Sztuczna inteligencja

Klasyfikacja bayesowska

Przemysław Klęsk, prezentujący Joanna Kołodziejczyk

Plan wykładu

Podział metod

Metody o motywacji:

1. geometrycznej (m.in. algorytm SVM[a])
2. probabilistycznej — czyli opartej na rachunku prawdopodobieństwa (m.in. naiwny klasyfikator bayesa)
3. mieszanej (m.in. drzewa decyzyjne CART, regresja logistyczna).

[a] ang. *Support Vector Machines*

Prawdopodobieństwo warunkowe

We wszystkich podejściach probabilistycznych elementarną rolę odgrywają ***prawdopodobieństwa warunkowe***.
W ramach krótkiego przypomnienia rozpoczynamy od omówienia tego pojęcia oraz kilku innych z nim powiązanych. Początkowo będziemy mówili o zdarzeniach losowych, stopniowo przechodząc do kontekstu zmiennych losowych i zbiorów danych.

Prawdopodobieństwo warunkowe

Niech A i B oznaczają podzbiory pewnej przestrzeni zdarzeń Ω, tj.: $A, B \subset \Omega$. Prawdopodobieństwo wystąpienia zdarzenia A, pod warunkiem że zaszło zdarzenie B, oblicza się następującym wzorem:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \tag{1}$$

gdzie $P(B) > 0$.
Innymi słowy, wśród zdarzeń elementarnych wspierających zdarzenie B patrzymy, jak często zachodzi także zdarzenie A, a zatem prawdopodobieństwo warunkowe $P(A|B)$ to iloraz miary przecięcia tych zdarzeń — $A \cap B$ (lub inaczej ich części wspólnej) w stosunku do miary zdarzenia B.

Interpretacja

W kontekście uczenia maszynowego lub eksploracji danych, możemy pytać np. o

- prawdopodobieństwo wystąpienia pewnej choroby pod warunkiem ustalonej płci (lub odwrotnie)
- prawdopodobieństwo, że grzyb jest trujący pod warunkiem cechy blaszkowatość.

Zarówno przed jak i za kreską warunkowania możemy rozpatrywać koniunkcje pewnych zdarzeń (co będzie oznaczane symbolem $\cap$ lub krócej przecinkiem). Warto nadmienić, że wspomaniane prawdopodobieństwa są zwykle utożsamiane z odpowiednimi częstościami odczytywanymi z tabelki z danymi, które badamy.

Przekształcenia

W niektórych zadaniach lub wyprowadzeniach przydatne mogą być dodatkowo poniższe przekształcenia manipulujące wzorem na prawdopodobieństwo warunkowe:

- przenoszenie zdarzenia B za kreskę warunkowania —

$$P(A,B|C) = \frac{P(A,B,C)}{P(C)} = \frac{P(A,B,C)\cdot P(B,C)}{P(C)\cdot P(B,C)} = P(A|B,C)P(B|C). \quad (2)$$

- przenoszenie zdarzenia B przed kreskę warunkowania —

$$P(A|B,C) = \frac{P(A,B,C)}{P(B,C)} = \frac{P(A,B,C)\cdot P(C)}{P(B,C)\cdot P(C)} = \frac{P(A,B|C)}{P(B|C)}. \quad (3)$$

Niezależność zdarzeń

W ramach rachunku prawdopodobieństwa istnieje pojęcie ***niezależności zdarzeń*** (definicja przedstawiona poniżej). Pojęcie to niesie ważne konsekwencje dla uczenia maszynowego w ogólności, a w szczególności dla klasyfikacji bayesowskiej.

Definition (niezależność zdarzeń)

Mówimy, że zdarzenia A i B są niezależne (piszemy $A \perp B$), wtedy i tylko wtedy, gdy prawdopodobieństwo ich iloczynu (wspólnego wystąpienia) jest równe iloczynowi prawdopodobieństw:

$$P(A \cap B) = P(A) \cdot P(B). \tag{4}$$

Niezależność zdarzeń

Jeżeli $A \perp B$, to możemy oczekiwać, że w odpowiednio dużej populacji zdarzenie A będzie pojawiało się z taką samą częstością w całej populacji jak i warunkowo w zdarzeniu B, oraz odwrotnie — B w przybliżeniu tak samo często w całej populacji, jak i w A. Należy także zwrócić uwagę, że jeżeli $A \perp B$, to:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A). \qquad (5)$$

Innymi słowy warunek mówiący o tym, że zaszło zdarzenie B nie wnosi dodatkowej informacji, która pomagałaby we wnioskowaniu o prawdopodobieństwie zajścia zdarzenia A.

