Zadanie klasyfikacji-metody statystyczne wykł. 6

Joanna Jędrzejowicz

Instytut Informatyki

Klasyfikacja oparta o twierdzenie Bayesa

'Naiwna' klasyfikacja Bayesowska - zakłada się, że wpływ każdego atrybutu na wartość zmiennej celu jest niezależna od pozostałych atrybutów.
Załóżmy, że X jest rekordem danych, którego klasa jest nieznana. Niech H oznacza hipotezę, że X pochodzi z ustalonej klasy C. Chcemy wyznaczyć prawdopodobieństwo warunkowe $P(H|X)$, tj. prawdopodobieństwo, że hipoteza H jest prawdziwa pod warunkiem że mamy do czynienia z obserwacją X, nazywamy je prawdopodobieństwem aposteriori.
Prawdopodobieństwo $P(X|H)$ nazywamy prawdopodobieństwem apriori.

Twierdzenia Bayesa

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Jak stosujemy twierdzenia Bayesa do klasyfikacji

- Załóżmy, że rekord danych $X = (x_1, \ldots, x_n)$ jest reprezentowany przez n wartości atrybutów $A_1, \ldots, A_n$.
- Załóżmy, że mamy m klas, odpowiednio $C_1, \ldots, C_m$. Naiwny klasyfikator Bayesa przypisze próbce X klasę C_i wtedy i tylko wtedy, gdy

$$P(C_i|X) > P(C_j|X) \; dla \; j \neq i$$

- Z twierdzenia Bayesa mamy

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Jak stosujemy twierdzenie Bayesa do klasyfikacji, cd.

- Wartość $P(X)$ jest stała dla wszystkich klas, więc możemy ją pominąć i należy maksymalizować wartość $P(X|C_i)P(C_i)$. Przyjmuje się założenie, że prawdopodobieństwa klas są identyczne $P(C_1) = \cdots = P(C_m)$. Można także przyjąć przybliżenie $P(C_i) = \frac{s_i}{s}$, gdzie s jest liczbą danych rekordów, s_i - liczba rekordów z klasy C_i.
- Przyjmuje się założenie o niezależnosci atrybutów, czyli

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

- Jeśli atrybut A_k ma wartośc kategoryczną, to $P(x_k|C_i) = \frac{s_{ik}}{s_i}$, gdzie s_{ik} jest liczbą rekordów z klasy C_i, które dla atrybutu A_k mają wartość x_k,

Jak stosujemy twierdzenie Bayesa do klasyfikacji, cd.

Jeśli atrybut A_k jest typu ciągłego, to aproksymuje się prawdopodobieństwo, przyjmując rozkład gaussowski:

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi}\sigma_{C_i}} \exp(-\frac{x_k - \mu_{C_i}^2}{2\sigma_{C_i}^2})$$

Przykład liczbowy

Załóżmy, że korzystajac z danych z poprzedniego wykładu należy sklasyfikować rekord:
X=(wiek<=30, dochod=sredni, student=tak, wycenaKredytu=dostateczna)
Mamy dwie klasy: kupujeKomputer=tak, kupujeKomputer=nie

$$P(C_1) = \frac{9}{14} = 0.643, \; P(C_2) = \frac{5}{14} = 0.357$$

Przykład liczbowy, cd

Wyliczamy prawdopodobieństwa warunkowe

$$P(wiek <= 30|kupujeKomputer = tak) = \frac{2}{9} = 0.222$$

$$P(wiek <= 30|kupujeKomputer = nie) = \frac{3}{5} = 0.6$$

$$P(dochod = sredni|kupujeKomputer = tak) = \frac{4}{9} = 0.444$$

$$P(dochod = sredni|kupujeKomputer = nie) = \frac{2}{5} = 0.4$$

$$P(student = tak|kupujeKomputer = tak) = \frac{6}{9} = 0.667$$

$$P(student = tak|kupujeKomputer = nie) = \frac{1}{5} = 0.2$$

Przykład, cd

$$P(wycenaKredytu = dostat|kupujeKomputer = tak) = \frac{6}{9} = 0.667$$

$$P(wycenaKredytu = dostat|kupujeKomputer = nie) = \frac{2}{5} = 0.4$$

$$P(X|kupujeKomp = tak) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$$

$$P(X|kupujeKomp = nie) = 0.600 \times 0.400 \times 0.2 \times 0.4 = 0.019$$

$$P(X|kupujeKomputer = tak)P(kupujeKomputer = tak) =$$

$$= 0.044 \times 0.643 = 0.028$$

Przykład, dokończenie

$$P(X|kupujeKomputer = nie)P(kupujeKomputer = nie) =$$

$$= 0.019 \times 0.357 = 0.007$$

Naiwny klasyfikator Bayesa przewiduje klasę kupujeKomputer=tak.
Sprawdzic, jak klasyfikuje ten przyklad drzewo decyzyjne z ID4

