Modele dyskryminacyjne i generatywne

Systemy uczące się - laboratorium

Mateusz Lango

Zakład Inteligentnych Systemów Wspomagania Decyzji
Wydział Informatyki i Telekomunikacji
Politechnika Poznańska

„Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”,
projekt finansowany ze środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20

Fundusze Europejskie Polska Cyfrowa

Rzeczpospolita Polska

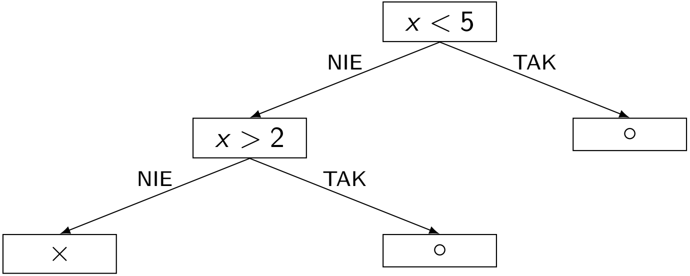
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Na ostatnich zajęciach: powtórka drzew decyzyjnych i zasada ERM

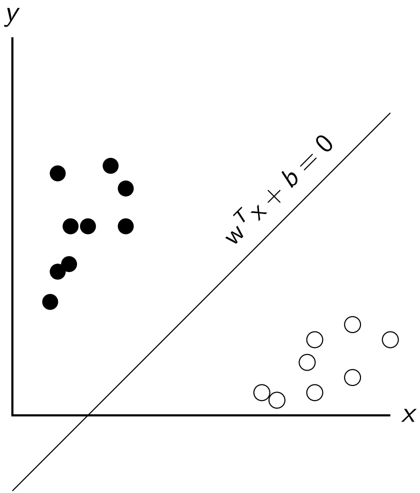
Zasada minimalizacji ryzyka empirycznego

$$\hat{f} = \arg\min_{g \in \mathcal{H}} \frac{1}{n} \sum_{\textit{dane uczace}} [L(g(X), Y)]$$

gdzie L to dla klasyfikacji zwykle błąd zero-jedynkowy.



Nowy pomysł na reprezentację wiedzy: klasyfikator liniowy



- Reprezentacja wiedzy w postaci nauczonych wag wyrażenia:

$$g(x) = w^T x + b$$

- Decyzję podejmuje się poprzez porównanie wartości $g(x)$ z pewnym ustalonym (zwykle: nieuczonym) progiem

$$\hat{y} = \begin{cases} 1 & g(x) > 0 \\ 0 & g(x) \leq 0 \end{cases}$$

Klasyfikator liniowy – jak się uczyć?

- Zasada minimalizacji ryzyka empirycznego (ang. *empirical risk minimization, ERM*):

$$\hat{f} = \arg\min_{g \in \mathcal{H}} \frac{1}{n} \sum_{\textit{dane uczace}} [L(g(X), Y)]$$

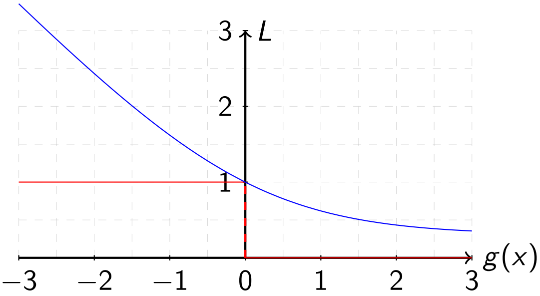
  gdzie L to błąd zero-jedynkowy.

- Klasa hipotez $\mathcal{H}$ zawiera wszystkie możliwe wektory wag

$$\mathcal{H} = \{w : w \in \mathbb{R}^{d+1}\}$$

- ERM w tej sytuacji jest NP-trudny, a także NP-trudne jest uzyskanie aproksymacji do pewnego stałego czynnika

A gdyby tak...



- Możemy spróbować zastąpić w algorytmie błąd 0-1 zastępczą funkcją straty, która będzie go z góry ograniczała
- Zastępczą funkcję straty wybieramy tak aby była wypukła, przez co prosta w optymalizacji
- W przyszłości: od zastępczej funkcji straty zwykle wymagamy dodatkowych własności teoretycznych np. własność kalibracji[a]
- Dzisiaj: zobaczymy, że takie funkcje naturalnie pojawiają się przy projektowaniu algorytmów **metodami statystycznymi**.