Zmienne losowe

Przejdźmy na chwilę do ogólniejszego kontekstu zmiennych losowych (zamiast zdarzeń). Rozważmy dla przykładu zmienne losowe: *wzrost człowieka* (H) o dyskretnych wartościach $\{\mathrm{m}, \acute{\mathrm{s}}, \mathrm{d}\}$ odpowiednio o znaczeniu mały, średni, duży, oraz *kolor oczu* (C) o wartościach $\{\mathrm{z}, \mathrm{n}, \mathrm{b}, \mathrm{s}\}$ reprezentujących popularne kolory (zielony, niebieski, brązowy, szary).

Niezależność zmiennych losowych

Aby rozpatrywane zmienne te były niezależne, wzór (4) musiałby zachodzić dla wszystkich możliwych podstawień par wartości do tych zmiennych, tj.:

$$\forall_{h\in\{\mathrm{m,ś,d}\}}\forall_{c\in\{\mathrm{z,n,b,s}\}} \quad P(H=h\cap C=c)=P(H=h)\cdot P(C=c). \tag{6}$$

Rozstrzygnięcie, czy dla rozpatrywanego przykładu powyższy zapis jest prawdziwy, wymagałoby dokładniejszego sprawdzenia. Niemniej warto sobie uświadomić, że można z łatwością wskazać wiele przykładów zmiennych, które *nie* są niezależne. Przykłady: płeć i wzrost człowieka (mężczyźni są statystycznie wyżsi od kobiet), wzrost i waga człowieka (ludzie wyżsi są statystycznie cięższi), cena paliwa i koszt pewnej usługi transportowej, itd.

Prawdopodobieństwo całkowite

Theorem

Dla każdego rozbicia przestrzeni zdarzeń Ω na rozłączne podzbiory $B_1, B_2, \ldots, B_n$ (każdy o dodatniej mierze prawdopodobieństwa), tj:

$$\bigcup_{i=1}^{n} B_i = \Omega,$$
$$\forall i \neq j \quad B_i \cap B_j = \emptyset,$$
$$\forall i \quad P(B_i) > 0,$$

prawdopodobieństwo całkowite dowolnego zdarzenia A możemy obliczać wg wzoru:

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n) \\ &= \sum_{i=1}^{n} P(A|B_i)P(B_i). \end{aligned} \tag{7}$$

Zadania

Pomogą one zrozumieć wyprowadzenie naiwnego klasyfikatora bayesowskiego.

1. Trzy fabryki produkują żarówki. Prawdopodobieństwo zdarzenia polegającego na tym, że wyprodukowana żarówka będzie świeciła dłużej niż 5 lat, wynoszą dla tych fabryk odpowiednio: 0.9, 0.8, 0.7. Prawdopodobieństwa napotkania na rynku żarówek z poszczególnych fabryk wynoszą odpowiednio: 0.3, 0.5, 0.2. Jakie jest prawdopodobieństwo, że losowo zakupiona żarówka będzie świeciła dłużej niż 5 lat?
2. Jeżeli wiemy, że pewna losowo zakupiona żarówka świeciła dłużej niż 5 lat, to jakie jest prawdopodobieństwo, że pochodzi ona z drugiej fabryki?

Zadanie 1

Pierwsze zadanie sprowadza się do bezpośredniego zastosowania wzoru (7).
Wystarczają podstawienia
$P(A|B_1) = 0.9$,
$P(A|B_2) = 0.8$,
$P(A|B_3) = 0.7$,
oraz $P(B_1) = 0.3$, $P(B_2) = 0.5$, $P(B_3) = 0.2$.

Zadanie 2

Drugie zadanie to niejako zadanie odwrotne, pytające o $P(B_2|A)$. Podchodząc ogólniej, wyprowadźmy wzór na $P(B_i|A)$:

$$\begin{aligned} P(B_i|A) &= \frac{P(B_i \cap A)}{P(A)} \\ &= \frac{P(B_i \cap A)}{P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)} \\ &= \frac{P(B_i \cap A)\frac{P(B_i)}{P(B_i)}}{P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)} \\ &= \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)}. \end{aligned} \tag{8}$$

Jak wskazuje ostateczny wzór, patrzymy na udział i-tego składnika w całej sumie obliczanej wg prawdopodobieństwa całkowitego.