Algorytm k-najbliższych sąsiadów

Algorytm k-najbliższych sąsiadów jest przykładem uczenia leniwego, to znaczy algorytm zaczyna działać, dopiero wtedy gdy potrzebne jest zaklasyfikowanie danego przykładu. Takie podejscie różni się od uczenie klasyfikatora na podstawie danych treningowych zastosowanego np. przy generowaniu drzew decyzyjnych.
Algorytm k-najbliższych sąsiadów przypisuje nowemu rekordowi klasę najbardziej 'podobnego' rekordu lub rekordów. Jak mierzyć podobieństwo? Załóżmy, że w zbiorze treningowym znajdują się rekordy złożone z wartości n atrybutów oraz wartości funkcji celu (inaczej, klasy).

Jakie problemy pojawiają się przy budowaniu klasyfikatora metodą najbliższych sąsiadów

- Ilu sąsiadów należy rozważać, czyli jaka jest wartość k,
- Jak określać odległość,
- Jak łaczyć informacje pochodzące z więcej niż jednej obserwacji,
- Czy wszystkie punkty powinny mieć jednakową wagę,
- Co zrobić z atrybutami kategorycznymi.

Metryki - przypomnienie

Dla atrybutów liczbowych:
odległość euklidesowa

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

metryka miasto (metryka Manhattan)

$$d(x, y) = \sum_i |x_i - y_i|$$

Metryka VDM

Metryka VDM (Value Diffrence Metric: Stanfill, Walz 1986) dla atrybutów nominalnaych (inaczej kategorycznych).
Niech: C oznacza zbiór klas, a - ustalony atrybut, x, y - wartości atrybutu

$$vdm_a(x, y) = \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|$$

gdzie $N_{a,x}$ oznacza liczbę przykładów, które dla atrybutu a przyjmują wartość x,
gdzie $N_{a,x,c}$ oznacza liczbę przykładów, które dla atrybutu a przyjmują wartość x i pochodzą z klasy c,

Metryka HEOM

Metryka HEOM (Heterogenenous Value Diffrence Metric) może być użyta dla przykładów, których część atrybutów jest liczbowa i część kategoryczna:

$$HVDM(x,y) = \sqrt{\sum_{a=1}^{m} d_a^2(x_a, y_a)}$$

$$d_a(x,y) = \begin{cases} 1 & \text{jezeli } x \text{ lub } y \text{ nieokreślone, inaczej} \\ vdm_a(x,y) & \text{atrybut } a \text{ kategoryczny} \\ diff_a(x,y) & \text{atrybut } a \text{ jest liczbowy} \end{cases}$$

$$diff_a(x,y) = \frac{|x-y|}{4 \cdot \sigma_a}$$

σ_a - odchylenie standardowe a

Algorytm k najbliższych sąsiadów - proste głosowanie

- Przed uruchomieniem algorytmu określ wartość k, czyli ile rekordów będzie decydowało o klasyfikacji nowego rekordu oraz sposób definiowania odległości,
- Porównaj nowy rekord z k najbliższymi sąsiadami, czyli z k rekordami które mają najmniejszą odległość od nowego rekordu. Uwaga: bierzemy pod uwagę wartości n atrybutów porównywanych rekordów, pomijamy klasy.
- Sprawdź klasy sąsiadów i przyjmując zasadę głosowania większościowego przypisz klasę dla nowego rekordu,

Głosowanie ważone

- Można poszczególnym sąsiadom przyporządkować wagi, tak aby 'bliższi' sąsiedzi mieli większy wpływ na decyzje klasyfikacyjną,
- W niektórych przypadkach odległość określa się w taki sposób, aby wybrane atrybuty odgrywały istotniejszą rolę niz inne.

