Algorytmy eksploracji danych: Wykład 9

Wojciech Kempa

Politechnika Śląska
Wydział Matematyki Stosowanej

Problem „częstości zero" (1)

- Stosując „naiwny" klasyfikator bayesowski możemy spotkać się w praktyce z tzw. **problemem „częstości zero"**.

Problem „częstości zero" (1)

- Stosując „naiwny" klasyfikator bayesowski możemy spotkać się w praktyce z tzw. **problemem „częstości zero"**.
- Mamy z nim do czynienia w sytuacji, w której co najmniej jedno z prawdopodobieństw $\mathbf{P}(A_j = x_j \,|\, C = C_i)$ oszacowane za pomocą częstości względnej na podstawie zbioru treningowego wynosi 0.

Problem „częstości zero" (1)

- Stosując „naiwny" klasyfikator bayesowski możemy spotkać się w praktyce z tzw. **problemem „częstości zero"**.
- Mamy z nim do czynienia w sytuacji, w której co najmniej jedno z prawdopodobieństw $\mathbf{P}(A_j = x_j \,|\, C = C_i)$ oszacowane za pomocą częstości względnej na podstawie zbioru treningowego wynosi 0.

- Problem „częstości zero" pojawia się w przypadku, gdy w zbiorze treningowym nie występuje ani jeden obiekt, dla którego $A_j = x_j$ oraz $C = C_i$.

Problem „częstości zero" (1)

- Stosując „naiwny" klasyfikator bayesowski możemy spotkać się w praktyce z tzw. **problemem „częstości zero"**.
- Mamy z nim do czynienia w sytuacji, w której co najmniej jedno z prawdopodobieństw $\mathbf{P}(A_j = x_j \,|\, C = C_i)$ oszacowane za pomocą częstości względnej na podstawie zbioru treningowego wynosi 0.
- Problem „częstości zero" pojawia się w przypadku, gdy w zbiorze treningowym nie występuje ani jeden obiekt, dla którego $A_j = x_j$ oraz $C = C_i$.
- W praktyce jednak uniemożliwia to przeprowadzenie klasyfikacji: klasa C_i atrybutu decyzyjnego jest w tym momencie „automatycznie" eliminowana jako potencjalna klasa obiektu X.

Problem „częstości zero" (2)

Eliminacja „częstości zero" dla danej klasy C_i polega na **modyfikacji wszystkich prawdopodobieństw** $\mathbf{P}(A_j = x_j \,|\, C = C_i)$ dla $i = 1, ..., m$ w następujący sposób:

$$\mathbf{P}(A_j = x_j \,|\, C = C_i) = \frac{n_{i,j} + \lambda}{n_i + \lambda \cdot m_j},$$

gdzie m_j oznacza liczbę różnych kategorii atrybutu A_j, a λ jest pewnym **współczynnikiem skalującym** (współczynnikiem korekcji). Najczęściej przyjmuje się w praktyce $\lambda = \frac{1}{n}$, gdzie n oznacza liczebność zbioru treningowego. Jeżeli $\lambda = 1$, to estymator prawdopodobieństwa warunkowego zdefiniowany wzorem (**??**) nazywamy **estymatorem Laplace'a**.

Metoda k-najbliższych „sąsiadów” (1)

- **Metoda k-najbliższych „sąsiadów”** (w skrócie metoda k-NN) należy do grupy klasyfikatorów opartych na analizie przypadku i jest realizowana w trybie, który można określić jako tryb “on-line”: nie konstruuje się w tym przypadku funkcji klasyfikacyjnej, a sama klasyfikacja odbywa się na bieżąco.

Metoda k-najbliższych „sąsiadów” (1)

- **Metoda k-najbliższych „sąsiadów”** (w skrócie metoda k-NN) należy do grupy klasyfikatorów opartych na analizie przypadku i jest realizowana w trybie, który można określić jako tryb “on-line”: nie konstruuje się w tym przypadku funkcji klasyfikacyjnej, a sama klasyfikacja odbywa się na bieżąco.

- W literaturze tego typu metody określa się jako tzw. „leniwe” metody uczące (ang. *lazy learning methods*).