Plan wykładu

Założenie naiwne

Naiwny klasyfikator Bayesa (NBC — ang. *Naive Bayes Classifier*) to klasyfikator probabilistyczny z tzw. ***założeniem naiwnym***, które mówi, że zmienne wejściowe są niezależne warunkowo w klasach decyzyjnych, tzn.:

$$\forall y \; \forall i \neq j \qquad X_i | Y = y \; \perp \; X_j | Y = y. \tag{9}$$

Oczywiście powyższe założenie rzadko kiedy jest spełnione dla rzeczywistych danych (lub wręcz prawie nigdy nie jest spełnione), co silnie akcentuje nazwa klasyfikatora. Niemniej, fakt ten nie przeszkadza w używaniu NBC w praktyce, i co więcej okazuje się, że klasyfikator ten sprawdza się bardzo dobrze dla wielu problemów.

Konsekwencje założenia

Założenie naiwne ma ważny walor matematyczny, ponieważ pozwala na zastąpienie *prawdopodobieństwa iloczynu* pewnych zdarzeń *iloczynem prawdopodobieństw*.
Niesie to ważne konsekwencje obliczeniowe, dzięki którym po pierwsze realizacja NBC jest w ogóle możliwa, a po drugie NBC radzi sobie dobrze z dużą liczbą zmiennych (cech, atrybutów) — mogą być ich setki czy nawet tysiące, i nie cierpi na tzw. *przekleństwo wymiarowości*[1]. Wraz ze wzrostem liczby zmiennych wejściowych złożoność (obliczeniowa i pamięciowa) NBC skaluje się liniowo, a nie wykładniczo.

[1] Przekleństwo wymiarowości — zjawisko występując w niektórych algorytmach uczenia maszynowego polegające na tym, że złożoność opracowywanego modelu pewnego zjawiska (np. liczba parametrów, które należy dobrać) skaluje się wykładniczo wraz z liczbą zmiennych (cech, atrybutów) opisujących to zjawisko.

Plan wykładu

NBC ze zmiennymi dyskretnymi

Przyjmiemy, że wszystkie zmienne są dyskretne (lub inaczej: skokowe, wyliczeniowe, kategoryczne), jak np. płeć, kolor oczu, wykształcenie, wystąpienie choroby, marka samochodu, itp.
Jeżeli któraś ze zmiennych nie jest dyskretna (np. wzrost, waga, prędkość, temperatura), a chcielibyśmy jej użyć, to istnieją różne techniki dokonujące dyskretyzacji takiej zmiennej, czyli zamiany jej ciągłych wartości na dyskretne (np. wzrost mały, średni, duży).

Zbiór uczący

Przypuśćmy, że do dyspozycji jest pewien zbiór danych uczących z dyskretnymi zmiennymi wejściowymi X_i, $i = 1, \ldots, n$, oraz z wyróżnioną wyróżnioną zmienną decyzyjną Y (zawierającą etykiety klas decyzyjnych). Przypuśćmy, że rozróżniamy K klas decyzyjnych, oznaczonych np. kolejnymi numerami naturalnymi $\{1, 2, \ldots, K\}$.

Zbiór danych uczących z dyskretnymi zmiennymi

Zwyczajowo tego typu zbiór przedstawia się w formie tabelki, gdzie wierszami pisane są przykłady uczące[2] zaś kolumnami zmienne (cechy, atrybuty), patrz schemat w Tab. 1.

Tablica: Poglądowy schemat tabelki reprezentującej dyskretny zbiór uczący z wyróżnioną zmienną decyzyjną. Przykłady uczące pisane wierszami, zmienne kolumnami.

