Zaawansowane uczenie maszynowe:
wykład 11

Paweł Cichosz

Zadanie klasyfikacji wieloetykietowej

Podstawowa koncepcja: wiele niewykluczających się etykiet klas przypisanych do jednego przykładu.

Zbiór klas: C – skończony zbiór możliwych etykiet.

Pojęcie docelowe: $c : X \to 2^C$ – przypisuje każdemu przykładowi podzbiór możliwych etykiet.

Model: $h : X \to 2^C$ – przypisuje każdemu przykładowi podzbiór możliwych etykiet.

Kodowanie binarne: $c(x)$, $h(x)$ reprezentowane jako $|C|$-elementowe wektory binarne, w których 1 na pozycji odpowiadającej etykiecie d oznacza odpowiednio, że $d \in c(x)$ lub $d \in h(x)$ – klasyfikacja wieloetykietowa może być potraktowana jako jednoczesna realizacja $|C|$ zadań klasyfikacji binarnej.

Metody transformacji klasyfikacji wieloetykietowej

Metoda binarnej relewantności: oddzielny model klasyfikacji binarnej dla każdej możliwej etykiety, predykcja zwraca wszystkie etykiety, dla których odpowiednie klasyfikatory binarne dały predykcję pozytywną.

Metoda łańcucha klasyfikatorów: sekwencyjnie tworzone i stosowane modele klasyfikacji binarnej, predykcja poprzedniego modelu staje się dodatkowym atrybutem dla kolejnego modelu.

Metoda zbioru potęgowego etykiet: model klasyfikacji wieloklasowej, dla którego klasami są wszystkie możliwe podzbiory zbioru etykiet (bezpośrednie potraktowanie klasyfikacji wieloetykietowej jako klasyfikacji wieloklasowej).

Metoda binarnej relewantności

Transformacja pojęcia: dla każdej etykiety $d \in \{d_1, d_2, \dots\}$:

$$c_d(x) = \begin{cases} 1 & \text{jeśli } d \in c(x) \\ 0 & \text{jeśli } d \notin c(x) \end{cases}$$

Tworzenie modelu: tworzone modele klasyfikacji binarnej h_d dla wszystkich $d \in \{d_1, d_2, \dots\}$.

Predykcja: wyznaczane predykcje $h_d(x)$ dla wszystkich $d \in \{d_1, d_2, \dots\}$, na tej podstawie:

$$h(x) = \{d \in C \mid h_d(x) = 1\}$$

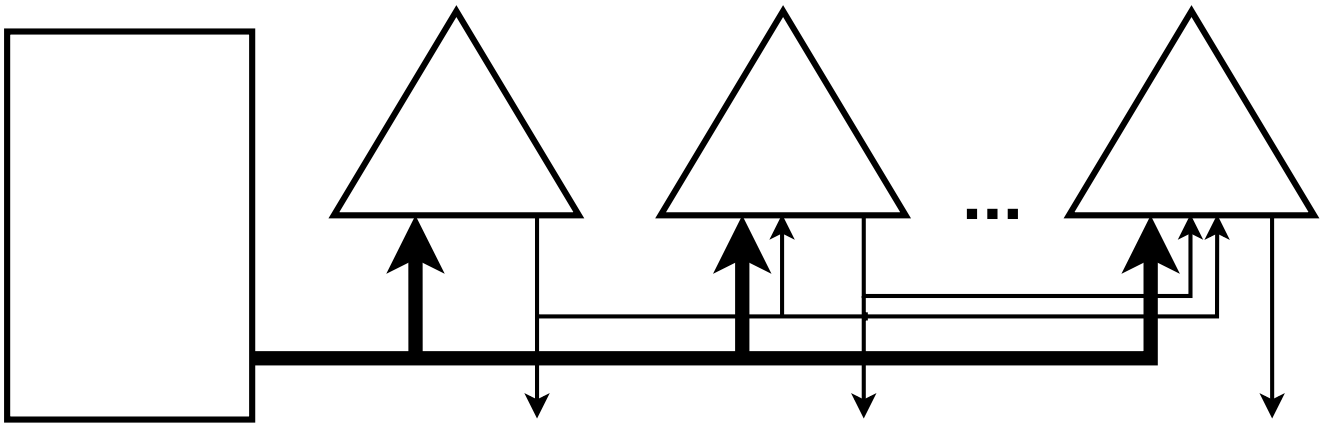
Ograniczenia: brak wykorzystania ewentualnych zależności między etykietami.