Wybór wartości k

Może nie być oczywistego, najlepszego rozwiązania..
Jeżeli przyjmie się małą wartośc k, to klasyfikacja może być pod nadmiernym wpływem punktów oddalonych lub 'szumów'; dla $k = 1$ proces uczenia może zakończyć sie 'przeuczeniem' i zapamiętaniem zbioru uczacego kosztem umiejętnosci generalizowania. Przy wyborze zbyt dużej wartości k, lokalnie ciekawe zachowanie może zostac przeoczone.
Często wartość k wyznacza się wykonując wiele eksperymentów dla ustalonych zbiorów danych.

Techniki ewaluacji modelu do zadania klasyfikacji- przypadek dwóch klas

Załóżmy, że mamy do czynienia z dwiema klasami,np. bank przydziela kredyt oraz bank nie przydziela kredytu. Nazwijmy te dwie klasy odpowiednio: pozytywną i negatywną, a liczbę odpowiednich rekordów próbką pozytywną i próbką negatywną. Określamy następujące 4 wartości:

- TP (true positive) - liczba rekordów prawidłowo zaklasyfikowanych jako pozytywne,
- FP (false positive) - liczba rekordów, które zostały błednie zaklasyfikowane jako pozytywne,
- TN (true negative) - liczba rekordów prawidłowo zaklasyfikowanych jako negatywne,
- FN (false negative) - liczba rekordów, które zostały błednie zaklasyfikowane jako negatywne,

Macierz błędu klasyfikacji

Macierz błędu klasyfikacji (ang. confusion matrix)

klasa			ogółem
pozytywne	TP	FN	P
negatywne	FP	TN	N

Wartości na przekątnej dotyczą prawidłowej klasyfikacji.
Kiedy klasyfikator idealny?

Macierz błędu klasyfikacji, cd

W niektórych zastosowaniach błędne klasyfikacje mogą mieć różny koszt.

- wynik badania medycznego- pozytywne=jest chory - gorszym błędem jest traktowanie chorego pacjenta jako zdrowego niż odwrotnie, czyli FN powinno być niskie
- udzielanie kredytu - pozytywne=klient wiarygodny; FP powinno byc niskie - bank nie chce dac kredytu klientowi ktory nie jest wiarygodny,
- wyszukiwarka jako klasyfikator - akceptujemy duze FN (trafne strony pominięte) i male FP (niestotne dla zapytania znalezione)

Współczynniki korzystające z macierzy błędu klasyfikacji

Często korzysta się z dodatkowych miar

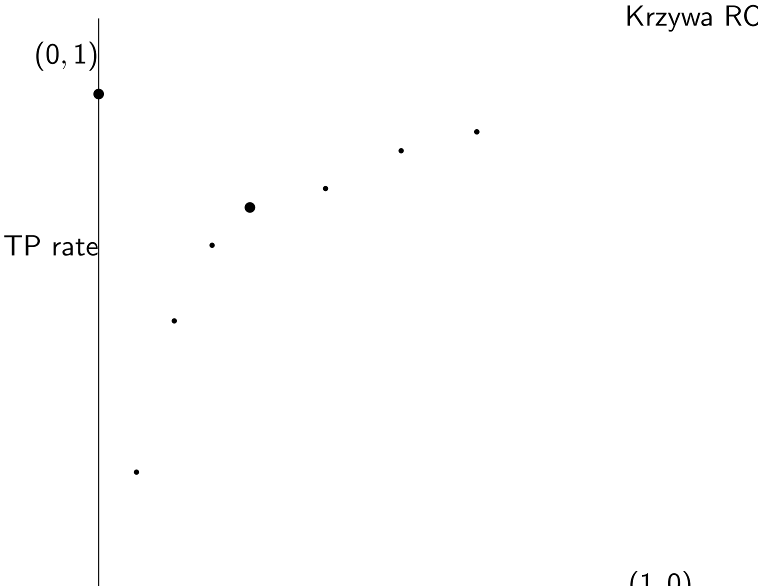
- czułość klasyfikatora (ang. sensitivity) $\frac{TP}{TP+FN}$
- specyficzność klasyfikatora (ang. specificity) $\frac{TN}{TN+FP}$
- współczynnik błędu $\frac{FN+FP}{TP+FN+FP+TN}$

jesli przyjąć nast. interpretację - pozytywna=ma daną chorobę, to

- czulośc mierzy na ile dobrze klasyfikator znalazl te przypadki
- specyficznosc - na ile dobrze znalazl przypadki, gdy nie ma choroby