X_1	X_2	$\cdots$	X_n	Y
3	1	$\cdots$	2	1
2	5	$\cdots$	4	2
1	4	$\cdots$	2	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

[2] inne możliwe nazwy: próbki, obserwacje, rekordy, punkty danych

Częstości a prawdopodobieństwo

Na podstawie tabelki uczącej znamy rozkłady prawdopodobieństwa (tak naprawdę rozkłady częstości) *wektorów* wejściowych w poszczególnych klasach, tj. $X = x|Y = y$. Przypuśćmy, że mamy za zadanie sklasyfikować pewien nowo przychodzący obiekt (wektor) postaci $x = (x_1, x_2, \ldots, x_n)$, gdzie x_i reprezentują konkretne wartości, np. $x = (2, 3, \ldots, 1)$. A zatem, chcemy wyznaczyć taką etykietę klasy (lub numer klasy) — y^*, która jest najbardziej prawdopodobna dla podanego wektora wejściowego x, co można zapisać jako:

$$y^* = \text{argmax}_{y \in \{1,\ldots,K\}} P(Y = y|X = x). \tag{10}$$

Prawdopodobieństwo klasy

Zgodnie z twierdzeniem Bayesa o prawdopodobieństwie całkowitym, możemy $P(Y = y|X = x)$ rozpisać jako:

$$P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y = y)}{P(X)}$$
$$= \frac{P(X = x|Y = y)P(Y = y)}{P(X = x|Y = 1)P(Y = 1) + \cdots + P(X = x|Y = K)P(Y = K)}. \quad (11)$$

Warto tu zwrócić uwagę, że mianownik w powyższym wzorze jest stały i niezależny od y, dla którego badamy $P(Y = y|X = x)$. A zatem możemy zignorować mianownik przy podejmowaniu decyzji o najbardziej prawdopodobnej klasie, innymi słowy zachodzi:

$$y^* = \text{argmax}_{y\in\{1,\ldots,K\}} P(Y = y|X = x)$$
$$= \text{argmax}_{y\in\{1,\ldots,K\}} P(X = x|Y = y)P(Y = y). \quad (12)$$

Prawdopodobieństwo klasy – cd.

Rozpiszmy pierwszy powyższy czynnik, wprowadzając założenie naiwne (przejście z linii pierwszej do drugiej):

$$\begin{aligned} P(X = x|Y = y) &= P(X_1 = x_1 \cap X_2 = x_2 \cap \cdots \cap X_n = x_n|Y = y) \\ &= P(X_1 = x_1|Y = y)P(X_2 = x_2|Y = y) \cdots P(X_n = x_n| \\ &= \prod_{j=1}^{n} P(X_j = x_j|Y = y). \end{aligned} \tag{13}$$

Prawdopodobieństwo klasy – cd.

A zatem, wychodząc od (12), interesujący nas końcowy **wzór dla wariantu dyskretnego**, pozwalający przyporządkować obiektowi $x = (x_1, x_2, \ldots, x_n)$ najbardziej prawdopodobną klasę, przyjmuje postać:

$$y^* = \text{argmax}_{y \in \{1,\ldots,K\}} \prod_{j=1}^{n} P(X_j = x_j | Y = y) P(Y = y). \qquad (14)$$

Prosty przykład obliczeń dla NBC ze zmiennymi dyskretnymi

Przypuśćmy, że chcemy zastosować NBC w celu rozpoznawania (lub przewidywania) nadciśnienia tętniczego krwi u ludzi po 40 roku życia. Rozpoznawanie chcemy oprzeć na trzech zmiennych wejściowych (cechach): płci, aktywności sportowej, paleniu. Przypuśćmy dalej, że do dyspozycji jest następująca tabelka z oznakowanymi przykładami uczącymi (uwaga: dane zostały wymyślone na potrzeby przykładu) czyli takimi, dla których znamy zarówno wektory cech wejściowych jak i etykietę klasy, ponieważ ta została np. określona przez lekarza. Dla uproszczenia każda zmienna X_i przyjmuje dwie możliwe wartości.

Dane uczące dla przykładu

Tablica: Sztuczne dane dla problemu rozpoznawania (przewidywania) nadciśnienia tętniczego krwi u ludzi po 40 roku życia.

	X_1 — płeć	X_2 — sport	X_3 — palenie	Y — nadciśnienie
1	M	$-$	$+$	$+$
2	M	$+$	$+$	$-$
3	K	$-$	$-$	$-$
4	M	$+$	$+$	$+$
5	K	$+$	$-$	$-$
6	K	$+$	$-$	$-$
7	K	$-$	$-$	$+$
8	K	$+$	$-$	$+$
9	M	$-$	$+$	$+$
10	M	$+$	$-$	$+$
11	K	$-$	$-$	$-$
12	M	$-$	$-$	$+$
13	K	$-$	$-$	$-$
14	K	$+$	$-$	$-$
15	M	$-$	$+$	$+$
16	K	$-$	$+$	$+$

