Laboratorium Metody Sztucznej Inteligencji

Klasyfikacja danych

Klasyfikacja

Klasyfikacja jest to proces przewidywania przez klasyfikator (algorytm) przynależności punktów do konkretnej klasy.
Jednym z najprostszych przykładów klasyfikacji jest detekcja spamu w skrzynkach pocztowych. Jest to klasyfikacja binarna ze względu na podział wyłącznie na dwie klasy, tj. spam i nie spam. Klasyfikator wykorzystuje dane uczące do nauczenia się i zrozumienia jak zmienne wejściowe są powiązane z klasami. W tym przypadku znane wiadomości email będące i nie będące spamem są używane jako dane uczące. Jeśli proces uczenia klasyfikatora zostanie poprawnie przeprowadzony, klasyfikator będzie mógł służyć do rozpoznawania nowych, nieznanych wiadomości email.
Istnieje wiele algorytmów klasyfikacji, jednak nie jest możliwe stwierdzenie, który z nich jest najlepszy. Zależy to przede wszystkim od rodzaju danych, które będą przetwarzane.

Proces klasyfikacji

1. Przygotowanie danych (import, przetwarzanie, usunięcie pustych rekordów, etc.). W tym etapie dane dzielone są na dwie albo trzy części:
   a) dane uczące – wykorzystywane do uczenia klasyfikatora
   b) dane walidacyjne (w bardziej złożonych przypadkach) – służą do oceny jakości klasyfikatora podczas jego tworzenia
   c) dane testowe – wykorzystywane do ostatecznej oceny jakości klasyfikatora
2. Stworzenie klasyfikatora
3. Ocena jakości klasyfikatora na podstawie danych testowych

Dane uczące i testowe

W celu oceny jakości klasyfikacji dane powinny zostać podzielone na dwie grupy, dane uczące i dane testowe. Istnieje wiele metod podziału danych na zbiory uczące i testowe, np. metoda K-fold, Leave-one-out, Bootstrap, etc. Klasyfikator powinien być uczony danymi ze zbioru uczącego, a następnie sprawdzany na danych ze zbioru testowego. Aby określić jakość klasyfikatora należy sprawdzić jaki procent obserwacji został poprawnie przyporządkowany do danej klasy (w zbiorze testowym).

Holdout

Polega na losowym podziale zbioru danych na podzbiór uczący i testowy

Walidacja krzyżowa K-krotna (k-fold)

Walidacja krzyżowa K-fold polega na podziale zbioru danych na K podzbiorów. Każdy z podzbiorów jest kolejno wykorzystywany jako zbiór testowy, pozostałe K-1 podzbiory tworzą zbiór uczący. W celu określenia jakości klasyfikacji proces uczenia i testowania powtarzany jest K- krotnie, a błąd klasyfikacji obliczany jest jako suma błędnych klasyfikacji dla wszystkich K-wersji klasyfikatora podzielona przez liczbę obserwacji oryginalnego zbioru. Jest to równoznaczne z obliczeniem wartości średniej z błędów uzyskanych dla każdej wersji klasyfikatora osobno.

Leave-one-out

Szczególną odmianą walidacji krzyżowej K-fold jest walidacja krzyżowa typu leave-one-out. Polega ona na każdorazowym wykluczeniu pojedynczej obserwacji ze zbioru oryginalnego, która jest wykorzystywana jako obserwacja testowa, natomiast pozostałe obserwacje stanowią zbiór uczący. Proces uczenia i testowania powtarza się zatem tyle razy ile obserwacji zawiera oryginalny zbiór danych, a błąd obliczany jest jako suma obserwacji błędnie zaklasyfikowanych do odpowiedniej grupy podzielona przez liczebność oryginalnego zbioru danych.

Wskaźniki jakości klasyfikacji

W przypadku klasyfikacji binarnej najczęściej wykorzystywanymi wskaźnikami oceny jakości klasyfikacji są dokładność (*accuracy)*, czułość (*sensitivity)* i swoistość (*specificity)*. Do ich wyznaczenia korzystamy z tzw. tablicy pomyłek (*confusion matrix)*. Jej skonstruowanie wymaga zdefiniowania klasy pozytywnej i negatywnej. Na przykład w przypadku grupy pacjentów chorych i zdrowych do grupy pozytywnej zaliczymy chorych, a do grupy negatywnej – zdrowych.

