Kolejny rzut oka na pandemię

Załączony zbiór danych zawiera wyniki ankiety dotyczącej stosunku do COVID-19, przeprowadzonej w Indiach przez TKR College of Engineering & Technology. Naszym zadaniem będzie przygotowanie klasyfikatora mającego na celu próbę predykcji, czy ankietowany byłby gotowy przyjąć szczepionkę chroniącą przed wiadomym wirusem.

PRZYGOTOWANIE DANYCH

Zbiór zawiera 46 kolumn. Podobnie jak w przypadku zadania z klasteryzacji, nie wszystkie będą nam potrzebne, inne zaś wymagają przetworzenia lub usunięcia.

- Jedna z kolumn zawiera odpowiedź na pytanie "If a vaccine to prevent COVID-19 was offered to you today, would you choose to be vaccinated?". Ta kolumna to nasz cel predykcyjny. W oryginalnym zbiorze zawiera ona 4 różne wartości "Yes, Definitely", "Yes, Probably", "No, Probably Not", "No, Definitely Not". Dla uproszczenia zadania zamieńmy ją na dwie tylko wartości: „Yes” i „No”. W zależności od preferencji można uznać, że "Yes, Probably" = „Yes” (bliższe rzeczywistości) albo "Yes, Probably" = „No” (poprawia rozkład danych między klasami).
- Kilka kolejnych kolumn zawiera pytania o to, czy na decyzję o szczepieniu wpłynęłaby rekomendacja od przyjaciela lub urzędnika, oraz pytanie o obawy dotyczące skutków ubocznych szczepienia. Usuńmy również te kolumny – są w oczywisty sposób skorelowane z celem predykcyjnym, więc jeżeli je zostawimy, to klasyfikator nie odkryje żadnej ciekawej zależności.
- Część kolumn zawiera tylko jedną unikatową wartość – je też usuwamy, bo w żaden sposób nie pomogą nam odróżniać od siebie obserwacji.

KLASYFIKATOR K-NN

Jednym z dwóch klasyfikatorów, które planujemy wykorzystać w zadaniu jest przedstawiany już wcześniej klasyfikator k-NN. Tym razem mamy jednak do czynienia z danymi kategorycznymi, a nie wektorem liczb. By móc skorzystać z takiej reprezentacji obserwacji musimy zdefiniować własną miarę niepodobieństwa (opisującą dystans między obserwacjami).

- Zaprojektuj funkcję opisującą jak podobne są dwie obserwacje ze zbioru. Wcześniej przyjrzyj się wszystkich cechom, które występują w zbiorze i zastanów jak je w niej uwzględnić. Być może warto rozważyć funkcję zależącą od kilku parametrów (np. jaką wagę mają różnice w odczuwanych objawach, jaką wagę mają różnice w wieku, etc.).
- Przyjmij też stałą k=5. Zbiór jest niewielki, więc może się okazać nieco zbyt duża, ale pozwoli na uzyskanie nieco bardziej obrazowej krzywej ROC – a na tym będziemy się koncentrować w zadaniu.

KLASYFIKATOR RANDOM FOREST

Drugim z rozważanych klasyfikatorów będzie RandomForest (jak zwykle można skorzystać z gotowej implementacji np. ze scikit-learn). W tym przypadku domyślne wartości parametrów są zazwyczaj dobrym wyborem (*gini impurity* jako podstawa do generowania drzew, branie pod uwagę $\sqrt{N}$ cech w jednym drzewie, gdzie N to liczba wszystkich cech w zbiorze). Warto tylko zwiększyć samą liczbę drzew (np. do 500) – w tym przypadku granicą jest tylko dostępny czas na obliczenia, im więcej drzew tym lepiej (ale każde kolejne daje mniejszą korzyść).

DIAGNOSTYKA

Głównym celem zadania jest obserwacja jak oba klasyfikatory radzą sobie z postawionym problemem.

- Procedura testowa będzie analogiczna do tej użytej w zadaniu o k-NN, ale z jedną różnicą – zamiast losowo dzielić zbiór na część treningową i testową wykorzystamy *5-fold cross-validation* (istnieje gotowa implementacja w scikit-learn).
    - Pamiętaj, że samo *cross-validation* też warto powtórzyć kilkakrotnie – efekt zależy przecież od permutacji obserwacji!
    - Tym razem nie będziemy stroili parametrów, nie jest więc potrzebny drugi podział zbioru treningowego na treningowy właściwy i walidacyjny.
- Chcemy zobaczyć jak wyglądać będzie krzywa ROC i powierzchnia pod tą krzywą (AUC).
    - W tym celu musimy mieć możliwość ustalania różnych progów czułości dla obu metod.
        - Zakładamy, że *positive* = brak chęci na przyjęcie szczepienia (na pierwszy rzut oka jest to rzadsze, bardziej nietypowe zjawisko).
        - W przypadku k-NN próg ustalamy poprzez definiowanie ilu sąsiadów musi być *positive* by klasyfikowana obserwacja była uznana za *positive* (domyślnie jest to 50%, ale przecież można przyjąć inne wartości).
        - W przypadku RF ustalamy jaki % drzew musi zwrócić *positive*, by zaklasyfikować tak obserwację (analogicznie – domyślnie jest to 50%, ale przecież można modyfikować tą wartość).
    - W przypadku k-NN przygotujmy krzywą dla dwóch różnych miar niepodobieństwa. W przypadku RF dla wersji domyślnej, oraz takiej gdzie wszystkie drzewa uczą się na wszystkich cechach (zamiast na ich losowych podzbiorze) i gdzie wyłączony jest mechanizm boostrapowania przy wyborze obserwacji uczących dla danego drzewa. Łącznie 4 warianty.
        - Zwizualizuj krzywe ROC. Pamiętaj, by były średnią z kilku podejść, wraz z odpowiednio oznaczonym odchyleniem standardowym.
        - Czy i jakie widzisz między nimi różnice?
        - Jakie ich punkty wydają się dobrym miejscem na ustalenie progu czułości?
            - Jaki jest w nich *precision*?
            - Jaki *recall*?
            - Jakie *accuracy*?
            - Pamiętaj o podaniu odchylenia standardowego tych wielkości!
- RF jest metodą o przyzwoitej interpretowalności. Istnieje szereg technik pozwalających na ustalanie, które cechy miały największe znaczenie przy podejmowaniu decyzji. Najprostszą jest skorzystanie z tzw. *Gini importance* (ponownie: jest gotowe w scikit-learn).
    - Które cechy ze zbioru zostały uznane za najważniejsze przez podstawowy wariant RF?