Metoda k-najbliższych „sąsiadów" (1)

- **Metoda k-najbliższych „sąsiadów"** (w skrócie metoda k-NN) należy do grupy klasyfikatorów opartych na analizie przypadku i jest realizowana w trybie, który można określić jako tryb "on-line": nie konstruuje się w tym przypadku funkcji klasyfikacyjnej, a sama klasyfikacja odbywa się na bieżąco.
- W literaturze tego typu metody określa się jako tzw. „leniwe" metody uczące (ang. *lazy learning methods*).
- Zastosowanie metody k-NN zasadniczo wymaga danych (atrybutów) ilościowych (liczbowych). Model klasyfikacyjny stanowi sam zbiór treningowy.

Metoda k-najbliższych „sąsiadów" (2)

- W najprostszej wersji metody, algorytmie $1-$NN ($k = 1$), obiekt X klasyfikuje się poprzez wybór ze zbioru treningowego D pojedynczego obiektu położonego najbliżej obiektu X.

Metoda k-najbliższych „sąsiadów” (2)

- W najprostszej wersji metody, algorytmie $1-$NN ($k = 1$), obiekt X klasyfikuje się poprzez wybór ze zbioru treningowego D pojedynczego obiektu położonego najbliżej obiektu X.
- Obiekt X zalicza się następnie do klasy atrybutu decyzyjnego reprezentowanej przez obiekt wybrany ze zbioru D.

Metoda k-najbliższych „sąsiadów” (2)

- W najprostszej wersji metody, algorytmie $1-$NN ($k = 1$), obiekt X klasyfikuje się poprzez wybór ze zbioru treningowego D pojedynczego obiektu położonego najbliżej obiektu X.
- Obiekt X zalicza się następnie do klasy atrybutu decyzyjnego reprezentowanej przez obiekt wybrany ze zbioru D.
- W ogólnym wypadku postępuje się w następujący sposób. Oblicza się odległość obiektu X od wszystkich obiektów zbioru treningowego. Następnie ze zbioru D wybiera się k najbliższych „sąsiadów” obiektu X (czyli k obiektów położonych najbliżej obiektu X). Obiektowi X przyporządkowuje się tę klasę atrybutu decyzyjnego C, która jest najliczniej reprezentowana w zbiorze k najbliższych „sąsiadów” tego obiektu.

Kluczowe znaczenie w metodzie k-najbliższych „sąsiadów" ma wybór miary odległości pomiędzy obiektami. W praktyce stosuje się zazwyczaj jeden z wariantów **metryki (odległości) Minkowskiego** postaci

$$D(X,Y) \stackrel{def}{=} \left(\sum_{i=1}^{r} |x_i - y_i|^p\right)^{1/p},$$

gdzie $X = (x_1, ..., x_r)$, $Y = (y_1, ..., y_r)$, natomiast $p \geq 1$. W przypadku $p = 1$ otrzymujemy odległość miejską (Manhattan), natomiast dla $p = 2$ - odległość euklidesową.

Sieci bayesowskie (1)

- **Sieć bayesowska** to acykliczny graf skierowany składający się z wierzchołków (tzw. **węzłów sieci**) i łączących je krawędzi.

Sieci bayesowskie (1)

- **Sieć bayesowska** to acykliczny graf skierowany składający się z wierzchołków (tzw. **węzłów sieci**) i łączących je krawędzi.
- Pojedynczy węzeł sieci odpowiada dokładnie jednemu z atrybutów (zmiennych losowych typu dyskretnego) A_i, $i = 1, ..., s$ (s jest liczbą węzłów).

- **Sieć bayesowska** to acykliczny graf skierowany składający się z wierzchołków (tzw. **węzłów sieci**) i łączących je krawędzi.
- Pojedynczy węzeł sieci odpowiada dokładnie jednemu z atrybutów (zmiennych losowych typu dyskretnego) A_i, $i = 1, ..., s$ (s jest liczbą węzłów).
- Zakładać będziemy, że wszystkie zmienne (atrybuty) A_i, $i = 1, ..., s$, określone są na pewnej wspólnej przestrzeni probabilistycznej $(\Omega, \mathcal{R}, \mathbf{P})$, przy czym przez V_i oznaczać będziemy zbiór wartości zmiennej losowej A_i.

