obserwacji, a $x_i \in \mathbb{R}^p$ to p wymiarowe zmienne niezależne.

- Zbiór etykiet: $\{1, \ldots, K\}$,
- Skonstruowana reguła klasyfikacji: $\hat{f}(x)$

Załóżmy, że mamy dwie klasy ($K = 2$). Jak w takim przypadku użyć regresji liniowej?

Klasyfikacja binarna

Obserwujemy pary (x_i, y_i) dla $i = 1, 2, \ldots, n$, gdzie y_i oznacza klasę i-tej obserwacji, a $x_i \in \mathbb{R}^p$ to p wymiarowe zmienne niezależne.

- Zbiór etykiet: $\{1, \ldots, K\}$,
- Skonstruowana reguła klasyfikacji: $\hat{f}(x)$

Załóżmy, że mamy dwie klasy ($K = 2$). Jak w takim przypadku użyć regresji liniowej?

- Znaleźć współczynniki regresji liniowej rozwiązując problem:

$$\hat{\beta}_0, \hat{\beta} = \text{argmin}_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2$$

Klasyfikacja binarna

Obserwujemy pary (x_i, y_i) dla $i = 1, 2, \ldots, n$, gdzie y_i oznacza klasę i-tej obserwacji, a $x_i \in \mathbb{R}^p$ to p wymiarowe zmienne niezależne.

- Zbiór etykiet: $\{1, \ldots, K\}$,
- Skonstruowana reguła klasyfikacji: $\hat{f}(x)$

Załóżmy, że mamy dwie klasy ($K = 2$). Jak w takim przypadku użyć regresji liniowej?

- Znaleźć współczynniki regresji liniowej rozwiązując problem:

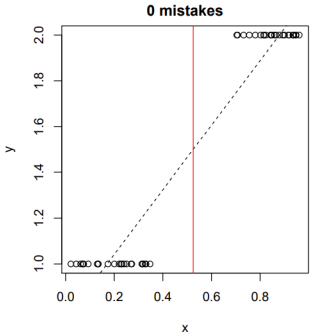
$$\hat{\beta}_0, \hat{\beta} = \text{argmin}_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2.$$

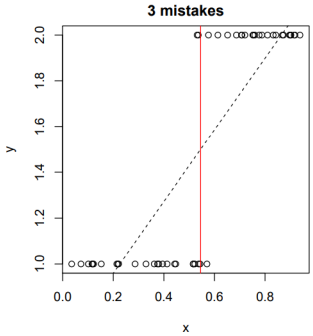
- Dokonać predykcji klasy za pomocą wzoru:

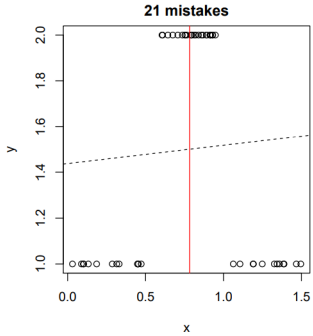
$$\hat{f}^{LS}(x_0) = \begin{cases} 1, & \hat{\beta}_0 + x_0^T \hat{\beta} \leqslant 1.5 \\ 2, & \hat{\beta}_0 + x_0^T \hat{\beta} > 1.5 \end{cases}$$

Klasyfikacja binarna

Przykład:







Klasyfikacja dla dowolnego K

Jak sobie poradzić, gdy mamy więcej niż dwie klasy?

Załóżmy więc, że mamy K klas. Niech $Y \in \mathbb{R}^{n \times K}$ oznaczmy macierz indykatorów, gdzie $Y_{ij} = 1$, jeżeli $y_i = j$ (obserwacja i jest w klasie j) i $Y_{ij} = 0$, gdy $y_i \neq j$ (obserwacja i nie jest w klasie j).

Klasyfikacja dla dowolnego K

Jak sobie poradzić, gdy mamy więcej niż dwie klasy?