[a] patrz przedmiot: „Teoria uczenia maszynowego”

Modele generatywne

- Zadanie „uczenia się z danych” możemy przeformułować na zadanie *estymowania* rozkładu danych $P(\vec{x}, y)$ (gdzie $\vec{x}$ jest wektorem cech)
- Znając rozkład danych można:
  - wygenerować/dolosować więcej danych
  - poznać zależności między cechami i uzyskać ich rozkłady
  - skonstruować rozkład warunkowy:

    $$P(y|\vec{x}) = \frac{P(\vec{x}, y)}{P(\vec{x})} = \frac{P(\vec{x}, y)}{\sum_y P(\vec{x}, y)}$$

    i wybierając najbardziej prawdopodobną klasę uzyskać klasyfikator.
  - korzystać z innych zalet przetwarzania rozkładów prawdopodobieństwa
- Zadanie estymacji rozkładu możemy rozwiązywać na wiele sposób: gotowe rozwiązania w statystyce

Zasada maksymalnej wiarygodności (MLE)

- Zasada maksymalnej wiarygodności wybiera/estymuje/uczy się parametrów rozkładu poprzez wybranie takich wartości które maksymalizują funkcje wiarygodności:

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\, L(\theta\,; \mathrm{x})$$

gdzie (upraszając) $L(\theta\,; \mathrm{x})$ to prawdopodobieństwo uzyskania próbki wg. modelu o parametrach θ.

- W praktyce maksymalizujemy logarytm z tej funkcji
- Dla modeli generatywnych:

$$\max \ln P(\vec{X}, \vec{y}) \quad \Rightarrow \quad \max \ln \prod_{i=1}^{n} P(\vec{x}_i, y_i) \quad \Rightarrow \quad \max \sum_{i=1}^{n} \ln P(\vec{x}_i, y_i)$$

Przykład obliczeń: uproszczony „Play golf?” [Quinlan ’86]

Outlook	Windy	Play?
słonecznie	false	$\circ$
słonecznie	true	$\circ$
pochmurnie	false	$\star$
deszcz	false	$\star$
deszcz	false	$\star$
deszcz	true	$\circ$
pochmurnie	true	$\star$
słonecznie	false	$\circ$
słonecznie	false	$\star$
deszcz	false	$\star$
słonecznie	true	$\star$
pochmurnie	true	$\star$
pochmurnie	false	$\star$
deszcz	true	$\circ$

Problem

Wyestymuj łączny rozkład prawdopodobieństwa tych danych, zgodnie z zasadą maksymalnej wiarygodności, a następnie oblicz prawdopodobieństwo, że $P(y = \star | \textit{słonecznie}, \textit{false})$.
Do jakiej klasy zostałby przydzielony przykład (słonecznie, false) wg. klasyfikatora skonstruowanego na tym prawdopodobieństwie?

Rozwiązanie 1: założenie o normalności danych

Klątwa wymiarowości to zespół zjawisk polegający na wykładniczym wzroście trudności rozwiązywanego problemu uczenia się w zależności od wymiaru przestrzeni.

- Rozkład łączny $P(\vec{x}, y)$ możemy rozbić na dwie składowe korzystając z reguły łańcuchowej

$$P(\vec{x}, y) = P(\vec{x}|y)P(y)$$

- Zakładając rozkład normalny cech pod warunkiem klasy:

$$P(\vec{x}, y) = N(\vec{x}|\vec{\mu}_y, \Sigma_y)P(y)$$

$$N(\vec{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^\top \Sigma^{-1}(\vec{x} - \vec{\mu})\right)$$

- Dla każdej klasy potrzebujemy d średnich (dla każdej cechy) oraz kowariancje między parami cech (rzędu d^2)

Problem

Dlaczego zakładać akurat normalny rozkład danych?