Sieci bayesowskie (1)

- **Sieć bayesowska** to acykliczny graf skierowany składający się z wierzchołków (tzw. **węzłów sieci**) i łączących je krawędzi.
- Pojedynczy węzeł sieci odpowiada dokładnie jednemu z atrybutów (zmiennych losowych typu dyskretnego) A_i, $i = 1, ..., s$ (s jest liczbą węzłów).
- Zakładać będziemy, że wszystkie zmienne (atrybuty) A_i, $i = 1, ..., s$, określone są na pewnej wspólnej przestrzeni probabilistycznej $(\Omega, \mathcal{R}, \mathsf{P})$, przy czym przez V_i oznaczać będziemy zbiór wartości zmiennej losowej A_i.
- Krawędź sieci skierowaną od węzła A_i do węzła A_j (oznaczamy ją przez $A_i \rightarrow A_j$) utożsamiamy z pewnego rodzaju przyczynową zależnością zmiennej losowej A_j od zmiennej losowej A_i.

Sieci bayesowskie (2)

Następnik i poprzednik węzła

Mówimy, że w sieci bayesowskiej węzeł A_j jest **następnikiem** węzła A_i (lub też, równoważnie, że węzeł A_i jest **poprzednikiem** węzła A_j), jeżeli spełniony jest jeden z następujących warunków:

- w sieci istnieje krawędź skierowana od węzła A_i do węzła A_j (zatem $A_i \to A_j$);
- w sieci istnieje krawędź skierowana od węzła A_i do pewnego węzła $A_k \neq A_j$, a węzeł A_j jest następnikiem węzła A_k.

Sieci bayesowskie (3)

- Dla dowolnego węzła A_i zbiór numerów węzłów sieci będących jego bezpośrednimi poprzednikami oznaczamy przez U_i.

Sieci bayesowskie (3)

- Dla dowolnego węzła A_i zbiór numerów węzłów sieci będących jego bezpośrednimi poprzednikami oznaczamy przez U_i.
- Zbiór bezpośrednich poprzedników węzła A_i nazywamy **„rodzicami"** węzła A_i i oznaczamy przez **Pa**(A_i).

Sieci bayesowskie (3)

- Dla dowolnego węzła A_i zbiór numerów węzłów sieci będących jego bezpośrednimi poprzednikami oznaczamy przez U_i.
- Zbiór bezpośrednich poprzedników węzła A_i nazywamy **„rodzicami"** węzła A_i i oznaczamy przez $\mathbf{Pa}(A_i)$.
- Zbiór numerów węzłów sieci będących bezpośrednimi następnikami węzła A_i oznaczamy przez S_i.

Sieci bayesowskie (4)

- Wprowadzenie relacji następstwa węzła umożliwia interpretację występujących w sieci bayesowskiej krawędzi.

Sieci bayesowskie (4)

- Wprowadzenie relacji następstwa węzła umożliwia interpretację występujących w sieci bayesowskiej krawędzi.
- W sieci bayesowskiej zakłada się, że **każdy węzeł jest warunkowo niezależny od wszystkich węzłów, które nie są jego następnikami, przy danych wartościach „rodziców" tego węzła.**

Sieci bayesowskie (4)

- Wprowadzenie relacji następstwa węzła umożliwia interpretację występujących w sieci bayesowskiej krawędzi.
- W sieci bayesowskiej zakłada się, że **każdy węzeł jest warunkowo niezależny od wszystkich węzłów, które nie są jego następnikami, przy danych wartościach „rodziców” tego węzła.**

- Ponieważ $U_i \subseteq \{1, \ldots, s\} \setminus S_i$, warunek ten możemy zapisać w następujący sposób:

$$\mathbf{P}(A_i \mid \{1, \ldots, s\} \setminus S_i) = \mathbf{P}(A_i \mid U_i) \tag{1}$$

dla wszystkich $i = 1, \ldots, s$.

Sieci bayesowskie (5)

- Sieć bayesowska sama w sobie nie „wpływa” na występowanie (powstawanie) zależności między zmiennymi lub też brak takich zależności. Traktować należy ją jedynie jako ilustrację przyjętych przez nas założeń dotyczących powiązań pomiędzy zmiennymi (węzłami), które nie muszą być w praktyce spełnione.