Załóżmy więc, że mamy K klas. Przez $Y \in \mathbb{R}^{n \times K}$ oznaczmy macierz indykatorów, gdzie $Y_{ij} = 1$, jeżeli $y_i = j$ (obserwacja i jest w klasie j) i $Y_{ij} = 0$, gdy $y_i \neq j$ (obserwacja i nie jest w klasie j).

Przykład

Dysponujemy pięcioma obserwacjami i trzema klasami. Jak odkodować macierz Y?

$$Y = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Klasyfikacja dla dowolnego K

Jak skonstruować regułę klasyfikacyjną?

- Regresujemy każdą kolumnę $Y_j \in \mathbb{R}^n$ na predyktory:

$$\hat{\beta}_{j,0}, \hat{\beta}_j = \text{argmin}_{\beta_{j,0} \in \mathbb{R}, \beta_j \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_{ij} - \beta_{0,j} - \beta_j^T x_i)^2$$

Klasyfikacja dla dowolnego K

Jak skonstruować regułę klasyfikacyjną?

- Regresujemy każdą kolumnę $Y_j \in \mathbb{R}^n$ na predyktory:

$$\hat{\beta}_{j,0}, \hat{\beta}_j = \text{argmin}_{\beta_{j,0} \in \mathbb{R}, \beta_j \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_{ij} - \beta_{0,j} - \beta_j^T x_i)^2$$

- Dla nowego wektora wejście $x_0 \in \mathbb{R}^p$ obliczamy:

$$\hat{\beta}_{0,j} + x_0^T \hat{\beta}_j, \quad j = 1, \ldots, K$$

Klasyfikacja dla dowolnego K

Jak skonstruować regułę klasyfikacyjną?

- Regresujemy każdą kolumnę $Y_j \in \mathbb{R}^n$ na predyktory:

$$\hat{\beta}_{j,0}, \hat{\beta}_j = \text{argmin}_{\beta_{j,0} \in \mathbb{R}, \beta_j \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_{ij} - \beta_{0,j} - \beta_j^T x_i)^2$$

- Dla nowego wektora wejście $x_0 \in \mathbb{R}^p$ obliczamy:

$$\hat{\beta}_{0,j} + x_0^T \hat{\beta}_j, \quad j = 1, \ldots, K$$

- Wybieramy klasę, która rozwiązuje problem:

$$\hat{f}^{LS}(x_0) = \text{argmax}_{j=1,\ldots,K} \hat{\beta}_{0,j} + x_0^T \hat{\beta}_j$$

Klasyfikacja dla dowolnego K

Jak skonstruować regułę klasyfikacyjną?

- Regresujemy każdą kolumnę $Y_j \in \mathbb{R}^n$ na predyktory:

$$\hat{\beta}_{j,0}, \hat{\beta}_j = \text{argmin}_{\beta_{j,0} \in \mathbb{R}, \beta_j \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_{ij} - \beta_{0,j} - \beta_j^T x_i)^2$$

- Dla nowego wektora wejście $x_0 \in \mathbb{R}^p$ obliczamy:

$$\hat{\beta}_{0,j} + x_0^T \hat{\beta}_j, \quad j = 1, \ldots, K$$

- Wybieramy klasę, która rozwiązuje problem:

$$\hat{f}^{LS}(x_0) = \text{argmax}_{j=1,\ldots,K} \hat{\beta}_{0,j} + x_0^T \hat{\beta}_j$$

Granice decyzyjności pomiędzy dowolnymi klasami j, k :

$$\hat{\beta}_{0,j} + x^T \hat{\beta}_j = \hat{\beta}_{0,k} + x^T \hat{\beta}_k$$

Klasyfikacja dla dowolnego K

Jak skonstruować regułę klasyfikacyjną?