Zadanie

Problem

Zakładając rozkład normalny cech pod warunkiem klasy:

$$P(\vec{x}, y) = N(\vec{x}|\vec{\mu}_y, \Sigma_y)P(y)$$

dokonaj estymacji tego klasyfikatora zgodnie z zasadą maksymalnej wiarygodności, a następnie podaj wyrażenie na $P(+|x_1 = 3, x_2 = 0)$.

x_1	x_2	y
1	-2	+
2	0	+
3	2	+
1	5	-
1	-5	-

Liniowa analiza dyskryminacyjna

- Ze względu na liczbę parametrów potrzebną do estymowania macierzy kowariancji, często wprowadzamy założenie że macierz kowariancji jest *wspólna* dla każdej klasy

$$P(\vec{x}, y) = N(\vec{x}|\vec{\mu}_y, \Sigma)P(y)$$

- Dla każdej klasy potrzebujemy d średnich (dla każdej cechy) oraz dodatkowo jedną macierz kowariancji (rzędu d^2)
- Klasyfikator uzyskany z tak wyrażonego rozkładu nazywamy liniową analizą dyskryminacyjną (LDA)[1]

[1] Zwróć uwagę, że na wykładzie jest ona wyprowadzana jako metoda nadzorowanej redukcji wymiarowości. Prowadzi to również do efektywniejszej implementacji o ile nie interesują nas prawdopodobieństwa, a jedynie uzyskanie predyktora.

Liniowa analiza dyskryminacyjna - estymacja

- Zgodnie z założeniami LDA:

$$P(\vec{x}, y) = N(\vec{x}|\vec{\mu}_y, \Sigma)P(y)$$

- Zgodnie z zasadą maksymalnej wiarygodności:

$$\max \sum_{i=1}^{n} \ln P(\vec{x}_i, y_i) = \sum_{i=1}^{n} \ln N(\vec{x}|\vec{\mu}_y, \Sigma)P(y)$$

- Przyrównując pochodną do 0 otrzymujemy:
  - $P(y) = \frac{n_y}{n}$ – liczba przykładów z klasy y podzielić przez liczbę wszystkich przykładów
  - $\mu_k = \frac{1}{n_k} \sum_{y_i = k} x_i$ dla każdej klasy k
  - $\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{y_i})(x_i - \mu_{y_i})^T$ (tylko 1, współdzielony między klasami) [2]

[2] Standardowo korzystamy z nieobciążonych estymatorów $\Sigma = \frac{1}{n-|C|} \sum_{k=1}^{|C|} \sum_{y_i=k} (x_i - \mu_{y_i})(x_i - \mu_{y_i})^T$ gdzie $|C|$ oznacza liczbę klas

Rozwiązanie 2: założenie o warunkowej niezależności cech

- Rozkład łączny $P(\vec{x}, y)$ możemy rozbić na dwie składowe korzystając z reguły łańcuchowej

$$P(\vec{x}, y) = P(\vec{x}|y)P(y)$$

- Wprowadzając założenie o warunkowej niezależności cech

$$P(\vec{x}, y) = P(\vec{x}|y)P(y) = P(x_1, x_2, \ldots, x_d|y)P(y) = P(x_1|y)P(x_2|y)\cdots P(x_n|y)P(y)$$

- Przekształcając na rozkład warunkowy uzyskujemy:

$$P(y|\vec{x}) = \frac{P(\vec{x}, y)}{\sum_y P(\vec{x}, y)} = \frac{\prod_{i=1}^{d} P(x_i|y)P(y)}{\sum_y \prod_{i=1}^{d} P(x_i|y)P(y)}$$

- W literaturze uczenia maszynowego ten model nazywamy klasyfikatorem naiwnego Bayesa, w literaturze statystycznej...

Rozwiązanie 2: założenie o warunkowej niezależności cech

- Rozkład łączny $P(\vec{x}, y)$ możemy rozbić na dwie składowe korzystając z reguły łańcuchowej

$$P(\vec{x}, y) = P(\vec{x}|y)P(y)$$

- Wprowadzając założenie o warunkowej niezależności cech

$$P(\vec{x}, y) = P(\vec{x}|y)P(y) = P(x_1, x_2, \ldots, x_d|y)P(y) = P(x_1|y)P(x_2|y)\cdots P(x_n|y)P(y)$$

- Przekształcając na rozkład warunkowy uzyskujemy:

$$P(y|\vec{x}) = \frac{P(\vec{x}, y)}{\sum_y P(\vec{x}, y)} = \frac{\prod_{i=1}^{d} P(x_i|y)P(y)}{\sum_y \prod_{i=1}^{d} P(x_i|y)P(y)}$$

- W literaturze uczenia maszynowego ten model nazywamy klasyfikatorem naiwnego Bayesa, w literaturze statystycznej...

