DRZEWA DECYZYJNE – wstęp teoretyczny

Modele drzew klasyfikacyjnych i regresyjnych (CART, ang. *Classification and Regression Trees*), jak sama nazwa mówi, służą zarówno do rozwiązywania problemów regresyjnych (gdzie zmienną zależną jest cecha ilościowa – ciągła/liczbowa) jak i klasyfikacyjnych (zmienna zależna jakościowa – kategoryczna). Najogólniej, celem analizy z zastosowaniem algorytmu budowy drzew decyzyjnych jest znalezienie zbioru logicznych warunków podziału, typu *jeżeli, to*, prowadzących do jednoznacznego zaklasyfikowania obiektów.

Drzewa decyzyjne służą do wyboru deskryptorów o największym wpływie na modelowaną wielkość (najbardziej znaczących). Technika ta polega na „wzrastaniu drzewa" tj. dzielenia związków na wzajemnie wykluczające się grupy – węzły (ang. *nodes*). Linie łączące węzły nazywa się gałęziami (ang. *branches*). Algorytm rozpoczyna się od węzła głównego – korzenia (ang. *root*) – zawierającego wszystkie związki, które następni dzielone są na węzły podrzędne. Końcowe węzły, które nie podlegają podziałom to liście (ang. *leaves*). Każdy podział określa reguła (próg) uwzględniająca wartości wybranego na danym etapie deskryptora.

Zarówno w przypadku klasycznych modeli jakościowych (SAR, ang. *Structure-Activity Relationships*), jak również modeli ilościowych (QSAR, ang. *Quantitative Structure-Activity Relationships*) związki dzielone są na dwa zbiory – uczący (wykorzystywany do opracowania drzewa decyzyjnego) oraz walidacyjny (służący do oceny zdolności predykcyjnych drzewa decyzyjnego).

W przypadku **drzew klasyfikacyjnych** deskryptory wybierane są pod kątem najmniejszego prawdopodobieństwa błędnej klasyfikacji, co oznacza, że binarny podział wykonywany z opracowaną regułą powinien prowadzić do maksymalnie dwóch jednorodnych grup związków. Prawdopodobieństwo błędnej klasyfikacji mierzy się za pomocą indexu Giniego, wyrażonego wzorem:

$$G = 1 - \sum_{j=1}^{c} \left(\frac{n_j}{n}\right)^2$$

gdzie n_j jest liczbą związków z klasy j zawartych w węźle.

Do weryfikacji zdolności predykcyjnych modeli jakościowych służą miary statystyczne:

Czułość(Sensitivity = TP/(TP+FN)

Swoistość/Specyficzność(Specificity) = TN/(FP+TN)

Precyzja(Precision) = TP/(TP+FP)

$\boldsymbol{F1}\ (harmonic\ mean\ of\ precision\&sensitivity) = (2 \times TP)/(2 \times TP + FP + FN)$

$\boldsymbol{Balanced\ accuracy}\ = (Sensitivity + Specificity)/2$

$\boldsymbol{Balanced\ error}\ = 1 - Balanced\ accuracy$

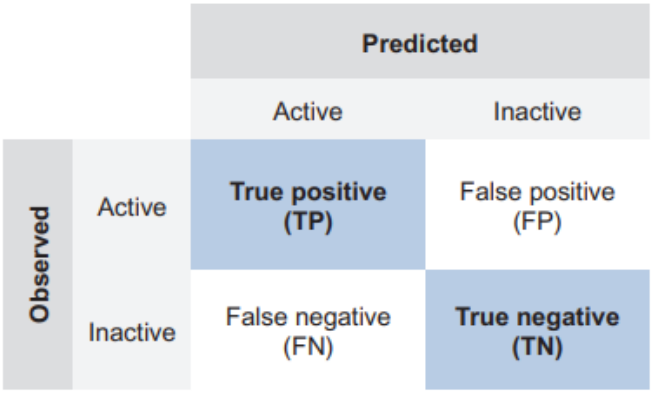
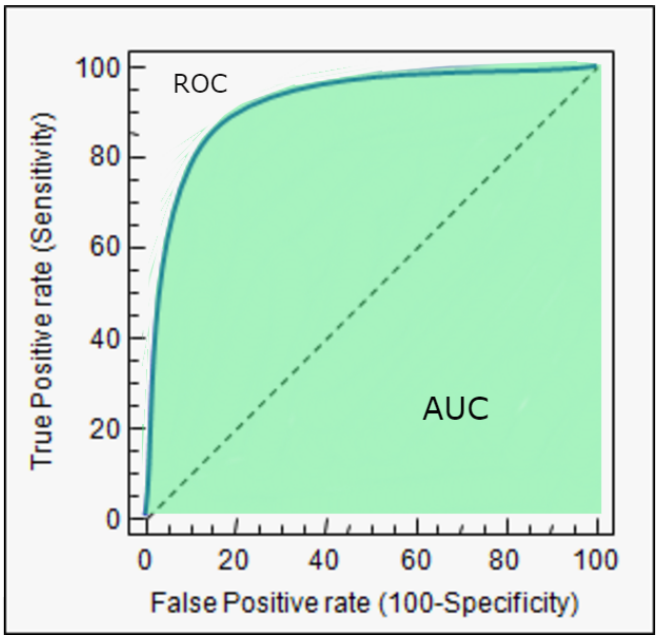


Figure 1. Confusion matrix describing the performance of a classification model (or 'classifier') on a set of test data for which the true values are known.

