Kowalczewski Mikołaj
Roszkowski Piotr

Dokumentacja wstępna do projektu z przedmiotu Zaawansowane Uczenie Maszynowe

Temat nr 21: Nie-całkiem-naiwny klasyfikator bayesowski typu AODE

Interpretacja tematu projektu

Celem projektu jest implementacja funkcji do tworzenia modelu oraz predykcji przy użyciu nie-całkiem-naiwnego klasyfikatora bayesowskiego typu AODE (*averaged one-dependence estimators*) w języku R, a następnie porównanie działania ze standardowym klasyfikatorem bayesowskim oraz wybranymi algorytmami klasyfikacji dostępnymi w R (lasem losowym, SVM oraz algorytmem k-NN), w którym zadanie będzie wykonywane - w tym celu zostanie użyte kilka zestawów danych pobranych z repozytorium UCI.

Opis użytych algorytmów

Naiwny klasyfikator bayesowski

Naiwny klasyfikator Bayesa to klasyfikator opierający swoje przewidywania na rachunku prawdopodobieństwa wykorzystujący twierdzenie Bayesa:

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)}$$

lub inaczej:

$$P(X|Y) = \frac{P(Y \wedge X) \cdot P(X)}{P(Y) \cdot P(X)} = \frac{P(Y \wedge X)}{P(Y)}$$

Jego „naiwność” polega na założeniu o niezależności atrybutów, tzn.
nie istnieje pomiędzy ich wartościami żaden związek.

Jeżeli chce się określić prawdopodobieństwo przyporządkowania danej pozycji do klasy $c = d$, na podstawie wartości atrybutów: $a_1 = v_1, a_2 = v_2, a_3 = v_3$, …, $a_n = v_n$ to zamiast liczyć $P(c = d \mid a_1 = v_1, \ldots, a_n = v_n)$ można obliczyć $P(a_1 = v_1, \ldots, a_n = v_n \mid c = d)$, a dzięki założeniu o niezależności atrybutów można założyć:
$P(a_1 = v_i, \ldots, a_n = v_n \mid c = d) = \prod_{i=1}^{n} P(a_i = v_i \mid c = d)$
Na etapie klasyfikacji wybieramy opcję o najwyższym prawdopodobieństwie zajścia.

Nie-całkiem-naiwny klasyfikator bayesowski typu AODE

Opierające się na strukturze sieci bayesowskich rozwiązanie, gdzie uwzględnia się ewentualne powiązanie pomiędzy atrybutami, czyli:
$P(a_1 = v_i, \ldots, a_n = v_n) = \prod_{i=1}^{n} P(a_i = v_i \mid a_{U_i} = v_{U_i})$ gdzie U_i oznacza zbiór atrybutów, które w sieci bayesowskiej wpływają na wartość atrybutu a_i

Klasyfikator bayesowski typu AODE zakłada, że przy obliczaniu:

$$P(a_1 = v_i, \ldots, a_n = v_n \mid c = d)$$

stosuje się uśrednione prawdopodobieństwo obliczane dla każdego atrybutu a_j przy założeniu, że pozostałe atrybuty zależą tylko od c, oraz od tegoż atrybuty a_j

Pozwala to ograniczyć sytuacje, gdy atrybuty, których wartości w istotnym stopniu zależą od siebie, wpływają na ostateczny wynik tak samo, jak atrybuty niezależne.

Las losowy

Metoda klasyfikacji z użyciem lasu losowego należy do rodziny metod używających drzew decyzyjnych, czyli struktur do przechowywania informacji złożonych z węzłów, w których sprawdzane są warunki dotyczące danej obserwacji i liści, które zawiera klasę odpowiadającą pewnemu zestawowi atrybutów.

Tworzenie drzew składających się na las losowy przebiega w następujący sposób:

- losujemy ze zwracaniem podzbioru danych ze zbioru danych uczących,
- tworzymy drzewo dla wylosowanego podzbioru - losowana jest pewna liczba zmiennych objaśniających oraz znajdowany jest najlepszy podział z wykorzystaniem tych zmiennych;
- klasa dla przypadków w zbiorze testowym odpowiada klasom wynikającym ze stworzonych drzew ze zbioru uczącego - wybierana jest opcja dominujaca;
- jeśli liczba drzew osiągnie wartość maksymalną (zdefiniowaną jako parametr algorytmu) lub błąd w próbie testowej przestanie maleć, uczenie należy zakończyć, w przeciwnym wypadku wracamy do pierwszego kroku.

Maszyna wektorów nośnych (SVM)

Istotą algorytmu opartego na SVM jest wyznaczenie hiperpłaszczyzny, która daje największy margines pomiędzy zbiorami danych. W zależności od tego, czy dane są separowalne lub nie, margines ten może być twardy (nie dopuszcza żadnych odstępstw) lub miękki (odstępstwa od klasyfikacji skutkują karą, jednak ciągle są dopuszczalne).