Uczenie

Uczenie NBC w wariancie dyskretnym polega na wyznaczeniu i zapamiętaniu (w pewnej strukturze danych, np. w tablicy lub słowniku) wszystkich możliwych prawdopodobieństw, które mogą być potrzebne jako czynniki we wzorze (14). Utożsamiając prawdopodobieństwa z częstościami występującymi w Tab. 2, byłyby to następujący zestaw:

$$P(Y=-) = 7/16$$

$$P(X_1 = M|Y=-) = 1/7 \qquad P(X_1 = K|Y=-) = 6/7$$
$$P(X_2 = -|Y=-) = 3/7 \qquad P(X_2 = +|Y=-) = 4/7$$
$$P(X_3 = -|Y=-) = 6/7 \qquad P(X_3 = +|Y=-) = 1/7$$

$$P(Y=+) = 9/16$$

$$P(X_1 = M|Y=+) = 6/9 \qquad P(X_1 = K|Y=+) = 3/9$$
$$P(X_2 = -|Y=+) = 6/9 \qquad P(X_2 = +|Y=+) = 3/9$$
$$P(X_3 = -|Y=+) = 4/9 \qquad P(X_3 = +|Y=+) = 5/9$$

Klasyfikacja

Sklasyfikujemy teraz dwa przykładowe nowo przychodzące obiekty: $(M, -, +)$ oraz $(K, -, -)$. Wzór (14) nakazuje nam „przejść" po wszystkich klasach decyzyjnych, dla każdej z nich obliczyć odpowiedni iloczyn prawdopodobieństw, i wreszcie wybrać jako odpowiedź tę klasę, dla której iloczyn jest największy.

Dla klasy: $y = -$, i obiektu $(M, -, +)$

$$P(X_1 = M|Y = -) \cdot P(X_2 = -|Y = -) \cdot P(X_3 = +|Y = -) \cdot P(Y = -)$$
$$= \frac{1}{7} \cdot \frac{3}{7} \cdot \frac{1}{7} \cdot \frac{7}{16} = \frac{3}{784} \approx 0.0038265,$$

Dla klasy $y = +$ i obiektu $(M, -, +)$

$$P(X_1 = M|Y = +) \cdot P(X_2 = -|Y = +) \cdot P(X_3 = +|Y = +) \cdot P(Y = +)$$
$$= \frac{6}{9} \cdot \frac{6}{9} \cdot \frac{5}{9} \cdot \frac{9}{16} = \frac{1\,620}{11\,664} \approx 0.1388889.$$

Jako, że druga z powyższych liczb jest większa, odpowiedzią NBC jest w tym przypadku $y^* = +$.

Dyskusja

Warto zwrócić uwagę, że obliczone wartości nie stanowią miar prawdopodobieństwa i nie sumują się do jedności. Powodem jest wspomniane wcześniej pominięcie mianownika (patrz (11)), który nie ma wpływu na decyzję. Jeżeli jednak z jakiegoś powodu zależy nam na wyznaczeniu liczb, które byłyby miarami prawdopodobieństwa (np. po to aby, poznać siłę wskazania na rzecz danej klasy na tle innych), to jako wspomniany mianownik należy przyjąć sumę obliczonych iloczynów (w zgodzie z założeniem naiwnym). W rozważanym przykładzie można by wówczas napisać:

$$P\big(Y=-|X=(M,-,+)\big)=\frac{\frac{3}{784}}{\frac{3}{784}+\frac{1\,620}{11\,664}}\approx 0.0268123,$$

$$P\big(Y=+|X=(M,-,+)\big)=\frac{\frac{1\,620}{11\,664}}{\frac{3}{784}+\frac{1\,620}{11\,664}}\approx 0.9731877.$$

Przykład cd.