Krzywe ROC

Wyniki różnych klasyfikatorów na tym samym zbiorze danych lub wyniki sprawdzianu krzyżowego często przestawia się graficznie przy pomocy wykresu ROC (ang. receiver operating characteristic).
Oś x-ów zawiera wartości $\frac{FP}{TN+FP}$ (FP-rate), os y-ów $\frac{TP}{TP+FN}$ (TP-rate).
Krzywa ROC bada stosunek *true positives* do *false positives*.
Krzywe ROC, jako technika analizy danych, zostały wprowadzone podczas II wojny światowej do analizy danych pochodzących z radarów.


Krzywa RC
$(0, 1)$
TP rate
$(1, 0)$

Krzywe ROC

Krzywych ROC używa się także do scharakteryzowania jednego klasyfikatora dla różnych wartości *progu* lub *parametru*, np. określającego od jakiej wartosci klasyfikujemy przykład jako pozytywny: rekord jest w klasie pozytywnej jeżeli wyliczone prawdopodobieństwo jest większe od wartości progu.

Wtedy dla każdej wartości progu mamy macierz błędu oraz punkt ROC, zaś dla różnych wartości progu otrzymujemy krzywą ROC złożoną z punktów ROC dla kolejnych wartości progu.

Krzywa ROC - przykład

	klasa	prawdop=próg	TP-rate	FP-rate
1	+	0.9	0.2	0
2	+	0.8	0.4	0
3	-	0.75	0.4	0.2
4	-	0.7	0.4	0.4
5	+	0.65	0.6	0.4
6	-	0.6	0.6	0.6
7	+	0.5	0.8	0.6
8	+	0.4	1.0	0.6
9	-	0.3	1.0	0.8
10	-	0.2	1.0	1.0

Macierz błędu klasyfikacji dla wybranych progów

próg 0.9

klasa			ogółem
pozytywne	1	4	5
negatywne	0	5	5

$$Tp{-}rate = \frac{TP}{TP+FN} = \frac{1}{5} = 0.2$$

$$Fp - rate = \frac{FP}{FP+TN} = 0$$

próg 0.65

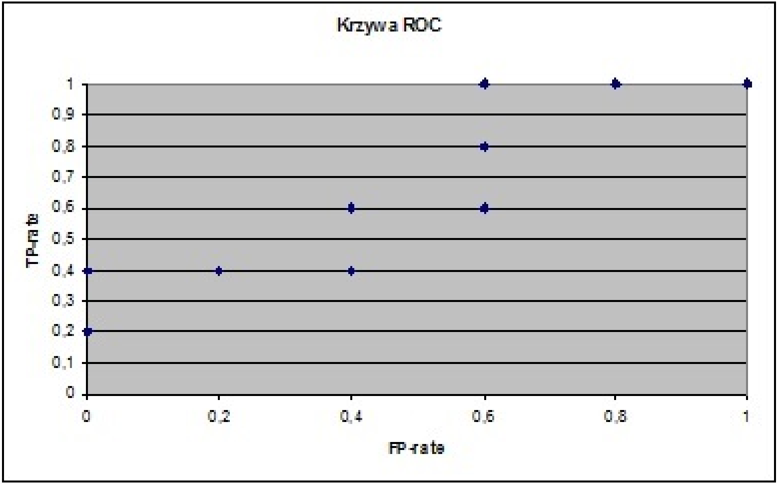
klasa			ogółem
pozytywne	3	2	5
negatywne	2	3	5

$$Tp{-}rate = \frac{TP}{TP+FN} = \frac{3}{5} = 0.6$$

$$Fp{-}rate = \frac{FP}{FP+TN} = \frac{2}{5} = 0.4$$

Przyklad

Otrzymalimy funkcję schodkową - zwiększając liczbę wartosci parametru otrzymujemy krzywą wygładzoną.



Różne miary oceniające klasyfikator

nazwa	wzór	opis
TP-rate	$\frac{TP}{P}$	proporcja pozytywnych przykładów sklasyfikowanych prawidlowo jako pozytyw
FP -rate	$\frac{FP}{N}$	proporcja negatywnych przykładów sklasyfikowanych błędnie jako pozytywn
dokładność	$\frac{TP+TN}{P+N}$	proporcja prawidlowo sklasyfikowanych