		Wiedza ekspercka	
		Pozytywna	Negatywna
Wynik klasyfikacji	Pozytywna	TP	FP
	Negatywna	FN	TN

Dokładność to najbardziej intuicyjna miara jakości klasyfikacji. Określa jaka część wszystkich obserwacji została sklasyfikowana poprawnie

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Czułość to udział poprawnie sklasyfikowanych obserwacji należących do klasy pozytywnej

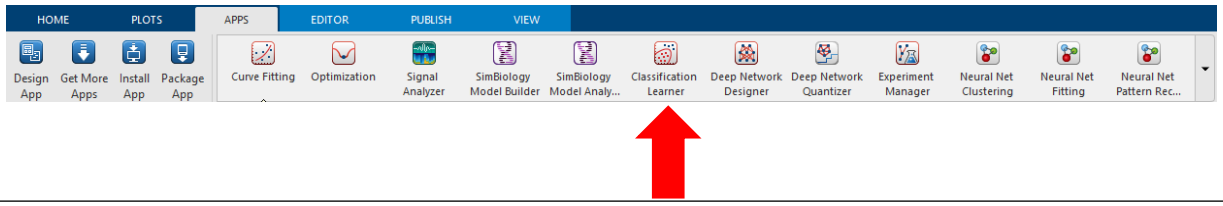
$$Sensitivity = \frac{TP}{TP + FN}$$

Swoistość to udział poprawnie sklasyfikowanych obserwacji należących do klasy negatywnej

$$Specificity = \frac{TN}{TN + FP}$$

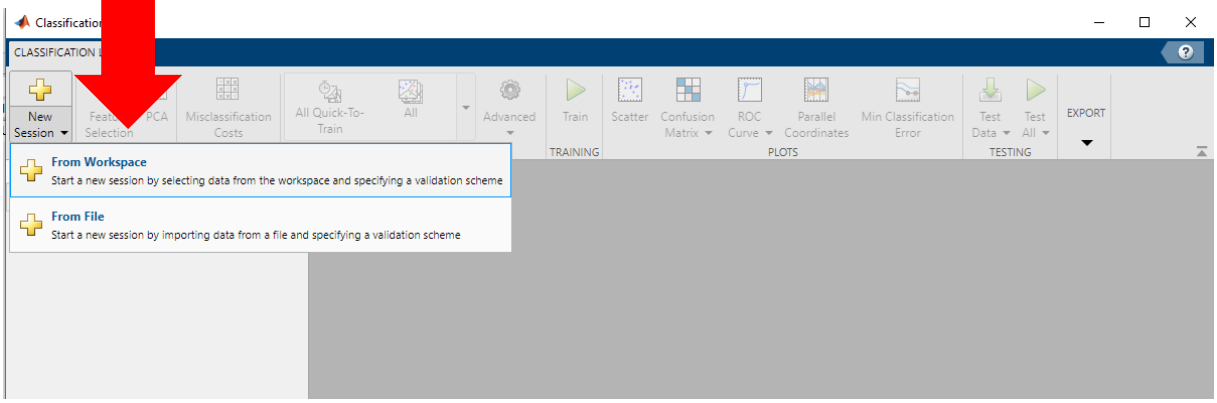
Wprowadzenie do MATLAB Classification Learner (zrzuty ekranu z wersji 2021a)

1. Uruchomienie aplikacji



Aplikację możemy uruchomić poleceniem `classificationLearner` lub wybierając ją w menu aplikacji MATLABa

