Wprowadzenie do sztucznej inteligencji - ćwiczenie 4

Igor Kraszewski
310164

Warszawa, Grudzień 2022

Spis treści

1. Zadanie

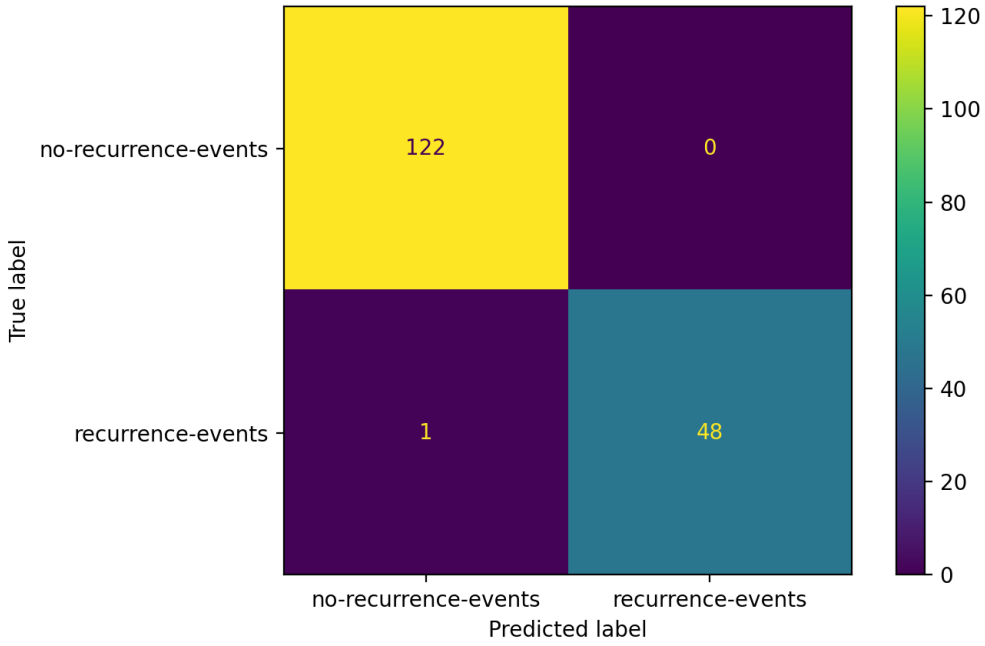
Zaimplementować klasyfikator ID3 (drzewo decyzyjne). Atrybuty nominalne, testy tożsamościowe. Podać dokładność i macierz pomyłek na zbiorach: Breast cancer i mushroom. Dlaczego na jednym zbiorze jest znacznie lepszy wynik niż na drugim? Do potwierdzenia lub odrzucenia postawionych hipotez konieczne może być przeprowadzenie dodatkowych eksperymentów ze zmodyfikowanymi zbiorami danych. Sformułować i spisać wnioski.