Rozwiązanie 2: założenie o warunkowej niezależności cech

- Rozkład łączny $P(\vec{x}, y)$ możemy rozbić na dwie składowe korzystając z reguły łańcuchowej

$$P(\vec{x}, y) = P(\vec{x}|y)P(y)$$

- Wprowadzając założenie o warunkowej niezależności cech

$$P(\vec{x}, y) = P(\vec{x}|y)P(y) = P(x_1, x_2, \dots, x_d|y)P(y) = P(x_1|y)P(x_2|y)\cdots P(x_n|y)P(y)$$

- Przekształcając na rozkład warunkowy uzyskujemy:

$$P(y|\vec{x}) = \frac{P(\vec{x}, y)}{\sum_y P(\vec{x}, y)} = \frac{\prod_{i=1}^d P(x_i|y)P(y)}{\sum_y \prod_{i=1}^d P(x_i|y)P(y)}$$

- W literaturze uczenia maszynowego ten model nazywamy klasyfikatorem naiwnego Bayesa, w literaturze statystycznej...

Naiwny Bayes - estymacja

- Zgodnie z założeniami NB:

$$P(\vec{x}, y) = \prod_{i=1}^{d} P(x_i|y)P(y)$$

- Zgodnie z zasadą maksymalnej wiarygodności:

$$\max \sum_{i=1}^{n} \ln P(\vec{x}_i, y_i) = \sum_{i=1}^{n} \ln \prod_{i=1}^{d} P(x_i|y)P(y) = \sum_{i=1}^{n} \left(\ln P(y) + \sum_{i=1}^{d} \ln P(x_i|y) \right)$$

- Przyrównując pochodną do 0 otrzymujemy:
  - $P(y) = \frac{n_y}{n}$ – liczba przykładów z klasy y podzielić przez liczbę wszystkich przykładów
  - Dla cech estymatory założonych (jednowymiarowych) rozkładów np.
    - dla cech ciągłych i rozkładu normalnego: zwykła średnia i wariancja
    - dla cech nominalnych i rozkładu kategorycznego: zwykłe zliczanie

Przykład obliczeń: uproszczony „Play golf?” [Quinlan ’86]

Outlook	Windy	Play?
słonecznie	false	$\circ$
słonecznie	true	$\circ$
pochmurnie	false	$\star$
deszcz	false	$\star$
deszcz	false	$\star$
deszcz	true	$\circ$
pochmurnie	true	$\star$
słonecznie	false	$\circ$
słonecznie	false	$\star$
deszcz	false	$\star$
słonecznie	true	$\star$
pochmurnie	true	$\star$
pochmurnie	false	$\star$
deszcz	true	$\circ$

Problem

Wyznacz prawdopodobieństwo $P(y = \star | \mathit{słonecznie}, \mathit{false})$ zgodnie z klasyfikatorem naiwnego Bayesa, a następnie odpowiedz na pytania:

- *Jakich rozkładów prawdopodobieństwa nie możemy się nauczyć (tj. zamodelować)?*
- *Czy dostrzegasz jakieś wady zaproponowanego podejścia?*
- *Ile parametrów ma ten klasyfikator?*
- *Zakładając klasyfikację binarną i d cech binarnych, podaj wzór na liczbę parametrów tego klasyfikatora.*

Modele generatywne - podsumowanie

- Modelują rozkład łączny danych
- Odtwarzają proces generowania danych np. „najpierw wybrano dla przykładu klasę z $P(y)$, a potem wygenerowano cechy z rozkładu normalnego tej klasy"
- Klasyfikator uzyskujemy poprzez przekształcenie do rozkładu warunkowego np. regułą Bayesa
  - W zasadzie nic nie stoi na przeszkodzie by przekształcić rozkład tak by przewidywał dowolną zmienną x_i zamiast y!

  $\Rightarrow$ Czy nie rozwiązujemy trudniejszego problemu żeby rozwiązać prostszy?