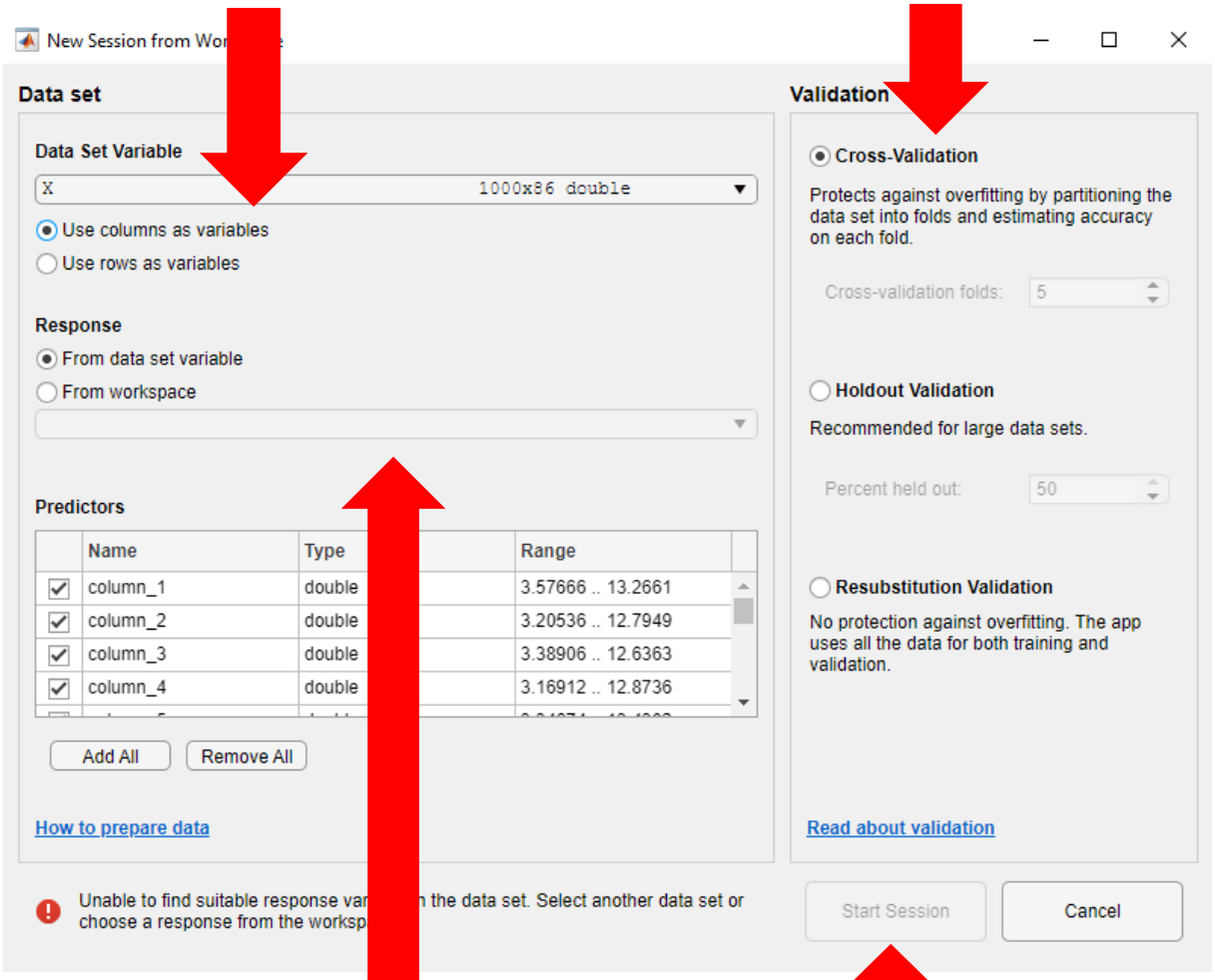
Możemy wybrać dane z pliku lub z obszaru roboczego. Ponieważ dane są zapisane w formacie struct musimy najpierw „rozdzielić” je na poszczególne części



3. Przygotowanie danych

Wybieramy czy zmienne znajdują się w wierszach czy kolumnach macierzy

Wybieramy metodę walidacji



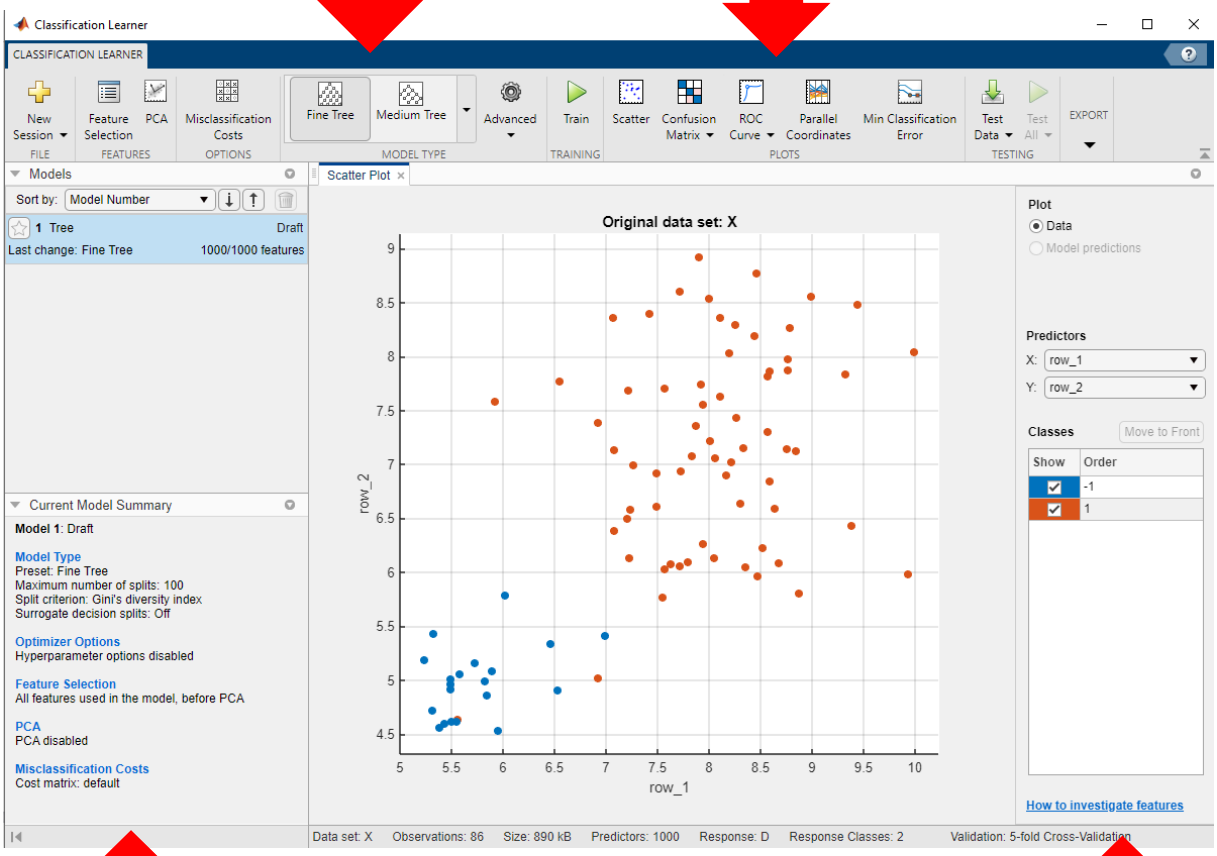
Wybieramy zmienną (w macierzy danych lub w przestrzeni roboczej) zawierającą etykiety klas