Metoda łańcucha klasyfikatorów

Transformacja pojęcia: dla każdej etykiety $d \in \{d_1, d_2, \dots\}$:

$$c_d(x) = \begin{cases} 1 & \text{jeśli } d \in c(x) \\ 0 & \text{jeśli } d \notin c(x) \end{cases}$$

Tworzenie modelu: sekwencyjnie tworzone modele klasyfikacji binarnej h_d dla wszystkich $d \in \{d_1, d_2, \dots\}$, przy czym do zestawu atrybutów wejściowych dla modelu h_{d_i} dołączane są predykcje modeli h_{d_j} dla $j < i$.



Metoda łańcucha klasyfikatorów

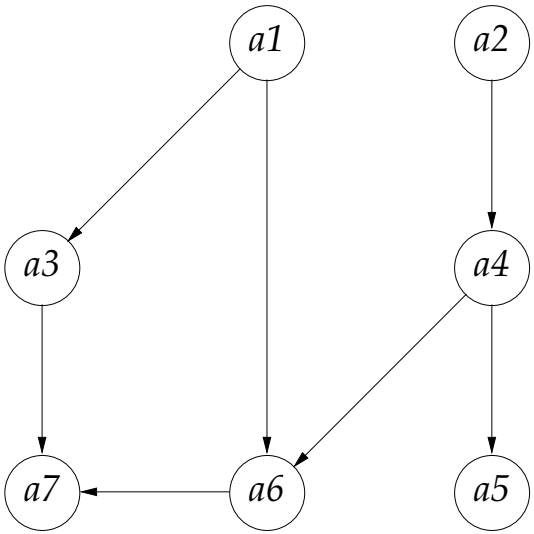
Predykcja: wyznaczane sekwencyjnie predykcje $h_d(x)$ dla wszystkich $d \in \{d_1, d_2, \dots\}$, na tej podstawie:

$$h(x) = \{d \in C \mid h_d(x) = 1\}$$

Kolejność etykiet: arbitralna, losowa lub dobrana heurystycznie (np. według łatwości predykcji, częstości występowania, zależności).

Warianty rozwinięte: m.in. bayesowski łańcuch klasyfikatorów, zespół łańcuchów klasyfikatorów.

Sieci bayesowskie



Węzły: odpowiadają atrybutom $a_1, a_2, \ldots, a_n$ określonym na dziedzinie.

Krawędzie: $a_i \to a_j$ reprezentuje bezpośrednią zależność przyczynową a_j od a_i.

Prawdopodobieństwa warunkowe: w każdym węźle a_i przechowywane prawdopodobieństwa warunkowe jego wartości przy danych wartościach jego bezpośrednich poprzedników U_i:

$$P(a_i = v_i \mid a_j = v_j, j \in U_i)$$

Warunkowa niezależność: każdy węzeł warunkowo niezależny od wszystkich węzłów niebędących jego następnikami, przy danych wartościach jego bezpośrednich poprzedników.

Identyfikacja struktury sieci bayesowskiej (Chow-Liu)

Wyznaczenie wag potencjalnych krawędzi: informacja wzajemna $I_S(a_1, a_2)$.

Wyznaczenie zbioru krawędzi: drzewo rozpinające o maksymalnej sumie wag (np. algorytm Kruskala, algorytm Prima).

Ustalenie zwrotów krawędzi: od wybranego węzła początkowego.

Szczegóły, inne podejścia: poza zakresem wykładu.

Bayesowski łańcuch klasyfikatorów

Identyfikacja zależności klas: sieć bayesowska opisująca zależności między etykietami na zbiorze trenującym.

Kolejność etykiet w łańcuchu: porządek topologiczny węzłów sieci bayesowskiej opisującej zależności między etykietami.

Możliwe uproszczenie: uwzględnienie na wejściu każdego modelu tylko tych etykiet odpowiadających wcześniejszym modelom, od których etykieta tego modelu okazała się bezpośrednio zależna.

Zespół łańcuchów klasyfikatorów

Tworzenie modelu: oddzielne łańcuchy klasyfikatorów tworzone dla wielu losowych permutacji klas.

Predykcja: agregacja predykcji dla poszczególnych klas przez głosowanie odpowiednich modeli ze wszystkich łańcuchów.