Postępując analogicznie dla drugiego obiektu $(K, -, -)$, można przekonać się, że interesujące nas iloczyny wynoszą odpowiednio 108/343 i 8/81, które po znormalizowaniu do prawdopodobieństw przełożyłyby sie na:

$$P\big(Y = -|X = (K, -, -)\big) = \frac{\frac{108}{343}}{\frac{108}{343} + \frac{8}{81}} \approx 0.7612252,$$

$$P\big(Y = +|X = (K, -, -)\big) = \frac{\frac{8}{81}}{\frac{108}{343} + \frac{8}{81}} \approx 0.2387748.$$

Analiza

Patrząc ponownie na główny wzór (14), warto zwrócić uwagę na rolę, jaką pełni w nim czynnik $P(Y = y)$ nazywamy *prawdopodobieństwem a priori* klasy. W rozważanym przykładzie rozkład prawdopodobieństw a priori dla klas wynosił: $P(Y = -) = 7/16$, $P(Y = +) = 9/16$. Daje to pewną przewagę klasie $Y = +$ przy obliczaniu odpowiedzi klasyfikatora, ale jest to przewaga drobna — rozkład klas jest bliski równomiernemu. Jeżeli natomiast rozważylibyśmy problem wykrywania pewnego bardzo rzadkiego zjawiska (np. pożaru w monitorowanym obiekcie, obecności rzadkiego wirusa w całej populacji, itp.), to rozkład a priori byłby daleki od równomiernego[3] i rozpoznanie klasy o małym $P(Y = y)$ musiałoby się wiązać z wysokimi wartościami wielu innych czynników występujących we wzorze (14).

[3] Mówimy wówczas o tzw. danych niezrównoważonych (ang. *imbalanced data*).

Klasyfikator bezregułowy - zero-rule classifier

Nakazuje on odpowiadać zawsze klasą najczęstszą w rozkładzie a priori, nie zwracając uwagi na cechy badanego obiektu. Klasyfikator ten należy traktować jako punkt odniesienia, gdy zastanawiamy się, na ile dobry klasyfikator uzyskaliśmy dla naszego problemu.

Dla przykładu powiedzmy, że zajmujemy się problemem wykrywania wiadomości e-mail będących spamem i zbiór uczący, zebrany np. wśród pracowników uczelni, wskazuje na rozkład a priori: $P(Y = \text{nie-spam}) = 0.2$, $P(Y = \text{spam}) = 0.8$. Wówczas klasyfikator bezregułowy klasyfikowałby „na ślepo" wszystkie przychodzące wiadomości jako spam. W takiej sytuacji od dowolnego opracowanego klasyfikatora — bayesowskiego, sieci neuronowej, drzewa CART, maszyny SVM, itd. — wymagamy, aby miał on dokładność rozpoznawania powyżej 0.8 (mowa tu o dokładności zmierzonej na zbiorze testowym nie widzianym podczas uczenia). W przeciwnym razie nie byłoby żadnego zysku z uczenia maszynowego i stosowania klasyfikatora.

Ocena klasyfikatora

Dowolny opracowany klasyfikator powinien pod względem dokładności przewyższać klasyfikator bezregułowy (ang. *zero-rule classifier*).

Zerowe prawdopodobieństwo

Można zauważyć pewne niebezpieczeństwo obliczeniowe tkwiące we wzorze (14). Co, jeśli którykolwiek z czynników w tym wzorze byłby równy 0? Oczywiście, spowodowałoby to wyzerowanie całego wyniku niezależnie od tego, czy pozostałe czynniki były przeciętnie niskie czy też wysokie. Byłaby to sytuacja niepożądana. Kiedy mogłoby dojść do niej?
Zgodnie z tym, co powiedziano wcześniej, zwyczajowo utożsamia się prawdopodobieństwa z częstościami w zbiorze uczącym. I takie podejście nie jest błędne, jeżeli zbiór danych jest odpowiednio duży. Do sytuacji, o której mowa, mogłoby dojść wtedy, gdyby w zbiorze uczącym nie zaistniałaby realizacja pewnego zdarzenia, np. nigdy nie zaobserwowano $X_3 = 5|Y = 2$, a podczas testowania klasyfikatora pojawiłby się obiekt, dla którego trzecia cecha ma właśnie wartość 5.

Poprawka LaPlace'a

Istnieją różne techniki radzenia sobie z tym niebezpieczeństwem, polegających w ogólności na *wygładzaniu* rozkładów prawdopodobieństwa i tym samym unikaniu skrajnych prawdopodobieństw (zarówno zer jak i jedynek). Jednym z najbardziej popularnych jest tzw. ***poprawka LaPlace'a***

Poprawka

Przypuśćmy, że w m próbach zaobserowaliśmy k wystąpień pewnego zdarzenia A dotyczącego zmiennej o q unikalnych wartościach. Szacując prawdopodobieństwo na podstawie częstości, powinniśmy napisać $P(A) \approx k/m$. Stosując poprawkę LaPlace'a, oszacowanie przybiera postać

$$P(A) \approx \frac{k+1}{m+q}. \tag{15}$$

W szczególności dla zdarzeń binarnych powyższy wzór wynosi $\frac{k+1}{m+2}$.