- Regresujemy każdą kolumnę $Y_j \in \mathbb{R}^n$ na predyktory:

$$\hat{\beta}_{j,0}, \hat{\beta}_j = \text{argmin}_{\beta_{j,0} \in \mathbb{R}, \beta_j \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_{ij} - \beta_{0,j} - \beta_j^T x_i)^2$$

- Dla nowego wektora wejście $x_0 \in \mathbb{R}^p$ obliczamy:

$$\hat{\beta}_{0,j} + x_0^T \hat{\beta}_j, \quad j = 1, \ldots, K$$

- Wybieramy klasę, która rozwiązuje problem:

$$\hat{f}^{LS}(x_0) = \text{argmax}_{j=1,\ldots,K} \hat{\beta}_{0,j} + x_0^T \hat{\beta}_j$$

Granice decyzyjności pomiędzy dowolnymi klasami j, k :

$$\hat{\beta}_{0,j} + x^T \hat{\beta}_j = \hat{\beta}_{0,k} + x^T \hat{\beta}_k \iff \hat{\beta}_{0,j} - \hat{\beta}_{0,k} + (\hat{\beta}_j - \hat{\beta}_k)^T x = 0$$

Klasyfikacja dla dowolnego K

Jakie są przesłanki takiego podejścia?

Postrzeganie regresji jako estymacji warunkowej wartości oczekiwanej. Dla zmiennej losowej Y_k, $E(Y_k|X = x) = P(G = k|X = x)$.

Klasyfikacja dla dowolnego K

Jakie są przesłanki takiego podejścia?

Postrzeganie regresji jako estymacji warunkowej wartości oczekiwanej. Dla zmiennej losowej Y_k, $E(Y_k|X = x) = P(G = k|X = x)$.

Uwagi:

- Jeżeli model zawiera intercept, to $\sum_{k \in \mathcal{G}} \hat{f}_k(x) = 1$ dla dowolnego x.

Klasyfikacja dla dowolnego K

Jakie są przesłanki takiego podejścia?

Postrzeganie regresji jako estymacji warunkowej wartości oczekiwanej. Dla zmiennej losowej Y_k, $E(Y_k|X = x) = P(G = k|X = x)$.

Uwagi:

- Jeżeli model zawiera intercept, to $\sum_{k \in \mathcal{G}} \hat{f}_k(x) = 1$ dla dowolnego x.
- Jeśli dopuścimy regresję liniową na rozwinięciach bazowych $h(X)$ danych wejściowych, to podejście może prowadzić do zgodnych estymacji prawdopodobieństwa.

Inne spojrzenie na problem

- Skonstruowanie celów t_k dla każdego klasy.
- Zakładamy, że wektor zmiennych zależnych y_i, czyli i-ty wiersz macierzy Y dla obserwacji i przyjmuje wartość t_k, jeżeli $g_i = k$.
- Dopasowujemy model liniowy metodą najmniejszych kwadratów:

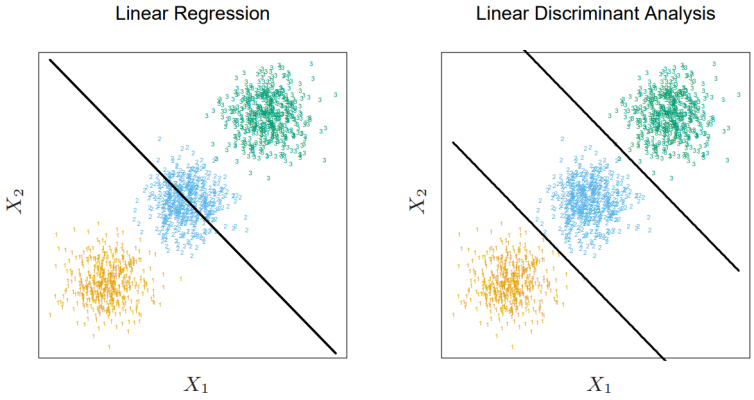
$$\min_B \sum_{i=1}^{N} \|y_i - [1, x_i^T]^T B^T\|^2.$$