Modele generatywne - podsumowanie

- Modelują rozkład łączny danych
- Odtwarzają proces generowania danych np. „najpierw wybrano dla przykładu klasę z $P(y)$, a potem wygenerowano cechy z rozkładu normalnego tej klasy”
- Klasyfikator uzyskujemy poprzez przekształcenie do rozkładu warunkowego np. regułą Bayesa
  - W zasadzie nic nie stoi na przeszkodzie by przekształcić rozkład tak by przewidywał dowolną zmienną x_i zamiast y!

  $\Rightarrow$ Czy nie rozwiązujemy trudniejszego problemu żeby rozwiązać prostszy?

Modele dyskryminacyjne

- Wyestymujmy bezpośrednio (i tylko) rozkład $P(y|\vec{x})$
- Zgodnie z zasadą maksymalnej wiarygodności:

$$\max \sum_{i=1}^{n} \ln P(y_i|\vec{x}_i)$$

- Dla problemu regresji możemy założyć, że $P(y|\vec{x})$ jest jednowymiarowym (!) rozkładem normalnym, a jego średnią możemy wyznaczyć wyrażeniem liniowym $\Rightarrow$ typowa regresja liniowa

$$P(y|\vec{x}) = N(y|\mu_x = w^T x + b, \sigma)$$

- Dla problemu klasyfikacji binarnej y jest zmienną $0/1$, co w naturalny sposób prowadzi nas do rozkładu Bernoulliego

$$P(y|\vec{x}) = B(y|p_x = ?w^T x + b?) = p_x^y (1 - p_x)^{(1-y)}$$

Modele dyskryminacyjne

- Wyestymujmy bezpośrednio (i tylko) rozkład $P(y|\vec{x})$
- Zgodnie z zasadą maksymalnej wiarygodności:

$$\max \sum_{i=1}^{n} \ln P(y_i|\vec{x}_i)$$

- Dla problemu regresji możemy założyć, że $P(y|\vec{x})$ jest jednowymiarowym (!) rozkładem normalnym, a jego średnią możemy wyznaczyć wyrażeniem liniowym $\Rightarrow$ typowa regresja liniowa

$$P(y|\vec{x}) = N(y|\mu_x = w^T x + b, \sigma)$$

- Dla problemu klasyfikacji binarnej y jest zmienną 0/1, co w naturalny sposób prowadzi nas do rozkładu Bernoulliego

$$P(y|\vec{x}) = B(y|p_x = ?w^T x + b?) = p_x^y (1 - p_x)^{(1-y)}$$

Modele dyskryminacyjne

- Wyestymujmy bezpośrednio (i tylko) rozkład $P(y|\vec{x})$
- Zgodnie z zasadą maksymalnej wiarygodności:

$$\max \sum_{i=1}^{n} \ln P(y_i|\vec{x}_i)$$

- Dla problemu regresji możemy założyć, że $P(y|\vec{x})$ jest jednowymiarowym (!) rozkładem normalnym, a jego średnią możemy wyznaczyć wyrażeniem liniowym $\Rightarrow$ typowa regresja liniowa

$$P(y|\vec{x}) = N(y|\mu_x = w^T x + b, \sigma)$$

- Dla problemu klasyfikacji binarnej y jest zmienną $0/1$, co w naturalny sposób prowadzi nas do rozkładu Bernoulliego

$$P(y|\vec{x}) = B(y|p_x =?w^T x + b?) = p_x^y (1 - p_x)^{(1-y)}$$

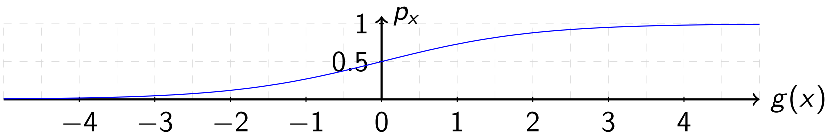
Regresja logistyczna

Załóżmy, że logarytm z szansy jest funkcją liniową[3]:

$$\mathrm{logit}(p_x) = \ln \frac{p_x}{1 - p_x} = w^T x + b$$

Wyrażając prawdopodobieństwo p_x w zależności od wyniku wyrażenia liniowego $g(x) = w^T x + b$ otrzymujemy:

$$p_x = \frac{1}{1 + e^{-g(x)}}$$



[3]Na przykład: jeśli dane mają warunkowo rozkład normalny ze współdzieloną macierzą kowariancji to tak rzeczywiście jest.