Poprawka LaPlace'a – cd

Należy mieć świadomość, że dla małych zbiorów danych poprawka LaPlace'a zwykle psuje nieznacznie dokładność uczącą klasyfikatora, czyli jego zdolność do bezbłędnego odtworzenia etykiet danych uczących. Niemniej, jednocześnie (w takich sytuacjach) poprawka ta poprawia dokładność testową, czyli zdolność do uogólniania (generalizacji) dla niewidzianych obserwacji, a na tym właśnie elemencie zależy nam w uczeniu maszynowym.

Plan wykładu

NBC ze zmiennymi ciągłymi

Jeżeli chcielibyśmy używać naiwnego klasyfikatora Bayesa, pracując na zmiennych ciągłych (wzrost, temperatura, itp.) w sposób bezpośredni, tzn. nie dyskretyzując ich, to wzór (14) nie pozwala nam na to. Operuje on bowiem na prawdopodobieństwach pewnych zdarzeń rozumianych (mówiąc nieformalnie) w sposób gruboziarnisty, np. płeć = kobieta, kolor oczu = szary. Co, jeśli klasyfikacji ma podlegać człowiek np. o wzroście $188.7\,\mathrm{cm}$? Zwróćmy również uwagę, że poza powyższymi oczywistymi przykładami w wielu problemach istnieją zmienne o charakterze dyskretnym z natury rzeczy, a mimo to wolelibyśmy je traktować w sposób ciągły np. intensywność piskela o zbiorze wartości $\{0, 1 \ldots, 255\}$.

Zmienne ciągłe

Warto przypomnieć, że w rachunku prawdopodobieństwa dla zmiennych ciągłych rozróżniamy zwyczajowo dwie funkcje związane z rozkładem: funkcję *gęstości* rozkładu prawdopodobieństwa (PDF — ang. *probability density function*) oraz *dystrybuantę* zwaną także *kumulantą* (CDF — ang. *cumulative distribution function*). Wartości funkcji gęstości w punkcie nie mają jako takiego sensu probabilistycznego, a dopiero całki funkcji gęstości (obliczone nad przedziałami lub innymi zbiorami) stanowią miary prawdopodobieństwa pewnych zdarzeń. Np. jeżeli p oznacza funkcję gęstości pewnej skalarnej zmiennej X, to prawdopodobieństwa zdarzenia, że wartość (realizacja) tej zmiennej należy do przedziału $[a, b]$, możemy obliczyć następująco:

$$P(a \leqslant X \leqslant b) = \int_a^b p(x)\, dx. \tag{16}$$

Zmienne ciągłe

P funkcje gęstości całkują się nad całą dziedziną do jedynki, czyli np. dla gęstości jednowymiarowych mamy $\int_{-\infty}^{\infty} p(x)\,dx = 1$. Z kolei wartości funkcji dystrybuanty w punkcie mają sens probabilistyczny. Jeżeli oznaczyć dystrybuantę przez F, to:

$$F(a) = P(X \leqslant a) = \int_{-\infty}^{a} p(x)\,dx.$$

Można zapisać następujące związki pomiędzy gęstością a dystrybutantą:

- różniczka dystrybuanty = gęstość $\cdot$ przyrost:
$$dF(x) = p(x)\,dx,$$
- całka nieoznaczona z funkcji gęstości = dystrybuanta + stała:
$$\int p(x)\,dx = F(x) + C$$
- prawdopodobieństwo jako przyrost dystrybuanty:
$$\int_{}^{b}$$

NBC wariant ciągły

NBC wariant ciągły

Interesujący nas **wzór dla wariantu ciągłego** naiwnego klasyfikatora Bayesa to „kuzyn" wzoru (14), w którym w miejsce prawdopodobieństw wpisujemy wartości warunkowych funkcji *gęstości* w punkcie (z wyjątkiem prawdopodobieństw a priori klas — te pozostają bez zmian):

$$y^* = \text{argmax}_{y \in \{1,\ldots,K\}} \prod_{j=1}^{n} p_j(x_j|Y = y)P(Y = y). \qquad (17)$$

Ograniczenia

Należy być świadomym następujących ograniczeń związanych ze wzorem (17):