Sieci bayesowskie (5)

- Sieć bayesowska sama w sobie nie „wpływa” na występowanie (powstawanie) zależności między zmiennymi lub też brak takich zależności. Traktować należy ją jedynie jako ilustrację przyjętych przez nas założeń dotyczących powiązań pomiędzy zmiennymi (węzłami), które nie muszą być w praktyce spełnione.
- W praktyce rozpatrywać będziemy jednak wyłącznie sieci, dla których założenia te są spełnione. Sieci takie nazywać będziemy sieciami o **poprawnej strukturze**, co oznacza, że ich struktura poprawnie reprezentuje zależności występujące w rozważanej przez nas dziedzinie.

Sieci bayesowskie (6)

Sieć bayesowska o poprawnej strukturze

Mówimy, że sieć bayesowska ma poprawną strukturę, jeśli dla każdego węzła A_i spełniony jest warunek

$$\mathbf{P}\Big(\{A_i = v_i\} \mid \bigcap_{j \in \{1,\ldots,s\} \setminus S_i} \{A_j = v_j\}\Big) = \mathbf{P}\Big(\{A_i = v_i\} \mid \bigcap_{j \in U_i} \{x_j = v_j\}\Big) \qquad (2)$$

dla wszystkich wartości $v_i \in V_i$ oraz $v_j \in V_j$.

Warunek (2) mówi, że w sieci o poprawnej strukturze rozkład prawdopodobieństwa zmiennej losowej A_i zależy tylko od rozkładu prawdopodobieństwa jego bezpośrednich poprzedników (a jest niezależny od rozkładu prawdopodobieństwa węzłów „historycznie” wcześniejszych w sieci).

Sieci bayesowskie (7)

Reguła iloczynu dla sieci bayesowskiej

Jeśli sieć bayesowska ma poprawną strukturę, to

$$\mathbf{P}\left(\bigcap_{i=1}^{s}\{A_i = v_i\}\right) = \prod_{i=1}^{s} \mathbf{P}\left(\{A_i = v_i\} \mid \bigcap_{j \in U_i}\{A_j = v_j\}\right) \qquad (3)$$

dla wszystkich wartości $v_i \in V_i$, $i = 1, \ldots, s$.

Sieci bayesowskie (8)

- Rozpatrzmy następujący problem. W domu osoby A zainstalowano antywłamaniowy system alarmowy, reagujący jednak czasami na wstrząsy sejsmiczne.

- Rozpatrzmy następujący problem. W domu osoby A zainstalowano antywłamaniowy system alarmowy, reagujący jednak czasami na wstrząsy sejsmiczne.
- Osoba A ma dwóch niepracujących sąsiadów: osoby B i C, które obiecały jej powiadomić ją telefonicznie, gdy tylko usłyszą alarm, a osoba A będzie w tym czasie w pracy.

Sieci bayesowskie (8)

- Rozpatrzmy następujący problem. W domu osoby A zainstalowano antywłamaniowy system alarmowy, reagujący jednak czasami na wstrząsy sejsmiczne.
- Osoba A ma dwóch niepracujących sąsiadów: osoby B i C, które obiecały jej powiadomić ją telefonicznie, gdy tylko usłyszą alarm, a osoba A będzie w tym czasie w pracy.
- Osoba B prawie zawsze dzwoni do osoby A, gdy usłyszy alarm, lecz czasem myli dźwięk dzwoniącego telefonu z dźwiękiem alarmu i wtedy też dzwoni.