Poniżej kilka wskazówek ogólnych do tego ćwiczenia

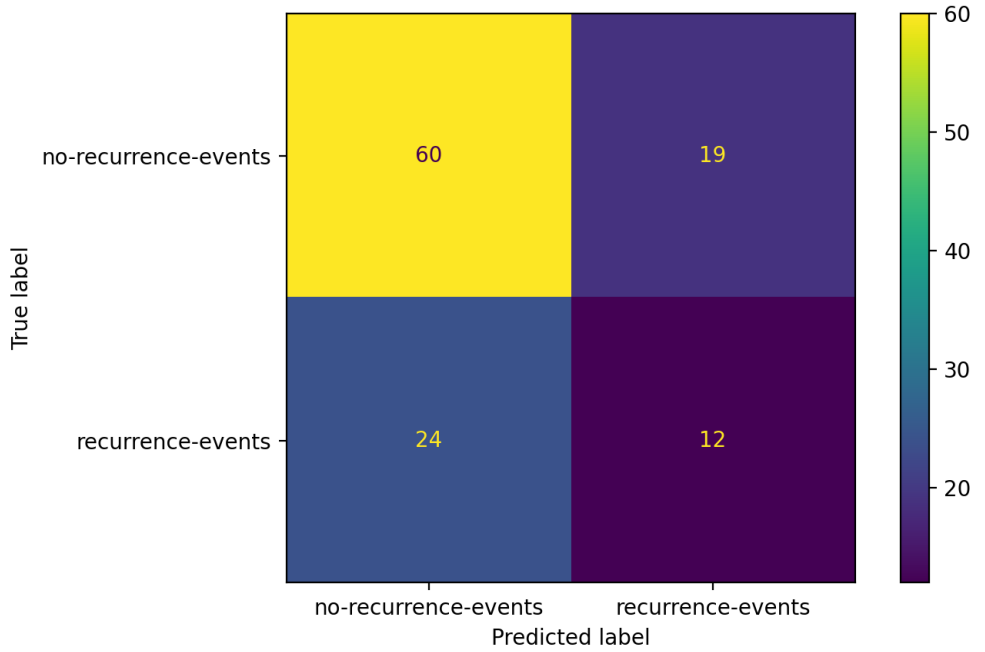
— Atrybuty nominalne - każdy atrybut może przyjmować jedną z kilku dozwolonych wartości, zakładamy, że wartość atrybutu to napis, np. "kot", "a", "20-34", "¿40".
— Testy tożsamościowe - jeżeli atrybut testowany w danym węźle ma np. 3 dozwolone wartości, np. a, b, c, to z węzła tego wychodzą 3 krawędzie oznaczone: a, b, c.
— Na tym ćwiczeniu klasyfikator trenuje się na zbiorze trenującym, a ocenia jego jakość na zbiorze testującym. Należy losowo podzielić zbiór danych na trenujący i testujący w stosunku 3:2.
— Jeżeli zbiór danych zawiera numery lub identyfikatory wierszy to należy je wyrzucić - nie chcemy uczyć się identyfikatorów wierszy.
— Brakujące wartości atrybutów taktujemy jako wartość, np. jeżeli symbol '?' oznacza brakującą wartość, a symbole 'a', 'b' wartości normalne, to z naszego punktu widzenia mamy 3 wartości normalne (fachowo: 3 wartości atrybutu): 'a', 'b', '?'.
— Tak naprawdę to nie musimy rozumieć dziedziny problemu - na wejściu mamy napisy, na wyjściu napisy, nie ważne czy klasyfikujemy sekwencje DNA, grzyby, czy samochody.
— Nazwa pliku ze zbiorem danych jest parametrem algorytmu klasyfikacji, kod klasyfikatora powinien być w stanie obsłużyć inny zbiór danych o tym samym rozkładzie kolumn (czyli nie należy wpisywać wartości atrybutów „na sztywno" w kodzie).
— W repozytorium ze zbiorami danych zwykle w plikach „.names" jest napisane, który atrybut to klasa (czyli wartości której kolumny mamy się nauczyć przewidywać).