Po ustawieniu wszystkich opcji wybieramy “Start session”

4. Konstrukcja i ocena modelu

Tu możemy wybrać kolejny model do wytrenowania. Po wybraniu klikamy „Train”

Tu możemy wybrać wizualizacje pozwalające ocenić zbiór danych i jakość klasyfikacji



Tu znajduje się podsumowanie aktualnie wybranego modelu

Tu możemy edytować przypisanie klas jako pozytywnej i negatywnej oraz parametry wizualizacji

Zadania do wykonania:

1. Dla 10 krotnej walidacji krzyżowej oraz klasyfikatora Simple Tree:
   a) Podać nazwy genów i określić ich funkcje
   b) Zamieścić wykresy Scatterplot dla przydzielonych par genów, określić czy dane są separowalne liniowo
   c) Zamieścić tablicę pomyłek (confusion matrix). Zaznaczyć gdzie znajduje się TP, TN, FP, FN.
   d) Podać wartości czułości, swoistości i dokładności. Czy widać jakiś związek pomiędzy jakością klasyfikacji a rozkładem zmiennych zaobserwowanym na wykresach scatterplot?
   e) Napisać co określają TPR, FNR, PPV oraz FDR

f) Czym charakteryzują się błędy pierwszego i drugiego rodzaju? Kiedy ważniejsza jest czułość, a kiedy swoistość w kontekście zdrowych i chorych pacjentów?
g) Podać wartości AUC. Napisać co ilustruje krzywa ROC, jakie informacje można z niej odczytać. Czy AUC jest dobrym wyznacznikiem jakości klasyfikatora?
h) W wykresie ROC zmienić klasę pozytywną na przeciwną, dlaczego wyniki AUC są inne? Dlaczego krzywa ROC się zmienia?

2. **Porównanie metod podziału na dane uczące i testowe**. Podczas wczytywania danych w kroku 3 (punkt 7) dostępne są trzy opcje podziału danych: metody K-fold i Holdout oraz resubstytucja. W sprawozdaniu porównać i zamieścić wyniki dla 5 krotnej walidacji krzyżowej, 10 krotnej walidacji krzyżowej, HoldOut 25, 40 i 50% oraz resubstytucji. Wyniki zamieścić w tabeli, która powinna zawierać: dokładność klasyfikatora (modelu, błąd, czułość, swoistość oraz AUC. Wyniki skomentować. Jaka jest najlepsza metoda podziału dla wykorzystywanych danych, z czego to wynika? Jak zmiana metody podziału na dane uczące i testowe wpływa na jakość klasyfikacji? Co to jest overfitting i jak się przed nim zabezpieczyć?
3. **Porównanie klasyfikatorów**. Wybrać 10 krotną walidację krzyżową, a następnie porównać działanie klasyfikatorów: SVM linear, SVM quadratic, Fine KNN, Medium KNN, Simple Tree, Medium Tree (Wybrać klasyfikator, a następnie, Train). Wyniki zamieścić w tabeli, która powinna zawierać: dokładność klasyfikatora (modelu), błąd, czułość, swoistość, AUC
4. Na czym polegają metody SVM, KNN oraz Decision Trees, (zamieścić własne ilustracje). Napisać czym klasyfikatory różnią się od siebie. Który z testowanych klasyfikatorów daje najlepsze wyniki i dlaczego? Czy dla innych danych ten sam klasyfikator będzie najlepszy?

Sprawozdanie powinno zawierać wyniki symulacji i wnioski (w tym odpowiedzi na pytania znajdujące się w instrukcji). Czas na wykonanie sprawozdania to dwa tygodnie od daty laboratorium.