Sieci bayesowskie (8)

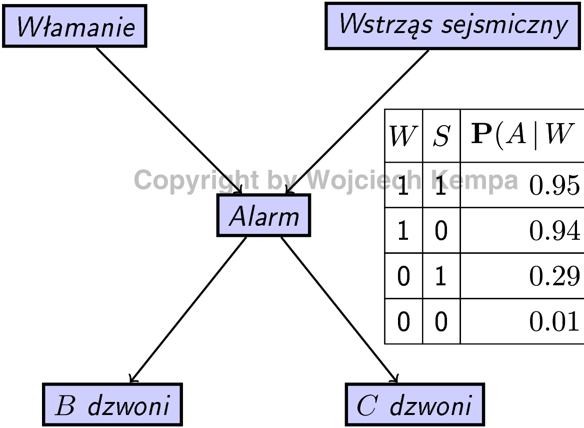
- Rozpatrzmy następujący problem. W domu osoby A zainstalowano antywłamaniowy system alarmowy, reagujący jednak czasami na wstrząsy sejsmiczne.
- Osoba A ma dwóch niepracujących sąsiadów: osoby B i C, które obiecały jej powiadomić ją telefonicznie, gdy tylko usłyszą alarm, a osoba A będzie w tym czasie w pracy.
- Osoba B prawie zawsze dzwoni do osoby A, gdy usłyszy alarm, lecz czasem myli dźwięk dzwoniącego telefonu z dźwiękiem alarmu i wtedy też dzwoni.
- Osoba C z upodobaniem słucha głośnej muzyki i z tego powodu często nie słyszy dźwięku alarmu.

Sieci bayesowskie (9)

- W zadaniu tym można określić następujące zmienne losowe (atrybuty): *Włamanie* (W), *Wstrząs sejsmiczny* (S), *Alarm* (A), B *dzwoni* (B), C *dzwoni* (C).

Sieci bayesowskie (9)

- W zadaniu tym można określić następujące zmienne losowe (atrybuty): *Włamanie* (W), *Wstrząs sejsmiczny* (S), *Alarm* (A), B *dzwoni* (B), C *dzwoni* (C).
- Podany problem można przedstawić za pomocą odpowiedniej sieci bayesowskiej, ustalając hipotetyczne wartości odpowiednich prawdopodobieństw.



$\mathbf{P}(W)$
0.001

$\mathbf{P}(S)$
0.002

W	S	$\mathbf{P}(A \mid W \cap S)$
1	1	0.95
1	0	0.94
0	1	0.29
0	0	0.01

Rysunek: 2: Sieć bayesowska ilustrująca rozważany problem

Sieci bayesowskie (11)

Prawdopodobieństwa zdarzeń $\{B \mid A\}$ oraz $\{C \mid A\}$ w zależności od zajścia zdarzenia A dane są w następujących tabelach:

A	$\mathbf{P}(B \mid A)$
1	0.90
0	0.05

A	$\mathbf{P}(C \mid A)$
1	0.65
0	0.01

Przykładowo, rozpatrzmy następujący rekord (zdarzenie):

$$\{W = 1, S = 0, A = 1\}.$$

Jakie jest prawdopodobieństwo, że osoba B zadzwoni do osoby A? Wykorzystując regułę iloczynu dla sieci bayesowskich, mamy

$$\begin{aligned}&\mathbf{P}(W = 1, S = 0, A = 1, B = 1)\\&= \mathbf{P}(W = 1)\mathbf{P}(S = 0)\mathbf{P}(A = 1 \,|\, W = 1, S = 0)\mathbf{P}(B = 1 \,|\, A = 1)\\&= 0.001 \cdot 0.998 \cdot 0.94 \cdot 0.90 \approx 0.0008.\end{aligned}$$

Odkrywanie asocjacji

- Jednym z ważnych obszarów współczesnej eksploracji danych jest metodologia **odkrywania asocjacji**.

Odkrywanie asocjacji

- Jednym z ważnych obszarów współczesnej eksploracji danych jest metodologia **odkrywania asocjacji**.
- Punktem wyjścia jest tu klasyczny problem związany z analizą koszyka zakupów. Które artykuły (usługi itp.) najczęściej kupowane są łącznie? Czy klient kupujący usługę A zdecyduje się przy okazji na kupno usługi B?

- Jednym z ważnych obszarów współczesnej eksploracji danych jest metodologia **odkrywania asocjacji**.
- Punktem wyjścia jest tu klasyczny problem związany z analizą koszyka zakupów. Które artykuły (usługi itp.) najczęściej kupowane są łącznie? Czy klient kupujący usługę A zdecyduje się przy okazji na kupno usługi B?
- Celem analizy asocjacyjnej (zwanej czasem także **analizą koszykową**) jest znalezienie pewnych wzorców zachowań klientów, a w konsekwencji uzyskanie tzw. **reguł asocjacyjnych**, czyli wzorców zachowań sformalizowanych matematycznie.