2. Wyniki

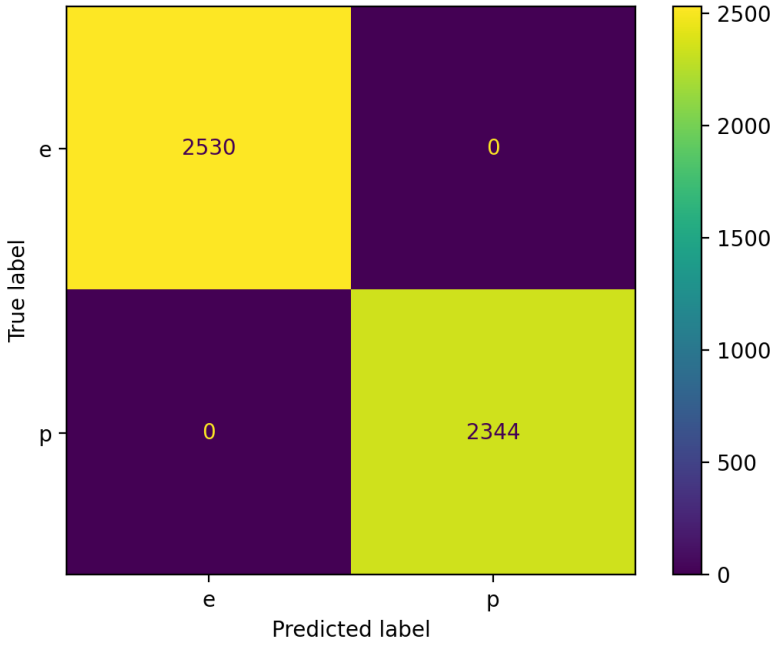
W przypadku zbioru danych "Breast Cancer" otrzymana dokładność na zbiorze treningowym to 99,42%, a na zbiorze testowym to około 61% (zależnie od uruchomienia wyniki wahają się między 59% a 63%, ponieważ została wprowadzona losowość w przypadku kiedy w zbiorze testowym występuje klasa, która nie występowała w zbiorze treningowym). Otrzymane odpowiednie macierze pomyłek [rys. 2.1 i rys. 2.2] pokazują, że w przypadku zbioru treningowego, model myli się tylko raz przyporządkowując klasę "no-recurrence-events", kiedy powinna wystąpić klasa "recurrence-events", co może wynikać z dwóch wystąpień o takich samych atrybutach, różniących się tylko klasą wyjściową. W przypadku zbioru testowego sytuacja wygląda dużo gorzej, model myli się więcej razy, głównie w przypadku przypisania klasy "no-recurrence-events", kiedy wystąpić powinna "recurrence-events", a to może być związane z niezbilansowanych pod względem liczby wystąpień danej klasy zbiorem danych. Dodatkowo błędy mogą wynikać z tego, że ten zbiór danych jest bardzo mały, a po podzieleniu go na zbiór treningowy i testowy w stosunku 3:2 mamy jeszcze mniejsze możliwości treningu. W związku z tym występują sytuacje gdzie w zbiorze testowym znajdują się wystąpienia zawierające w niektórych kolumnach klasy, które nie były obecne w zbiorze treningowym. Model nie działa dobrze, ale najlepiej radzi sobie z wykrywaniem klasy pozytywnej ("no-recurrence-events"), z wykrywaniem klasy negatywnej radzi sobie sporo gorzej. Najwięcej pomyłek polega na fałszywym przypisaniu klasy pozytywnej, jednak problem fałszywego przypisywania klasy negatywnej nie jest dużo mniejszy. Dla zbioru "Mushroom" otrzymana dokładność na zbiorze treningowym jak i testowym to 100%. Z otrzymanych odpowiednich macierzy pomyłek [rys. 2.4 i rys. 2.3] widać, że model nie mylił się ani razu.
Wyniki na zbiorze "Mushroom" są dużo lepsze niż na zbiorze "Breast Cancer". Wynika to z wielkości zbiorów. Pierwszy z nich jest prawie 28 razy większy od drugiego. Posiada on również sporo więcej atrybutów. Daje to możliwość głębszego wytrenowania, znalezienia ważniejszych i sensowniejszych zależności pomiędzy atrybutami a wyjściową klasą, oraz stworzenia modelu, który działa perfekcyjnie.



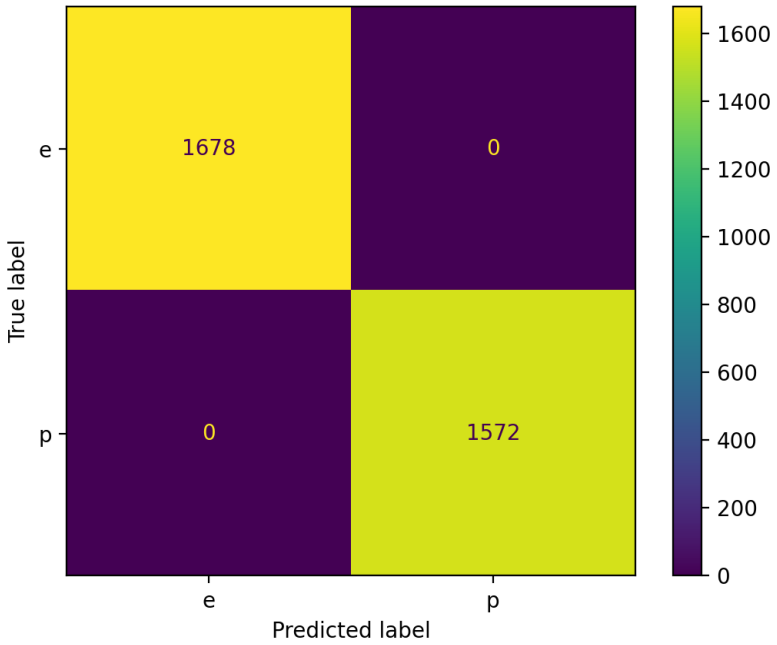
Rys. 2.1. Wyniki predykcji modelu dla danych "Breast Cancer" na zbiorze treningowym



Rys. 2.2. Wyniki predykcji modelu dla danych "Breast Cancer" na zbiorze testowym



Rys. 2.3. Wyniki predykcji modelu dla danych "Mushroom" na zbiorze treningowym



Rys. 2.4. Wyniki predykcji modelu dla danych "Mushroom" na zbiorze testowym