Wybór deskryptorów w przypadku **drzew regresyjnych** dokonywany jest przy pomocy metody najmniejszych kwadratów, czyli tak aby suma kwadratów różnic pomiędzy wartościami przewidywanymi przez model a zmierzonymi eksperymentalnie (tzw. rezyduałów) była jak najmniejsza.

Krzywa ROC to jeden ze sposobów wizualizacji jakości klasyfikacji, pokazujący zależności wskaźników TPR (True Positive Rate) oraz FPR (False Positive Rate). TPR przedstawia czułość, a FPR 1-specyficzność. W statystyce matematycznej krzywa ROC jest graficzną reprezentacją efektywności modelu predykcyjnego poprzez wykreślenie charakterystyki jakościowej klasyfikatorów binarnych powstałych z modelu przy zastosowaniu wielu różnych punktów odcięcia**.** Mówiąc inaczej – każdy punkt krzywej ROC odpowiada innej macierzy błędu uzyskanej przez modyfikowanie „cut-off point"(punkt w którym osiągnięta jest równowaga czułość-specyficzność) . Im więcej różnych punktów odcięcia zbadamy, tym więcej uzyskamy punktów na krzywej ROC.



Rys1. Przykładowa krzywa ROC z zaznaczonym AOC (Area under curve)

Na wykresie prawdopodobieństwo wzrasta od prawej do lewej strony. W prawym górnym rogu wykresu wynosi ono 0, więc model wszystkie przypadki uznaje za pozytywne – czułość wynosi więc 1 (100% rzeczywistych przypadków pozytywnych zostało za takie uznane) i 1-Swoistość też, ponieważ żaden przypadek nie został uznany za negatywny (1-0=1). Oprócz krzywej ROC na wykresie zaznaczona jest również przekątna – jest to teoretyczna linia klasyfikacji dokonywanych przez model losowy. Taki model równie często dokonuje błędnych i poprawnych klasyfikacji pozytywnych. Oczywiście nasz rzeczywisty model powinien mieć zdecydowanie lepsze parametry, więc obliczona dla niego krzywa ROC powinna znajdować się powyżej przekątnej jak na wykresie. Im wyżej tym lepiej. Bardzo popularnym podejściem jest wyliczanie pola pod wykresem krzywej ROC, oznaczanego jako AUC (ang. area under curve), i traktowanie go jako miarę dobroci i trafności danego modelu. W skrócie wartość ta oznacza jak dobrze model odróżnia i klasyfikuje klasy pozytywne i negatywne. Wartość wskaźnika AUC przyjmuje wartości z przedziału [0,1] i im większa wartość, tym lepszy model.

ZADANIE 1. Arkusz kalkulacyjny „**dane_nano1.xlsx**” zawiera dane eksperymentalne dot. karbonylacji białek dla dziesięciu nanomateriałów. Według tych danych nanomateriały te zostały podzielone na dwie klasy: aktywne (o silnej, średniej lub słabej zdolności do karbonylacji białek) oraz nieaktywne. Zbuduj model drzewa klasyfikacyjnego w celu przewidywania tego parametru dla pozostałych, niezbadanych eksperymentalnie związków (zbiór predykcyjny). Wykorzystaj dwa deskryptory: rozmiar („Size”) oraz powierzchnię („SSA”). Związki podziel na zbiór uczący i walidacyjny według algorytmu podziału 1:3 (test_size = 0.3, random_state=0). Oceń zdolności prognostyczne modelu na podstawie macierzy błędu oraz statystyk: czułości, specyficzności, precyzji, współczynnika F1, dokładności oraz błędu dokładności.

[Na dodatkowe 0,5 pkt na następnej wejściówce] Wykreśl krzywą ROC i wyznacz wartość AUC + krótka interpretacja.

ZADANIE 2. Arkusz kalkulacyjny „**dane_nano2.xlsx**” zawiera dane eksperymentalne dla 17 nanomateriałów – modelowaną wielkością jest NOAEC (ang. *no-observed-adverse-effect level*) tj. poziom niewywołujący zauważalnych szkodliwych skutków. Zbuduj model drzewa regresyjnego, aby przewidzieć aktywność pozostałych, niezmierzonych eksperymentalnie związków. Wykorzystaj 4 deskryptory: rozmiar, powierzchnię, obecność powłoki („SHELPRES”) i energię najniższego niezajętego orbitalu molekularnego rdzenia („LUMO_C”). Związki podziel na zbiór uczący i walidacyjny według algorytmu podziału 1:3 (test_size = 0.3, random_state=0). Pamiętaj o autoskalowaniu danych. W celu oceny prognostycznej modelu oblicz Q^2_{Ex} oraz $RMSE_{Ex}$.

Kod i sprawozdanie(lub kod z uwzględnionymi komentarzami) proszę wysłać na mail rafal.ziniewicz@phdstud.ug.edu.pl

Użyteczne biblioteki: pandas, numpy, matplotlib, sklearn-metrics, sklearn-tree, sklearn.tree-DecisionTreeClassifier, DecisionTreeRegression, sklearn.tree-plot_tree, sklearn.metrics-confusion_matrix, plot_confusion_matrix, classification_report, sklearn.model_selection-train_test_split.