W ogólności podczas liniowej klasyfikacji minimalizowana jest funkcja błędu, postaci chociażby:

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0,1-y_i(w\cdot x_i-b)\right)\right]+\lambda\|w\|^2.$$

gdzie:

- w jest wektorem szukanych współczynników hiperpłaszczyzny,
- b jest szukaną stałą w równaniu hiperpłaszczyzny,
- y są etykietami klas (1 lub -1),
- x jest wektorem danych punktów zbioru danych,
- lambda jest parametrem skalującym karę za punkty niepoprawnie sklasyfikowane (nie leżą po odpowiedniej stronie hiperpłaszczyzny) - im mniejsza wartość tego parametru, tym zadanie coraz bardziej przypomina SVM z twardym marginesem.

W przypadku występowania większej ilości klas niż dwie, sposób podziałów danych do klasyfikacji może odbyć się na dwa sposoby:

- jeden do jeden - w każdym wywołaniu dokonujemy klasyfikacji tylko dla obiektów dwóch klas, powtarzamy dla każdej kombinacji, a następnie wyniki uśredniamy;
- jeden do wielu - dla każdej klasy dokonujemy klasyfikacji wszystkich przypadków, gdzie możliwymi decyzjami jest: należy do klasy, nie należy do klasy;

Algorytm k najbliższych sąsiadów

W algorytmie k-NN w procesie uczenia zapamiętywane są jedynie wszystkie przypadki uczące. Natomiast klasyfikacja nowych przypadków przebiega w następujący sposób:

1. Oblicza się wartość odległości pomiędzy nowym przypadkiem a zapamiętanymi przypadkami uczącymi;
2. Po posortowaniu odległości rosnąco brane jest pierwszych k pozycji (k jest parametrem algorytmu) jako najbliżsi sąsiedzi;
3. Wśród tych przypadków wybieramy decyzję, która występuje najczęściej (w przypadku remisu wybieramy dowolną).

Odległość pomiędzy dwoma przypadkami możemy wyliczać jako np. odległość euklidesową (pierwiastek z sumy kwadratów różnic dla każdego atrybutu), odległość taksówkową (bezwzględne różnice pomiędzy poszczególnymi wartościami atrybutów) lub inne.
W realizacji zadania odległość będzie wyliczana jako liczba atrybutów, których wartość jest różna z uwagi na dyskretne dziedziny atrybutów wykorzystywanych w wybranych zbiorach danych.

Plan badań

Cel eksperymentów oraz charakterystyka zbiorów danych

Wszystkie zestawy danych będą pobierane z repozytorium UCI.

W pierwszej kolejności nastąpi weryfikacja poprawności zaimplementowanych algorytmów związanych z klasyfikatorami bayesowskimi. W tym celu zostaną użyte trywialne zestawy danych tj. z małą ilością przykładów i małą ilością atrybutów, np. zestaw danych “Balloons” z 16 przykładami oraz 4 atrybutami. W przypadku potwierdzenia dobrego wyniku klasyfikatorów zostaną one porównane z wcześniej omówionymi algorytmami dostępnymi w języku R dla bardziej złożonych zestawów danych:

- Connect-4 - 67757 przykładów, 42 atrybuty;
- Chess (King-Rook vs. King-Pawn) - 3196 przykładów, 36 atrybutów;

Wybrane zestawy danych są kompletne oraz posiadają dyskretną dziedzinę.

Parametry algorytmów, których wpływ na wyniki będzie badany

W przypadku braków (małej ilości) w wartościach atrybutów warunkowych przewiduje się stosowania wygładzenia Laplace’a - będzie badana jego obecność oraz waga na wpływ klasyfikacji.

Miary jakości i procedury oceny modeli

Do oceny modeli będą stosowane następujące miary jakości:

- miara F:

$$F = \frac{1}{\frac{\frac{1}{recall}+\frac{1}{precision}}{2}} = \frac{2 \cdot recall \cdot precision}{recall + precision}$$

  gdzie:
  - precision dla klasy A, jest to stosunek poprawnie sklasyfikowanych elementów z A do wszystkich, które nasz klasyfikator oznaczył jako A,
  - recall dla klasy A, jest to stosunek poprawnie rozpoznanych elementów z A do wszystkich, które powinien rozpoznać, czyli do całej klasy A;

- współczynnik kappa

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

  gdzie:
  - p_0 oznacza trafność zaobserwowana (observed) - trafność naszego klasyfikatora,
  - p_e oznacza trafność oczekiwaną (expected) - trafność, której możemy się spodziewać po klasyfikatorze losowym;
- ROC area - średnia arytmetyczna pól pod krzywymi ROC dla każdej klasy;

Modele będą sprawdzane metodą kroswalidacji (walidacji krzyżowej), tj. każdy zbiór danych testowych będzie podzielony na ileś równych części (ilość grup będzie uzależniona od liczebności zbioru danych i zostanie ustalona na etapie implementacji - zależy to od specyfiki zbiorów danych), następnie jedna z nich zostanie użyta jako dane testowe, a pozostałe jako dane uczące. Postępowanie to będzie kontynuowane aż każdy fragment będzie użyty jako dane testowe, a otrzymane miary jakości zostaną wtedy uśrednione.