- Klasyfikujemy przy użyciu funkcji:

$$\hat{G}(x) = \text{argmax}_k \|\hat{f}(x) - t_k\|^2.$$

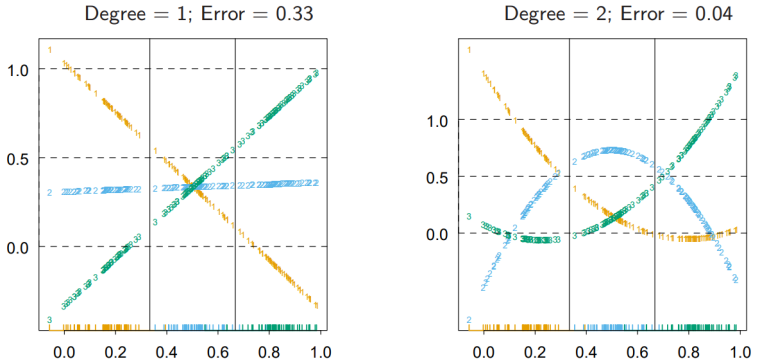
Problem maskowania

Problem maskowania pojawia się, gdy $K \geqslant 3$.



Problem maskowania

Skąd wynika problem maskowania?



Statystyczna teoria decyzji

Niech C będzie zmienną losową, wtedy klasyfikacji dokonujemy zgodnie z regułą:

$$f(x) = \text{argmax}_{j=1,\ldots,K} P(C = j | X = x).$$

Statystyczna teoria decyzji

Niech C będzie zmienną losową, wtedy klasyfikacji dokonujemy zgodnie z regułą:

$$f(x) = \text{argmax}_{j=1,\ldots,K} P(C = j | X = x).$$

Rozpisując prawdopodobieństwo warunkowe ze wzoru Bayes'a dostajemy:

$$P(C = j | X = x) = \frac{P(X = x | C = j) P(C = j)}{P(X = x)}.$$

Statystyczna teoria decyzji

Niech C będzie zmienną losową, wtedy klasyfikacji dokonujemy zgodnie z regułą:

$$f(x) = \text{argmax}_{j=1,\ldots,K} P(C = j | X = x).$$

Rozpisując prawdopodobieństwo warunkowe ze wzoru Bayes'a dostajemy:

$$P(C = j | X = x) = \frac{P(X = x | C = j) P(C = j)}{P(X = x)}.$$

Niech $\pi_j = P(C = j)$. Wówczas regułę tę można przekształcić do postaci:

$$f(x) = \text{argmax}_{j=1,\ldots,K} P(X = x | C = j) \cdot \pi_j.$$

Liniowa analiza dyskryminacyjna

Zakładamy, że dane przynależące do każdej klasy mają rozkład normalny:

$$h_j(x) = P(X = x|C = j) = \text{gęstość z rozkładu } N(\mu_j, \Sigma),$$

Liniowa analiza dyskryminacyjna

Zakładamy, że dane przynależące do każdej klasy mają rozkład normalny:

$$h_j(x) = P(X = x|C = j) = \text{gęstość z rozkładu } N(\mu_j, \Sigma),$$

co można również zapisać w postaci:

$$h_j(x) = \frac{1}{(2\pi)^{p/2}\det(\Sigma)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T\Sigma^{-1}(x-\mu_j)\right\}$$

Liniowa analiza dyskryminacyjna

Zakładamy, że dane przynależące do każdej klasy mają rozkład normalny:

$$h_j(x) = P(X = x|C = j) = \text{gęstość z rozkładu } N(\mu_j, \Sigma),$$

co można również zapisać w postaci:

$$h_j(x) = \frac{1}{(2\pi)^{p/2}\det(\Sigma)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T\Sigma^{-1}(x-\mu_j)\right\}$$

Uwaga: Zakładamy, że macierz kowariancji w każdej klasie jest taka sama, lecz dopuszczamy różne wektory średnich.

