Ćwiczenie 6 - Hurtownie danych i metody eksploracje danych

Regresja logistyczna i jej zastosowanie

Model regresji logistycznej jest budowany za pomocą klasy `Logistic` programu WEKA. Jako danych wejściowych zostanie użyta zmodyfikowana wersja zbioru danych `cereals` - dostępny pod adresem `http://lib.stat.cmu.edu/DASYL/` lub `http://www.data.miningconsultant.com`, w którym pole **RATING** (wartość odżywcza) zostało zdyskretyzowane poprzez odwzorowanie rekordów z wartościami większymi od 42 jako **"High"** (wysoka), podczas gdy mniejsze od 42 lub równe tej wartosc i będą oznaczone jako **"Low"** (niska). W ten sposób nasz model zostanie wykorzystany w celu klasyfikacji płatków śniadaniowych na takie, które mają wysoką lub niską wartość odżywczą. Zbiór danych składa się z trzech numerycznych zmiennych objaśniających: **PROTEIN** (białka),**SODIUM** (sód) i **FIBER** (błonnik).

Zbiór danych został podzielony na osobne pliki zawierające zbiory uczący i testujący. Plik zawierający zbiór testujący, `cereals-train.arff`, składa się z 24 rekordów i jest użyty do nauczenia naszego modelu regresji logistycznej. Plik jest wyważony — połowa rekordów przyjmuje wartość **"High"**, a druga połowa przyjmuje wartość **"Low"**. Średnie wartości dla zmiennych objaśniających **PROTEIN**, **SODIUM** i **FIBER** są równe odpowiednio 2,667, 146,875 i 2,458. Cały plik zawierający zbiór uczący jest pokazany poniżej.

Plik `cereals-trains.arff` zawierający zbiór uczący

@relation cereals-train.arff

©attribute PROTEIN numeric
©attribute SODIUM numeric
©attribute FIBER numeric
©attribute RATING {High, Low}

©data
3,200,3.000000,High
3,230,3.000000,High
3,200,3.000000,High

3,0,4.000000,High
4,150,2.000000,High
3,0,3.000000,High
4,260,9.000000,High
3,140,3.000000,High
2,0,3.000000,High
2,0,2.000000,High
3,80,1.000000,High
2,200,4.000000,High
2,180,1.500000,Low
4,150,3.000000,Low
2,140,2.000000,Low
4,95,3.000000,Low
1,220,0.000000,Low
2,180,0.000000,Low
3,140,4.000000,Low
4,170,2.000000,Low
2,200,l.000000,Low
3,250,1.500000,Low
2,200,1.000000,Low
l,140,0.000000,Low

Pliki zawierające zbiory testujący i uczący są w formacie ARFF, który jest standardową reprezentacją rekordów i atrybutów znalezionych w zbiorze danych dla programu WEKA. Słowo kluczowe **relation** poprzedza nazwę pliku, następnie znajduje się blok definiujący każdy atrybut (**attribute**) w zbiorze danych. Można zauważyć, że trzy zmienne opisujące są zdefiniowane jako numeryczne (**numeric**), podczas gdy zmienna celu **RATING** jest jakościowa. Blok danych data wymienia każdy rekord w nowej linii, które odpowiadają poszczególnym płatkom śniadaniowym. Na przykład pierwsza linia w bloku danych opisuje płatki śniadaniowe, które mają 3 gramy białka (**PROTEIN = 3,0**), 200 miligramów sodu (**SODIUM = 200**), 3 gramy błonnika (**FIBER = 3,0**) i wysoką wartość odżywczą (**RATING = High**). Załadujmy plik zawierający zbiór uczący i zbudujmy model:

1) Otwórz okno **Explorer** programu WEKA.

2) Na zakładce **Preprocess** (wstępne przetworzenie) kliknij **Open file** (otwórz plik) i określ ścieżkę do pliku wejściowego `cereals-train.arff`.

Okno Explorer programu WEKA pokazuje kilka charakterystyk pliku uczącego. Trzy zmienne objaśniające i zmienną klasy pokazano w ramce **Attributes** (atrybuty, po lewej stronie). Statystyki dla zmiennej **PROTEIN**, obejmujące zakres (1-4), średnią (2,667) i odchylenie standardowe (0,868) pokazano w ramce **Selected attribute** (wybrany atrybut, po prawej stronie). Ramka **Status** na dole okna informuje, że plik został poprawnie wczytany przez program WEKA.

1) Wybierz zakładkę **Classify** (klasyfikuj).

2) Wewnątrz ramki **Classifier** (klasyfikator) naciśnij przycisk **Choose** (wybierz).

3) Wybierz **Classifiers** (klasyfikatory) → **Functions** (funkcje) → **Logistic** (logistyczna) w hierarchii nawigacji.

4) W naszym doświadczeniu modelowania mamy osobne zbiory uczący i testujący, dlatego wewnątrz ramki **Test options** (opcje testowe) wybierz opcję **Use training set** (użyj zbioru uczącego).

5) Naciśnij **Start**, aby zbudować model.

WEKA buduje model regresji logistycznej i pokazuje wyniki w oknie **Classifier out-put** (wynik klasyfikacji). Chociaż wyniki (niepokazane) wskazują, że dokładność klasyfikacji modelu, mierzona względem zbioru uczącego, to 75zastosowaniem modelu do klasyfikacji nieznanych danych ze zbioru testującego. Iloraz szans i wartości współczynników regresji β_0, β_1, β_2, β_3 są również podawane przez model, co pokazano w tabeli poniżej.