1. zwracana przezeń wartość nie powinna być interpretowana jako prawdopodobieństwo, nawet po normalizacji, ze względu na mieszane czynniki p i P o różnym sensie probabilistycznym (wzór jedynie „przypomina" iloczyn prawdopodobieństw),
2. nadal w mocy jest założenie naiwne — wyprowadzeniu wzoru (17) (które pominęliśmy) gęstości *łącznych* warunkowych rozkładów prawdopodobieństw $p(\mathrm{x}|Y = y)$, gdzie $\mathrm{x} = (x_1, \ldots, x_n))$ jest wektorem w $\mathbb{R}^n$, należy zamienić na iloczyn gęstości dla pojedynczych zmiennych $p_j(x_j|Y = y)$,
3. aby móc używać wzoru (17) należy wyznaczyć za pomocą wybranego podejścia *estymaty* funkcji gęstości $p_j(x_j|Y = y)$ na podstawie danych uczących (np. przyjmując, że są one zgodne z rozkładami normalnymi).

Estymaty za pomocą rozkładów normalnych

Rozkład Gaussa

Rozkłady normalne (zwane także gaussowskimi) obserwujemy bardzo często w przyrodzie. Fakt ten można w dużej mierze można wytłumaczyć poprzez Centralne Twierdzenie Graniczne mówiące, że rozkład zmiennej losowej, która jest sumą innych niezależnych zmiennych losowych, zbiega szybko do rozkładu normalnego wraz z liczbą składników. Wiele rzeczywistych wielkości, które obserwujemy lub mierzymy, można często rozumieć właśnie jako wypadkową (lub sumę) pewnych drobnych, niskopoziomowych elementów lub przyczyn.

Rozkłady

W związku z powyższym argumentem popularnym podejściem do realizacji ciągłego NBC jest estymowanie rozkładów zmiennych ciągłych za pomocą rozkładów *normalnych*. Oczywiście należy być świadomym, że jest to uproszczenie, które może przekłamywać wpływ niektórych szczególnych zmiennych w obliczanym iloczynie, tzn. tych zmiennych, których rozkłady są dalekie od normalnych (chociażby rozkłady wielomodalne).

Funkcja gęstości

Funkcja gęstości

Wzór funkcji gęstości dla rozkładu normalnego jednej zmiennej ma postać

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{18}$$

gdzie parametry μ i σ oznaczają odpowiednio wartość *średnią* (lub oczekiwaną) oraz *odchylenie standardowe* rozkładu.

Funkcja gęstości

Parametry te można oszacować na podstawie skończonej próby, czyli na podstawie zbioru danych. Dla uproszczenia, przypuśćmy że wszystkie rozpatrywane zmienne wejściowe są ciągłe, i przypomnijmy przyjętą we wcześniejszym rozdziale notację dla zbioru danych postaci: $D = \left\{(\mathrm{x}_i, y_i)\right\}_{i=1,\ldots,m}$, gdzie $\mathrm{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in}) \in \mathbb{R}^n$ są wektorami cech rzeczywistoliczbowych, zaś y_i etykietami klas.
A zatem chcąc przygotować NBC w wariancie ciągłym gaussowskim, musimy wyznaczyć $2 \cdot n \cdot K$ parametrów — oznaczmy je jako μ_{jy} i σ_{jy} (z użyciem pary indeksów) — będących średnimi i odchyleniami standardowymi, dla wszystkich warunkowych rozkładów zmiennych $X_j|Y = y$, gdzie $j = 1, \ldots, n$, $y = 1, \ldots, K$.

Funkcja gęstości – cd

Oznaczając gęstość takiego wybranego rozkładu jako

$$p_j(x|Y=y) = \frac{1}{\sigma_{jy}\sqrt{2\pi}} e^{-\frac{(x-\mu_{jy})^2}{2\sigma_{jy}^2}}, \tag{19}$$

stosuje się poniższe wzory do wyznaczenia estymat odpowiednio średniej i odchylenia standardowego:

$$\mu_{jy} = \frac{1}{m} \sum_{\substack{i=1 \\ y_i=y}}^{m} x_{ij}, \tag{20}$$

$$\sigma_{jy} = \sqrt{\frac{1}{m-1} \sum_{\substack{i=1 \\ y_i=y}}^{m} (x_{ij} - \mu_{jy})^2}. \tag{21}$$