Ocena jakości klasyfikacji wieloetykietowej

Wieloetykietowa macierz pomyłek dla klas:

TP_d: $|\{x \in S \mid d \in c(x) \wedge d \in h(x)\}|$
TN_d: $|\{x \in S \mid d \notin c(x) \wedge d \notin h(x)\}|$
FP_d: $|\{x \in S \mid d \notin c(x) \wedge d \in h(x)\}|$
FN_d: $|\{x \in S \mid d \in c(x) \wedge d \notin h(x)\}|$

Wieloetykietowa macierz pomyłek dla przykładów:

TP_x: $|\{d \in C \mid d \in c(x) \wedge d \in h(x)\}|$
TN_x: $|\{d \in C \mid d \notin c(x) \wedge d \notin h(x)\}|$
FP_x: $|\{d \in C \mid d \notin c(x) \wedge d \in h(x)\}|$
FN_x: $|\{d \in C \mid d \in c(x) \wedge d \notin h(x)\}|$

Standardowe miary jakości oparte na macierzy pomyłek: współczynnik prawdziwych/fałszywych pozytywnych, precyzja, odzysk itp.

Agregacja oceny: mikro-/makro-uśrednianie.

Krzywe ROC/PR: dekompozycja „1 kontra reszta” (one vs. rest, OvR).

Klasyfikacja wieloklasowo-wielowyjściowa

Podstawowa koncepcja: wiele jednocześnie realizowanych zadań klasyfikacji wieloklasowej.

Zbiory klas: C_i – skończony zbiór możliwych etykiet pojęcia c_i dla $i = 1, 2, \ldots, m$.

Pojęcia docelowe: $c_i : X \to C_i$ dla $i = 1, 2, \ldots, m$.

Model: $h : X \to C_1 \times C_2 \times \cdots \times C_m$ – przypisuje każdemu przykładowi wektor klas.

Kodowanie binarne: $c(x)$, $h(x)$ reprezentowane jako $|C|$-elementowe wektory binarne, w których 1 na pozycji odpowiadającej etykiecie d oznacza odpowiednio, że $d \in c(x)$ lub $d \in h(x)$.

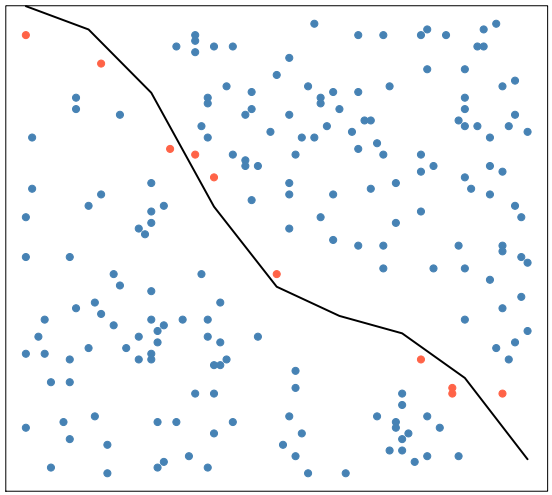
Aktywne uczenie się

Cel: ograniczenie pracochłonności lub kosztu dostarczenia danych etykietowanych w zadaniach klasyfikacji.

Założenie: dostępna duża pula przykładów nieetykietowanych i mały lub pusty zbiór przykładów etykietowanych.

Inicjalizacja: wykorzystanie dostępnego zbioru przykładów etykietowanych lub wybór przykładów z puli i żądanie dostarczenia ich etykiet w celu uzyskania początkowego zbioru trenującego do utworzenia pierwszego modelu.

Zapytania: iteracyjnie powtarzane żądania dostarczenia etykiet niewielkiej liczby wybranych przykładów z puli, dołączenie nowych etykietowanych przykładów do zbioru trenującego i utworzenie kolejnego modelu.



Scenariusz aktywnego uczenia się

1: wybierz początkowy zbiór przykładów trenujących T z puli P;
2: $P := P - T$;
3: uzyskaj etykiety klas dla T;
4: **repeat**
5: utwórz model h używając zbioru trenującego T;
6: wybierz zbiór przykładów do zapytania $Q \subset P$;
7: uzyskaj etykiety klas dla Q;
8: $T := T \cup Q$; $P := P - Q$;
9: **until** osiągnięto maksymalną liczbę iteracji lub $P = \emptyset$ lub spełnione są kryteria stopu;
10: **return** h.

Wybór przykładów do zapytań

Na podstawie niepewności: wybierane przykłady, dla których niepewność predykcji modelu jest największa (ufność jest najmniejsza).

Na podstawie odległości od granicy decyzyjnej: wybierane przykłady położone najbliżej granicy decyzyjnej.

Na podstawie zróżnicowania: wybierane przykłady, które powodują największe zróżnicowanie zbioru trenującego.