Zmienna	Współczynnik
1	-0.0423
2	-0.0107
3	0.9476
Przecięcie	-0.5478

Iloraz szans

Zmienna	OR
1	0.9586
2	0.9893
3	2.5795

1) Wewnątrz ramki **Test options** (opcje testowe) wybierz **Supplied test set** (dostarczony zbiór testujący). Naciśnij **Set** (ustaw).

2) Określ ścieżkę dostępu do pliku testowego `cereals-test.arff`.

3) Naciśnij przycisk **More options** (więcej opcji).

4) Wybierz **Output text predictors on the test set** (pokaż przewidywania dla zbioru testującego) i wybierz **OK**.

5) Wewnątrz **Result list** (lista wyników) wybierz z listy **Logistic model** (model logistyczny), naciskając prawym przyciskiem myszy. Następnie wybierz **Re-evaluate model on the current test set** (powtórnie oceń model na obecnym pliku testującym).

Znowu wyniki są prezentowane w oknie **Classifier output** (wynik klasyfikacji), jednak teraz wyniki pokazują, że model regresji logistycznej sklasyfikował poprawnie 62,5Dodatkowo model podaje właściwe przewidywania i prawdopodobieństwa, dla których sklasyfikował każdy rekord — patrz tabela poniżej.

inst#,	actual,	predicted,	error,	probability	distribution
1	1:High	2:Low	+	0.433	*0.567
2	l:High	2:Low	+	0,357	*0.643
3	l:High	l:High		*0.586	0,414
4	l:High	l:High		*0.578	0,422
5	2:Low	2:Low		0.431	*0.569
6	2:Low	2:Low		0.075	*0.925
7	2:Low	2:Low		0.251	*0.749
8	2:Low	l:High	+	*0.860	0.140

Na przykład, pierwszy przypadek jest przewidziany (sklasyfikowany) niepoprawnie jako **"Low"** z prawdopodobieństwem 0.567. Symbol plus (+) w kolumnie **"error"** wskazuje, że ta klasyfikacja jest niewłaściwa zgodnie z kryterium maksymalnego prawdopodobieństwa (*0,567). Obliczymy teraz oszacowaną funkcję logitową $g(x)$ dla tego przypadku zgodnie ze współczynnikami pokazanymi w tabeli 1. Jednak najpierw zbadamy plik zawierający dane testujące `cereals-test.arff` i ustalimy, że pierwszy rekord zawiera następujące pary atrybut-wartość: **PROTEIN = 4.0**, **SODIUM = 135**, **FIBER = 2.0**, i **RATING = High**. Dlatego też oszacowana wartość funkcji logitowej jest równa

$$g(x) = -0.5478 - 0.0423 \cdot 4 - 0.0107 \cdot 135 + 0.9476 \cdot 2 = -0.2663$$

a stąd

$$\pi(x) = \frac{e^{-0.2663}}{1 + e^{-0.2663}} = 0.43382$$

Zatem oszacowane prawdopodobieństwo, że płatki śniadaniowe z 4 gramami białka, 135 miligramami sodu i 2 gramami błonnika mają dużą wartość odżywczą jest równe około 43.4%. Zauważmy, że program WEKA podaje to samo prawdopodobieństwo (poza drobnymi zaokrągleniami) dla pierwszego przypadku w tabeli powyższej. Otrzymujemy więc, że model szacuje prawdopodobieństwo niskiej wartości odżywczej dla tych płatków jako $1 - \pi = 56.6\%$. Dlatego na podstawie powyższego prawdopodobieństwa stwierdzamy, że model niewłaściwie klasyfikuje rekord jako **"Low"**.

W pierwszej tabeli pokazano również ilorazy szans dla trzech ciągłych zmiennych objaśniających. Na przykład iloraz szans dla zmiennej **PROTEIN** to **OR**$= e^{b_1} = e^{-0.0423} = 0.9586$. Jest to interpretowane jako szansa wysokiej wartości odżywczej dla płatków z $x+1$ gramami białka w porównaniu z płatkami śniadaniowymi z x gramami białka.

2. Praca z innymi danymi

1. Przeprowadź analizę regresji logistycznej dla zbioru danych `breast-cancer`. Dziesięć numerycznych zmiennych objaśniających użyto w nim do przewidywania wystąpienia złośliwego raka piersi (**class = 1**) w odróżnieniu od łagodnego raka piersi (*class = 0*).

1) Która zmienna wydaje się nie być znaczącą zmienną objaśniającą dla typu raka piersi. Na jakiej podstawie możesz to stwierdzić?

2) Czy zmienne brane powyżej pod uwagę powinny zostć użyte do przewidywania typu raka piersi dla nowego nieznanego zbioru danych?

3) Jak należy traktować zmienne z p-wartościami około 0.05, 0.10, lub 0.15.

2. Dla zbioru danych `adult` dostępnego na stronie internetowej `www.datamining consultant.com` zbuduj model regresji logistycznej ze zmienną `wiek`^{2} i zmienną wskaźnikową `wiek` $33 - 65$.

1) Potwierdź, że kwadratowa składowa zapewnia wyższe oszacowanie prawdopodobieństwa wysokiego dochodu osoby 32-letniej niż osoby 20-letniej.

2) Znajdź postać oszacowania funkcji logitowej.

3) Znajdź prawdopodobieństwo wysokiego dochodu dla osoby, która pracuje 30, 40, 50 i 60 godzin tygodniowo.