Reguła asocjacyjna w naprostszym wypadku to po prostu relacja

$$X \longrightarrow Y,$$

gdzie X i Y są rozłącznymi zbiorami pewnych elementów. X nazywamy **poprzednikiem reguły**, zaś Y – **następnikiem**. W praktyce kluczowa jest ocena jakości skonstruowanej reguły.

Podstawowe pojęcia analizy asocjacji (2)

Stosuje się tu dwie zasadnicze miary:

- **wsparcie reguły** (ang. *support*), oznaczane $supp$, które jest procentowym udziałem transakcji zawierających lewą i prawą stronę (poprzednik i następnik) w zbiorze wszystkich analizowanych transakcji;
- **ufność reguły** (ang. *confidence*), oznaczane $conf$, którą definiuje się jako prawdopodobieństwo warunkowe wystąpienia reguły (procent transakcji zawierających X i Y w zbiorze wszystkich transakcji zawierających X).

Podstawowe pojęcia analizy asocjacji (3)

W procesie odkrywania reguł asocjacyjnych definiuje się często **minimalny próg wsparcia** (ang. *minimum support threshold*) $min\,supp$ oraz **minimalny próg ufności** (ang. *minimum confidence threshold*) $min\,conf$, które „zabezpieczają" jakość wszystkich branych pod uwagę reguł. Reguły spełniające zadany warunek minimalnego progu ufności i wsparcia nazywami **silnymi regułami asocjacyjnymi**.

Podstawowe pojęcia analizy asocjacji (4)

W praktyce regułę asocjacyjną zapisuje się zatem najczęściej w następujący sposób:

$$\text{mleko} \longrightarrow \text{płatki śniadaniowe} \quad [supp = 10\%, conf = 80\%],$$

co oznacza, że w pewnym zbiorze analizowanych przez nas transakcji 80% transakcji zawierających mleko zawierało również płatki śniadaniowe, przy czym transakcji zawierających jeden i drugi produkt w całym analizowanym zbiorze było 10%.

Podział reguł asocjacyjnych (1)

Ze względu na **typ przetwarzanych danych** reguły asocjacyjne dzielimy na

- binarne;
- ilościowe.

Reguła asocjacyjna jest **regułą binarną**, jeżeli obie jej strony (poprzednik i następnik) są zmiennymi binarnymi (przyjmującymi wartość 0 lub 1). Przykładową regułę binarną możemy sformułować, modyfikując regułę przytoczoną wyżej, w następujący sposób:

$$(\text{mleko} = 1) \longrightarrow (\text{płatki śniadaniowe} = 1) \quad [supp = 10\%, conf = 80\%]. \tag{4}$$

Podział reguł asocjacyjnych (2)

- W istocie obie te reguły są tożsame. Można jednak regułę sformułować nieco inaczej, na przykład

$$(\text{mleko} = 1) \longrightarrow (\text{płatki śniadaniowe} = 0) \quad [supp = 10\%, conf = 80\%],$$

co odnosi się do transakcji, w których klient kupujący mleko **nie kupił** równocześnie płatków śniadaniowych.

Podział reguł asocjacyjnych (2)

- W istocie obie te reguły są tożsame. Można jednak regułę sformułować nieco inaczej, na przykład

$$(\text{mleko} = 1) \longrightarrow (\text{płatki śniadaniowe} = 0) \quad [supp = 10\%, conf = 80\%],$$

co odnosi się do transakcji, w których klient kupujący mleko **nie kupił** równocześnie płatków śniadaniowych.

- **Regułą ilościową** nazywamy taką regułę asocjacyjną, w której zmienne są skategoryzowane i/lub ciągłe, np.

$$[(\text{wiek} \in [30, 40]) \wedge (\text{zamieszkanie} = \text{miasto})] \longrightarrow (\text{wykształcenie} = \text{wyższe})$$
$$[supp = 8\%, conf = 70\%]. \quad (5)$$