Na podstawie gęstości: wybierane przykłady leżące w obszarach o najwyższej gęstości.

Na podstawie wpływu na model: wybierane przykłady, które mogą w największym stopniu wpłynąć na zmianę modelu.

Miary niepewności

Entropia:

$$\mathrm{unc\_ent}(p) = \sum_d -P(d|x) \log P(d|x)$$

Najmniejsza ufność:

$$\mathrm{unc\_lc}(p) = 1 - P(d_1|x)$$

Margines ufności:

$$\mathrm{unc\_cm}(p) = 1 - (P(d_1|x) - P(d_2|x))$$

Iloraz ufności:

$$\mathrm{unc\_cr}(p) = \frac{P(d_2|x)}{P(d_1|x)}$$

- d_1 – klasa najbardziej prawdopodobna,
- d_2 – następna klasa najbardziej prawdopodobna.

Kryteria stopu

Na podstawie jakości predykcji: jakość predykcji jest wystarczająca lub przestała się poprawiać – trudne do zastosowania w praktyce (wymagałoby pozostawienia części danych etykietowanych do oceny modelu).

Na podstawie niepewności: niepewność predykcji modelu jest wystarczająco niska lub przestała spadać.

- Np. niska maksymalna niepewność mierzona za pomocą entropii.

Na podstawie stabilności: stabilność predykcji modelu jest wystarczająco wysoka.

- Np. wysoka korelacja między predykcjami prawdopodobieństw klas modeli z kolejnych iteracji.

Półnadzorowane uczenie się

Cel: Poprawienie jakości modeli klasyfikacji uzyskiwanych na podstawie małych zbiorów trenujących.

Założenie: dostępna duża pula przykładów nieetykietowanych i mały przykładów etykietowanych.

Możliwe podejścia:

- Wykorzystanie danych nieetykietowanych do poprawienia estymacji rozkładu prawdopodobieństwa klas: preferowanie modeli dających podobne predykcje dla podobnych przykładów nieetykietowanych.
- Wykorzystanie danych nieetykietowanych do poprawienia położenia granicy decyzyjnej: preferowanie modeli nierozdzielających skupisk przykładów nieetykietowanych.
- Wykorzystanie danych nieetykietowanych do doboru reprezentacji (np. selekcji lub transformacji atrybutów, funkcji jądrowej itp.).
- Wykorzystanie danych nieetykietowanych do generowania dodatkowych przykładów trenujących przez samoetykietowanie (*self-training*) lub współetykietowanie (*co-training*).

Samoetykietowanie

1: **repeat**
2: utwórz model h używając zbioru trenującego T;
3: wyznacz predykcje modelu h dla przykładów z puli P;
4: wybierz zbiór przykładów $S \subset P$ do samoetykietowania o najmniejszej niepewności predykcji;
5: $T := T \cup S$; $P := P - S$;
6: **until** osiągnięto maksymalną liczbę iteracji lub $P = \emptyset$ lub spełnione są kryteria stopu;
7: **return** h.

Współetykietowanie

- Oddzielne modele tworzone z użyciem różnych zestawów atrybutów, o których zakłada się, że są warunkowo niezależne względem klas i wystarczające do dobrej predykcji.
- Najmniej niepewne (najbardziej ufne) predykcje każdego modelu wykorzystywane jako etykiety dodatkowych przykładów dla pozostałych modeli.
- Modele wykorzystywane jako zespół, którego połączone predykcje wyznaczane są z uwzględnieniem ufności.
- Skuteczne jeśli poszczególne modele składowe są wystarczająco zróżnicowane (dają różne predykcje dla przykładów nieetykietowanych).

Inne warianty zadania klasyfikacji

Transfer learning: wykorzystywanie podczas tworzenia modelu wiedzy uzyskanej przy tworzeniu modeli dla innych pokrewnych zadań.

Multi-task learning: tworzenie jednego modelu dla wielu pokrewnych zadań.

Zero-, one-, few-shot learning: wykorzystanie dodatkowej wiedzy dziedzinowej (np. opisów klas lub podobieństwa między nimi) i danych nieetykietowanych do tworzenia modelu zdolnego do predykcji klas, które nie były reprezentowane w zbiorze trenującym, były reprezentowane przez pojedyncze przykłady lub były reprezentowane przez niewielką liczbę przykładów.

- Obecnie stosowane najczęściej z algorytmami głębokiego uczenia się, ale istnieją również inne podejścia (np. oparte na metodach bayesowskich lub jądrowych).
- Poza zakresem wykładu.