Zadanie

Problem

Próbuje się zamodelować prawdopodobieństwo ataku cybernetycznego w danym dniu przy użyciu liczby ataków z dnia poprzedniego. Otrzymano następujący model regresji logistycznej o współczynnikach $b = 0.5$ *oraz* $w = 0.1$*. Ile wynosi prawdopodobieństwo ataku, jeżeli wczoraj było ich 5?*

Regresja logistyczna - estymacja

Zgodnie z zasadą maksymalnej wiarygodności:

$$\max_{w,b} \sum_{i=1}^{n} \ln P(y_i|\vec{x}_i) = \sum_{i=1}^{n} \ln \left[p_{x_i}^{y_i} (1 - p_{x_i})^{(1-y_i)} \right]$$

$$= \sum_{i=1}^{n} \left(\ln p_{x_i}^{y_i} + \ln (1 - p_{x_i})^{(1-y_i)} \right)$$

$$= \sum_{i=1}^{n} \left(y_i \ln p_{x_i} + (1 - y_i) \ln (1 - p_{x_i}) \right)$$

gdzie $p_x = \frac{1}{1+e^{-(w^T x + b)}}$

Regresja logistyczna - estymacja

Zgodnie z zasadą maksymalnej wiarygodności:

$$\max_{w,b} \sum_{i=1}^{n} \ln P(y_i|\vec{x}_i) = \sum_{i=1}^{n} \ln \left[p_{x_i}^{y_i} (1 - p_{x_i})^{(1-y_i)} \right]$$

$$= \sum_{i=1}^{n} \left(\ln p_{x_i}^{y_i} + \ln (1 - p_{x_i})^{(1-y_i)} \right)$$

$$= \sum_{i=1}^{n} \left(y_i \ln p_{x_i} + (1 - y_i) \ln (1 - p_{x_i}) \right)$$

gdzie $p_x = \frac{1}{1+e^{-(w^T x+b)}}$

Regresja logistyczna - estymacja

Zgodnie z zasadą maksymalnej wiarygodności:

$$\begin{aligned}
\max_{w,b} \sum_{i=1}^{n} \ln P(y_i|\vec{x}_i) &= \sum_{i=1}^{n} \ln \left[p_{x_i}^{y_i} (1 - p_{x_i})^{(1-y_i)} \right] \\
&= \sum_{i=1}^{n} \left(\ln p_{x_i}^{y_i} + \ln(1 - p_{x_i})^{(1-y_i)} \right) \\
&= \sum_{i=1}^{n} \left(y_i \ln p_{x_i} + (1 - y_i) \ln(1 - p_{x_i}) \right)
\end{aligned}$$

gdzie $p_x = \frac{1}{1+e^{-(w^T x+b)}}$

Błąd logistyczny (*)

Przekształcając wzór dalej otrzymujemy:

$$
\begin{aligned}
\max_{w,b} \sum_{i=1}^{n} (y_i \ln p_{x_i} + (1 - y_i) \ln(1 - p_{x_i})) = \\
&= \sum_{i=1}^{n} \left(y_i \ln \frac{1}{1 + e^{-g(x)}} + (1 - y_i) \ln(1 - \frac{1}{1 + e^{-g(x)}}) \right) \\
&= \sum_{i=1}^{n} \left(y_i \ln \frac{1}{1 + e^{-g(x)}} + (1 - y_i) \ln(\frac{1}{1 + e^{g(x)}}) \right) \\
&= - \sum_{i=1}^{n} \left(y_i \ln(1 + e^{-g(x)}) + (1 - y_i) \ln(1 + e^{g(x)}) \right) \\
&= - \sum_{i=1}^{n} \ln \left(1 + e^{-y_i^* g(x)} \right)
\end{aligned}
$$

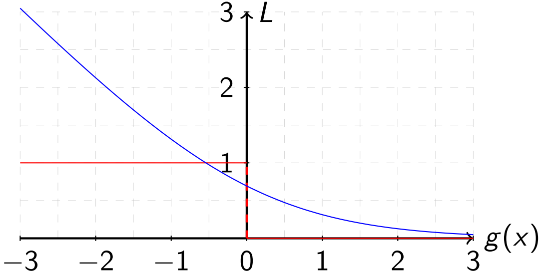
W ostatniej linijce zostało zmienione kodowanie klas $y_i^* \in \{-1, 1\}$.