Liniowa analiza dyskryminacyjna

Cel: znalezienie takiego j dla którego wyrażenie poniżej będzie największe

$$P(X = x|C = j) \cdot \pi_j = h_j(x) \cdot \pi_j$$

Liniowa analiza dyskryminacyjna

Cel: znalezienie takiego j, dla którego wyrażenie poniżej będzie największe

$$P(X = x|C = j) \cdot \pi_j = h_j(x) \cdot \pi_j$$

Logarytm jest funkcją monotoniczną, zatem możemy problem sprowadzić do szukania j, dla którego $\log(h_j(x)\pi_j)$ będzie największe.

Liniowa analiza dyskryminacyjna

Cel: znalezienie takiego j, dla którego wyrażenie poniżej będzie największe

$$P(X = x|C = j) \cdot \pi_j = h_j(x) \cdot \pi_j$$

Logarytm jest funkcją monotoniczną, zatem możemy problem sprowadzić do szukania j, dla którego $\log(h_j(x)\pi_j)$ będzie największe.

Reguła decyzyjna będzie postaci:

$$f^{LDA}(x) = \text{argmax}_{j=1,\ldots,K}\delta_j(x), \quad \text{gdzie} \quad \delta_j(x) = x^T\Sigma^{-1}\mu_j - \frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j + \log \pi_j.$$

Liniowa analiza dyskryminacyjna

Problem: Zazwyczaj nie znamy π_j, Σ oraz μ_j.

Liniowa analiza dyskryminacyjna

Problem: Zazwyczaj nie znamy π_j, Σ oraz μ_j.

Pomysł: Estymujemy je na danych treningowych $x_i \in \mathbb{R}^p$ i $y_i \in \{1, 2, \ldots, K\}$ korzystając ze wzorów (n_j oznacza liczbę punktów w klasie j):

- $\hat{\pi}_j = n_j/n$, proporcja obserwacji w klasie j,
- $\hat{\mu}_j = \frac{1}{n_j} \sum_{y_i=j} x_i$, centroidy j-tej klasy,
- $\hat{\Sigma} = \frac{1}{n-K} \sum_{j=1}^{K} \sum_{y_i=j} (x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T$, próbkowa macierz kowariancji,

Liniowa analiza dyskryminacyjna

Problem: Zazwyczaj nie znamy π_j, Σ oraz μ_j.

Pomysł: Estymujemy je na danych treningowych $x_i \in \mathbb{R}^p$ i $y_i \in \{1, 2, \ldots, K\}$ korzystając ze wzorów (n_j oznacza liczbę punktów w klasie j):

- $\hat{\pi}_j = n_j/n$, proporcja obserwacji w klasie j,
- $\hat{\mu}_j = \frac{1}{n_j} \sum_{y_i=j} x_i$, centroidy j-tej klasy,
- $\hat{\Sigma} = \frac{1}{n-K} \sum_{j=1}^{K} \sum_{y_i=j} (x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T$, próbkowa macierz kowariancji,

Estymowana funkcja dyskryminacyjna będzie postaci:

$$\hat{\delta}_j = x^T \hat{\Sigma}^{-1} \hat{\mu}_j - \frac{1}{2} \hat{\mu}_j^T \hat{\Sigma}^{-1} \hat{\mu}_j + \log \hat{\pi}_j,$$

a reguła decyzyjna:

$$\hat{f}^{LDA}(x) = \text{argmax}_{j=1,\ldots,K} \hat{\delta}_j(x).$$

Liniowa analiza dyskryminacyjna

Estymowana funkcja dyskryminacyjna będzie postaci:

$$\hat{\delta}_j = x^T \hat{\Sigma}^{-1} \hat{\mu}_j - \frac{1}{2} \hat{\mu}_j^T \hat{\Sigma}^{-1} \hat{\mu}_j + \log \hat{\pi}_j,$$

a reguła decyzyjna:

$$\hat{f}^{LDA}(x) = \text{argmax}_{j=1,\ldots,K} \hat{\delta}_j(x).$$

Uwaga: Jeżeli nie będziemy zakładać, iż macierze kowariancji są takie same dla każdej klasy, to zmianie ulegnie funkcja dyskryminacyjna:

$$\delta_k(x) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log \pi_k.$$

Dziękuję za uwagę.