Błąd logistyczny

Przekształcając wzór dalej otrzymujemy:

$$\min_{w,b} \sum_{i=1}^{n} -\left(y_i \ln p_{x_i} + (1 - y_i)\ln(1 - p_{x_i})\right) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i^* g(x)}\right)$$

przy czym zostało zmienione kodowanie klas $y_i^* \in \{-1, 1\}$.



- Błąd tej postaci nazywamy błędem logistycznym (ang. *logistic loss*).
- Zauważ, że formulacja jest ogólna i za $g(x)$ można wstawić inny model wiedzy niż wyrażenie liniowe, uzyskując inny algorytm wykorzystujący tę funkcję błędu.

Związek między Naiwnym Bayesem a Regresją Logistyczną

Wykonajmy kilka przekształceń wzoru na klasyfikator Naiwnego Bayesa.

$$P(Y=1|X) = \frac{P(X|Y=1)P(Y=1)}{P(X|Y=1)P(Y=1)+P(X|Y=0)P(Y=0)}$$

$$P(Y=1|X) = \frac{1}{\frac{P(X|Y=1)P(Y=1)+P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}}$$

O ile $\ln \frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}$ jest funkcją liniową to klasyfikator NB można zapisać jako LR!

Związek między Naiwnym Bayesem a Regresją Logistyczną

Wykonajmy kilka przekształceń wzoru na klasyfikator Naiwnego Bayesa.

$$P(Y=1|X) = \frac{P(X|Y=1)P(Y=1)}{P(X|Y=1)P(Y=1)+P(X|Y=0)P(Y=0)}$$

$$P(Y=1|X) = \frac{1}{\frac{P(X|Y=1)P(Y=1)+P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}}$$

O ile $\ln \frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}$ jest funkcją liniową to klasyfikator NB można zapisać jako LR!

Związek między Naiwnym Bayesem a Regresją Logistyczną

Wykonajmy kilka przekształceń wzoru na klasyfikator Naiwnego Bayesa.

$$P(Y=1|X) = \frac{P(X|Y=1)P(Y=1)}{P(X|Y=1)P(Y=1) + P(X|Y=0)P(Y=0)}$$

$$P(Y=1|X) = \frac{1}{1 + \frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}}$$

O ile $\ln \frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}$ jest funkcją liniową to klasyfikator NB można zapisać jako LR!

Związek między Naiwnym Bayesem a Regresją Logistyczną

Wykonajmy kilka przekształceń wzoru na klasyfikator Naiwnego Bayesa.

$$P(Y=1|X) = \frac{P(X|Y=1)P(Y=1)}{P(X|Y=1)P(Y=1)+P(X|Y=0)P(Y=0)}$$

$$P(Y=1|X) = \frac{1}{1+e^{\ln\left(\frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}\right)}}$$

O ile $\ln \frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}$ jest funkcją liniową to klasyfikator NB można zapisać jako LR!

Związek między Naiwnym Bayesem a Regresją Logistyczną

Wykonajmy kilka przekształceń wzoru na klasyfikator Naiwnego Bayesa.

$$P(Y = 1|X) = \frac{P(X|Y = 1)P(Y = 1)}{P(X|Y = 1)P(Y = 1) + P(X|Y = 0)P(Y = 0)}$$

$$P(Y = 1|X) = \frac{1}{1 + e^{\ln\left(\frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}\right)}}$$

$$P(Y = 1|X) = f\left(-\ln\frac{P(X|Y = 0)P(Y = 0)}{P(X|Y = 1)P(Y = 1)}\right)$$

O ile $\ln\frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}$ jest funkcją liniową to klasyfikator NB można zapisać jako LR!

Związek między Naiwnym Bayesem a Regresją Logistyczną

Wykonajmy kilka przekształceń wzoru na klasyfikator Naiwnego Bayesa.

$$P(Y=1|X) = \frac{P(X|Y=1)P(Y=1)}{P(X|Y=1)P(Y=1)+P(X|Y=0)P(Y=0)}$$

$$P(Y=1|X) = \frac{1}{1+e^{\ln\left(\frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}\right)}}$$

$$P(Y=1|X) = f\left(-\ln\frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}\right)$$

O ile $\ln\frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}$ jest funkcją liniową to klasyfikator NB można zapisać jako LR!

Związek między Naiwnym Bayesem a Regresją Logistyczną

Dla cech binarnych:

$$\ln \frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)} = \underbrace{\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) + \sum_{i=1}^{m} \ln \frac{P(X_i=0|Y=1)}{P(X_i=0|Y=0)}}_{w_0}$$

$$+ \sum_{i=1}^{m} \left(\underbrace{\left(\ln \frac{P(X_i=1|Y=1)}{P(X_i=1|Y=0)} - \ln \frac{P(X_i=0|Y=1)}{P(X_i=0|Y=0)} \right)}_{w_i} X_i \right) \tag{1}$$

$$= w_0 + \sum_{i=1}^{m} w_i x_i$$

Związek między LDA a Regresją Logistyczną

Dla rozkładu normalnego z jednostkową macierzą kowariancji:

$$\ln \frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)} = \underbrace{\ln \frac{P(1)}{P(0)} - \frac{1}{2}(||\mu_1||^2 + ||\mu_0||^2)}_{w_0} \tag{2}$$

$$+ \underbrace{(\mu_1^T - \mu_0^T)}_{w^T} x$$

Analogiczną zależność można pokazać dla rozkładu normalnego ze współdzieloną między klasami macierzą kowariancji.
W ogólności taki zapis nie jest możliwy dla dowolnego rozkładu!

Metody generatywne a dyskryminacyjne - podsumowanie

- Oba klasyfikatory mają tę samą formę (dla pewnych przypadków)
- Oba klasyfikatory używają zasady maksymalnej wiarygodności do wytrenowania parametrów
- Różnica:
  - NB/LDA optymalizuje MLE rozkładu łącznego $P(x, y)$
  - LR optymalizuje MLE rozkładu warunkowego $P(y|x)$
- Które z tych podejść jest lepsze?

Metody generatywne a dyskryminacyjne - podsumowanie

- Oba klasyfikatory mają tę samą formę (dla pewnych przypadków)
- Oba klasyfikatory używają zasady maksymalnej wiarygodności do wytrenowania parametrów
- Różnica:
  - NB/LDA optymalizuje MLE rozkładu łącznego $P(x, y)$
  - LR optymalizuje MLE rozkładu warunkowego $P(y|x)$
- Które z tych podejść jest lepsze?

Metody generatywne a dyskryminacyjne - podsumowanie

- Oba klasyfikatory mają tę samą formę (dla pewnych przypadków)
- Oba klasyfikatory używają zasady maksymalnej wiarygodności do wytrenowania parametrów
- Różnica:
  - NB/LDA optymalizuje MLE rozkładu łącznego $P(x, y)$
  - LR optymalizuje MLE rozkładu warunkowego $P(y|x)$
- Które z tych podejść jest lepsze?

Metody generatywne a dyskryminacyjne - podsumowanie

Twierdzenie (Ng, Jordan)

Niech h_{Gen} i h_{Dys} będą parą klasyfikatorów generatywny-dyskryminacyjny, a poprzez h^*_{Gen} i h^*_{Dys} oznaczmy ich wersje populacyjne[a]. Wtedy

$$\epsilon(h^*_{Dys}) \leq \epsilon(h^*_{Gen})$$

[a] wyobraź sobie nieskończony zbiór danych

Metody generatywne a dyskryminacyjne - podsumowanie

Problem

Wykorzystując poniższe pytania wskaż na wady i zalety podejść dyskryminacyjnych i generatywnych.

- *Łatwy do nauczenia?*
- *Łatwo dodać nową klasę?*
- *Łatwo obsłużyć brakujące dane?*
- *Łatwo wykorzystać niezaetykietowane dane podczas uczenia?*
- *Łatwo wykorzystać przetworzone cechy?*
- *Dobrze wykalibrowane prawdopodobieństwa?*

Zadanie

Problem

Przeanalizuj modele LDA, naiwnego Bayesa i regresji logistycznej w kontekście zasady minimalizacji ryzyka empirycznego. Jakie są klasy hipotez? Jak jest optymalizowana funkcja? Jaki algorytm optymalizacyjny może być stosowany?

Dziękuję za uwagę!

Fundusze Europejskie
Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego