Laboratorium Metody Sztucznej Inteligencji– Klasyfikacja danych

1) 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.

2) 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

3) 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).

Walidacja krzyżowa K-krotna (ang.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-oneout. 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.

4) Wskaźniki jakości klasyfikacji

W przypadku gdy klasyfikujemy dane do dwóch klas najczęściej wykorzystywanymi wskaźnikami oceny jakości klasyfikacji są skuteczność, czułość i specyficzność.

Zakłada się, że do grupy Pozytywnej zaliczamy te obserwacje na których wykrywaniu najbardziej nam zależy. Na przykład w przypadku grupy pacjentów chorych i zdrowych do grupy Pozytywnej zaliczymy chorych, a do grupy Negatywnej zdrowych.

		Wiedza eksperta	
		Positive	**Negative**
Wynik klasyfikacji	**Positive**	TP	FP
	Negative	FN	TN

Wiedza eksperta Positive lub Negative

Wynik klasyfikacji Positive TP, FP, Negative FN, TN

Skuteczność jest najbardziej intuicyjnym wskaźnikiem nie wyróżniającym żadnej z grup (pozytywnej i negatywnej). Jego wartośd wyznaczana jest przez iloraz wszystkich poprawnie zaklasyfikowanych obserwacji zbioru testowego do liczby wszystkich obserwacji należących do tego zbioru:

$$\boldsymbol{Skutecznosc} = \frac{\mathbf{TP+TN}}{\mathbf{TP+FP+TN+FN}}$$

Czułością nazywa się iloraz liczby poprawnie zaklasyfikowanych obserwacji przez klasyfikator do grupy pozytywnej do liczby wszystkich obserwacji , które wedle wiedzy eksperta należą do tej grupy:

$$\boldsymbol{Czulosc} = \frac{\mathbf{TP}}{\boldsymbol{TP+FN}}$$

Specyficznością natomiast nazywa się iloraz liczby poprawnie zaklasyfikowanych obserwacji przez klasyfikator do grupy negatywnej do liczby wszystkich obserwacji, które wedle wiedzy eksperta należą do tej grupy:

$$\boldsymbol{Specyficznosc} = \frac{\boldsymbol{TN}}{\boldsymbol{TN+FP}}$$

Krzywa ROC (Receiver Operating Characteristic) jest narzędziem do oceny poprawności klasyfikatora, zapewnia ona łączny opis jego czułości i specyficzności. Jakość klasyfikacji za pomocą krzywej ROC można ocenić wyliczając takie wskaźniki jak:

- Pole pod krzywą (AUC) (Area Under ROC Curve)

Im większe AUC tym lepiej: AUC = 1 (klasyfikator idealny), AUC = 0.5 (klasyfikator losowy), AUC < 0.5 (nieprawidłowy klasyfikator (gorszy niż losowy)).

Dane Data_PTC_FTC.mat

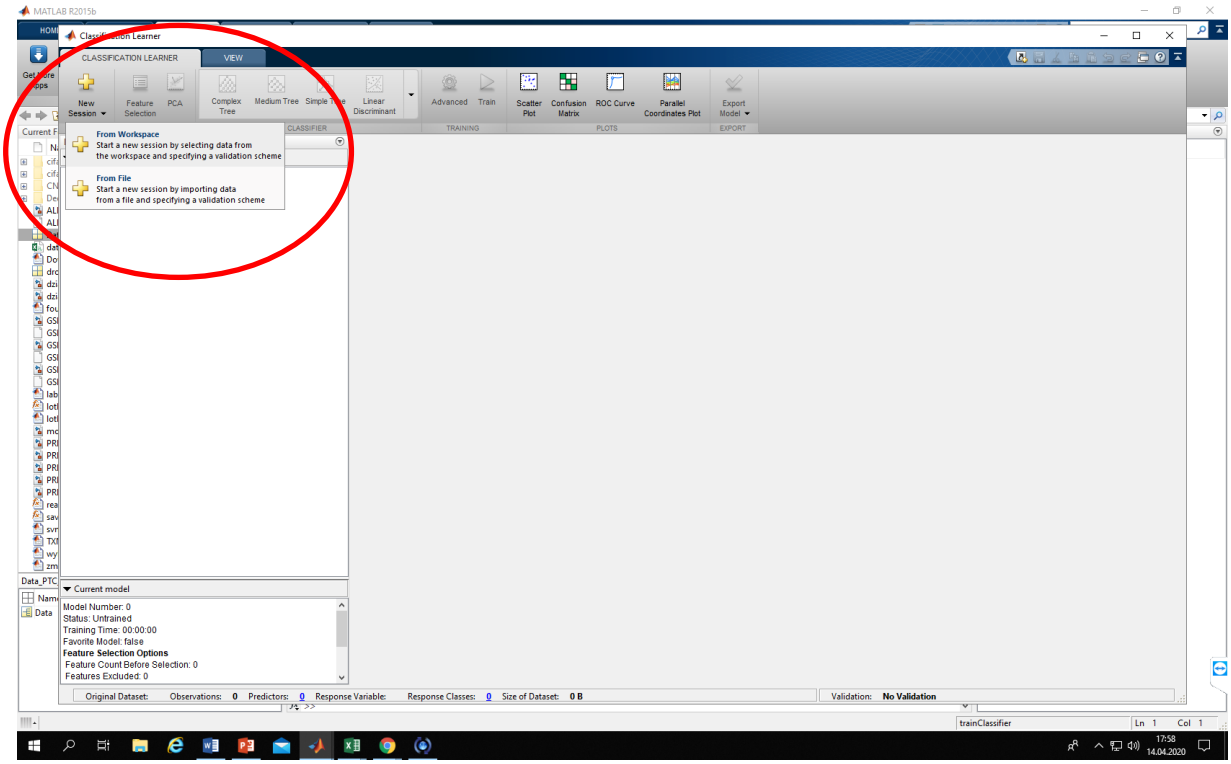
PTC (*ang. papillary thyroid carcinoma*) jest to nowotwór brodawkowaty tarczycy, pojawiający się najczęściej pomiędzy 20, a 40 rokiem życia. Zwykle nie daje większych objawów. Rokowanie określa się jako bardzo dobre, jednak w przypadku braku leczenia nowotwór prowadzi do śmierci. Dane, które będą wykorzystywane na laboratorium to dane dotyczące pacjentów chorych na nowotwór tarczycy i zdrowych przedstawiające wartości ekspresji 1000 genów. W pliku Dane.D znajdują się dane oznaczające przynależność do klas (wiedza ekspercka). Wartości określone jako 1 to pacjenci cierpiący na nowotwór tarczycy, -1 (na potrzeby ćwiczenia) pacjenci zdrowi. Plik Data.X zawiera dane dla 86 pacjentów z wartościami ekspresji 1000 genów. W pliku gene_names znajdują się nazwy prezentowanych genów.

Celem ćwiczenia jest porównanie metod klasyfikacji pozwalających na przyporządkowanie 86 pacjentów do konkretnej grupy chorych lub zdrowych na podstawie ekspresji wybranych genów.

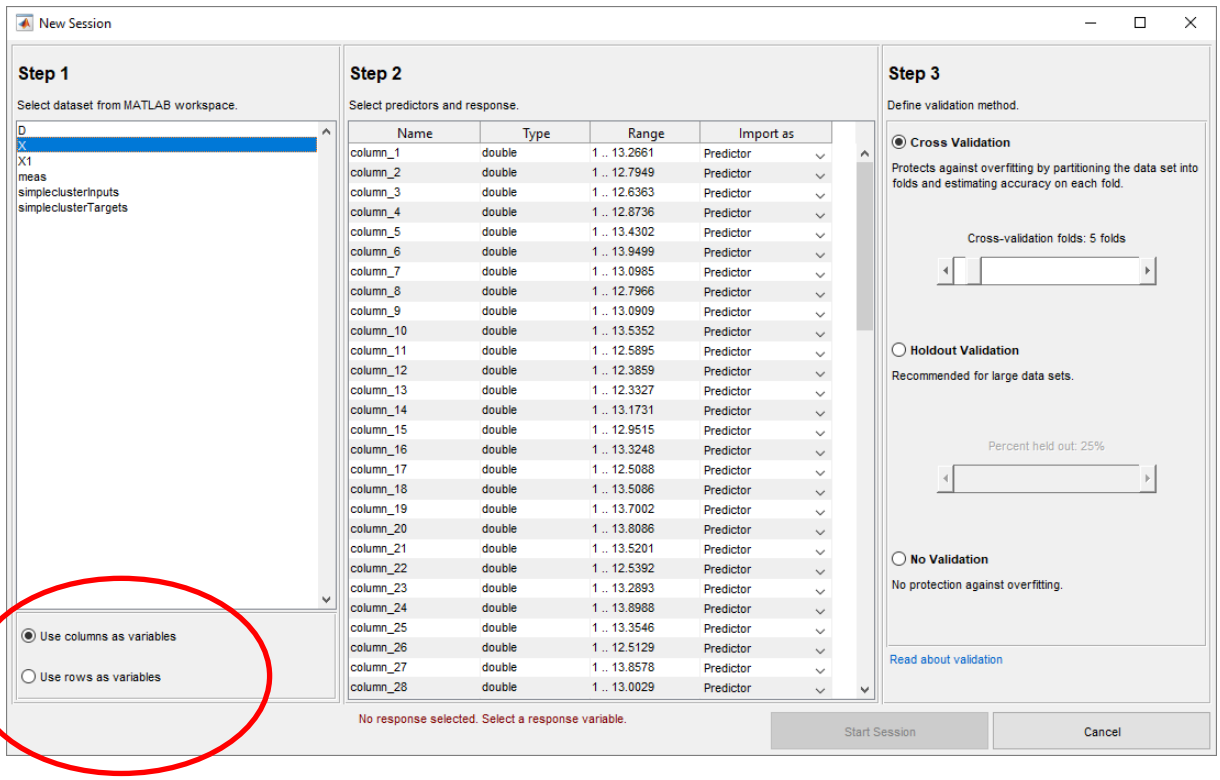
Wstęp do aplikacji Matlaba Classification Learner

1. Wczytać dane do przestrzeni roboczej MATLABA (plik Data_PTC_vs_FTC.mat), utworzyć zbiory X= Data.X, D=Data.D. W pliku gene_names znajdują się nazwy sprawdzanych genów.
2. Przygotować zbiór X do dalszej analizy. Do zbioru z wartościami ekspresji genów (X) dla 86 pacjentów, dodać do pierwszego wolnego wiersza wartości ze zbioru (D) oznaczające przynależność do klas (wiedza ekspercka). Wartości określone jako 1 to pacjenci cierpiący na nowotwór tarczycy, -1 (na potrzeby ćwiczenia) pacjenci zdrowi
3. Otworzyć aplikację ClassificationLearner poprzez wpisanie odpowiedniej komendy lub kliknąć

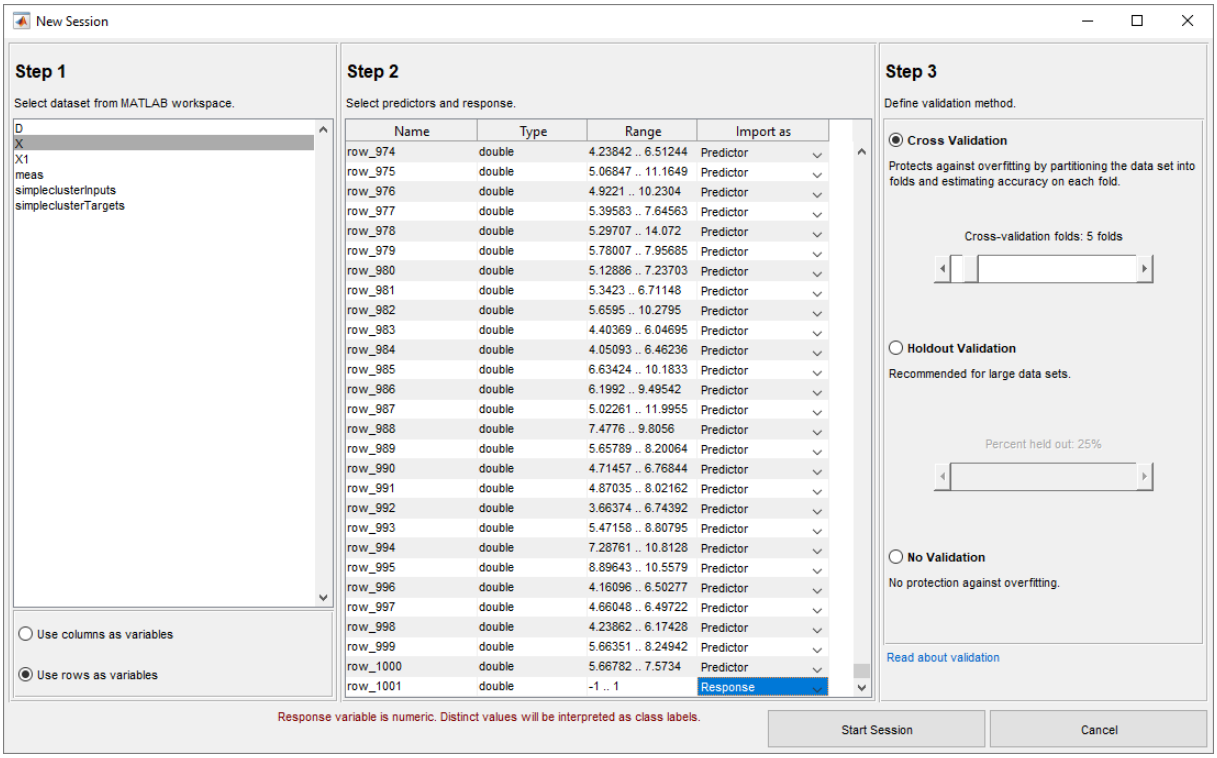
4. Po otwarciu aplikacji wybrać opcję New Session, a następnie From WorkSpace



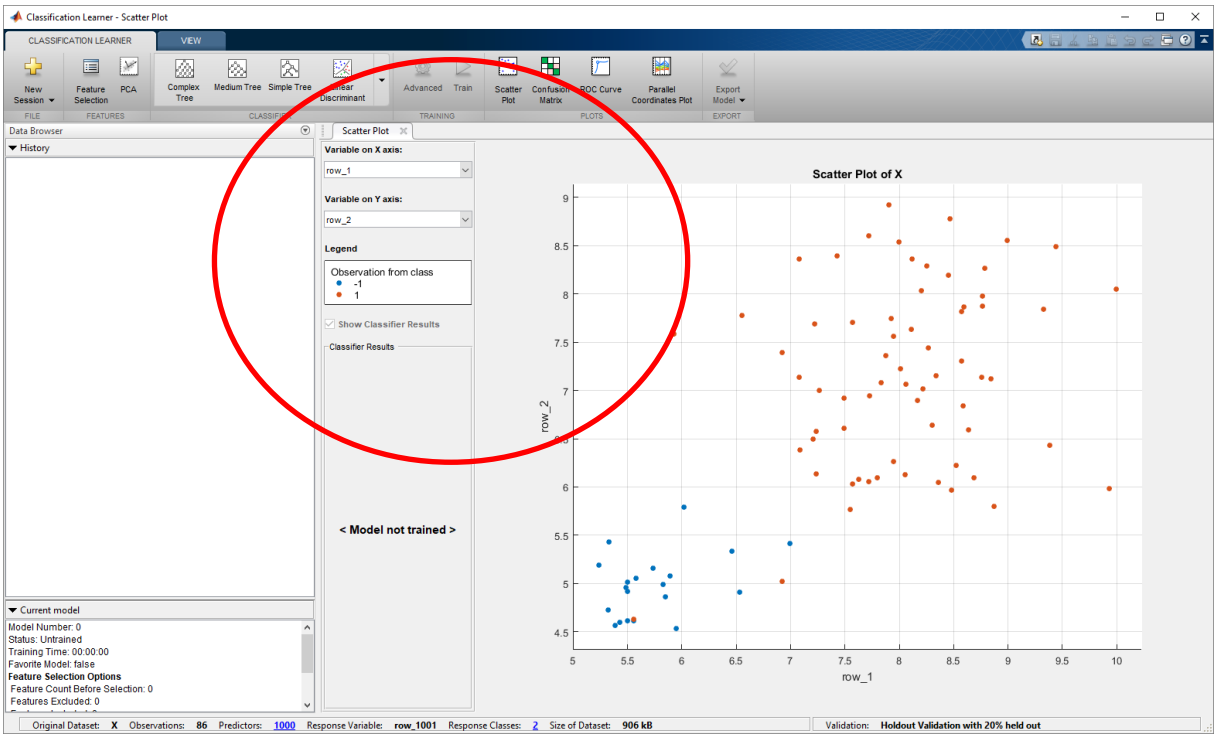
5. W kroku pierwszym Step 1, wybrać odpowiedni zbiór danych (X). Następnie wybrać odpowiednią opcję: kolumny jako zmienne lub wiersze jako zmienne. Należy pamiętać, że zbiór danych zawiera dane zawierające wartości ekspresji 1000 genów dla 86 pacjentów oraz wiersz z wartościami przynależności do danej klasy.



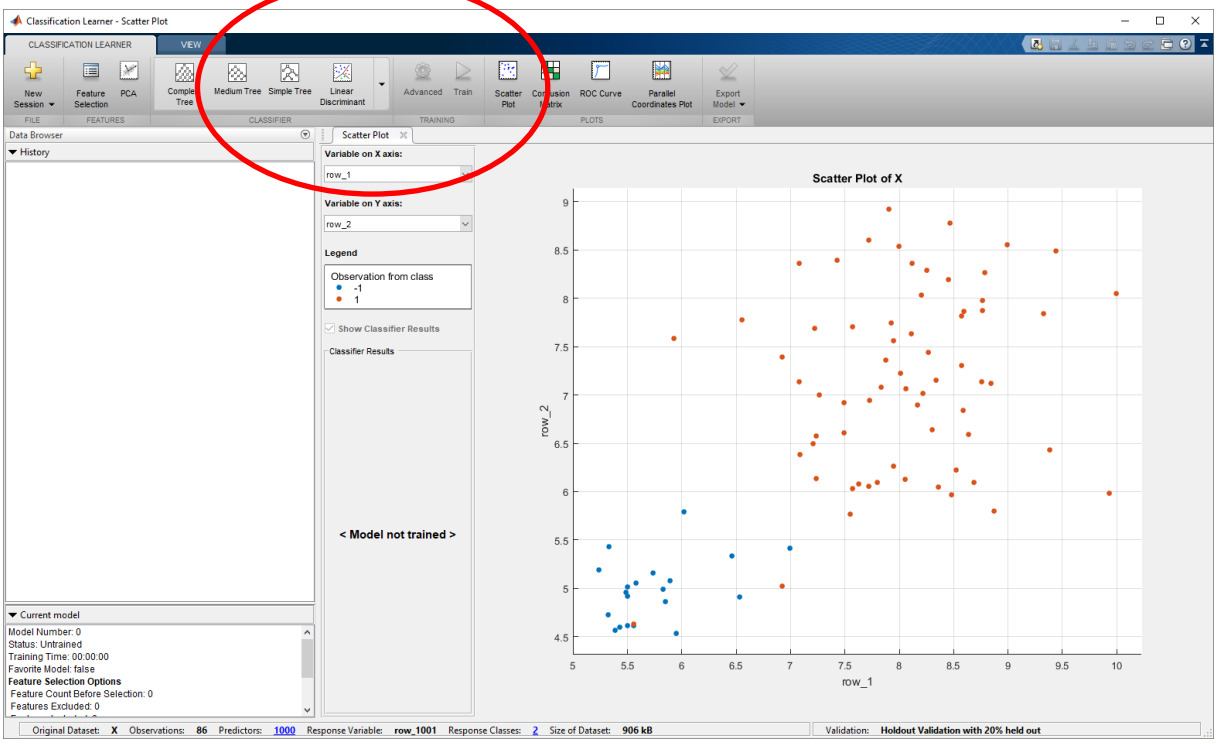
6. Po wybraniu odpowiedniej opcji, należy określić, zmienne jako predictor lub response. Predictor są to dane do zbudowania klasyfikatora, response określają dane z wiedzy eksperckiej. W aplikacji Classficiation Learner możliwe jest użycie tylko jednego zbioru, w którym określa się predictors oraz response, dlatego konieczne było dodanie do zbioru X wartości ze zbioru D zawierających dane z wiedzy eksperckiej (punkt 2), tak aby otrzymać jeden zbiór danych.



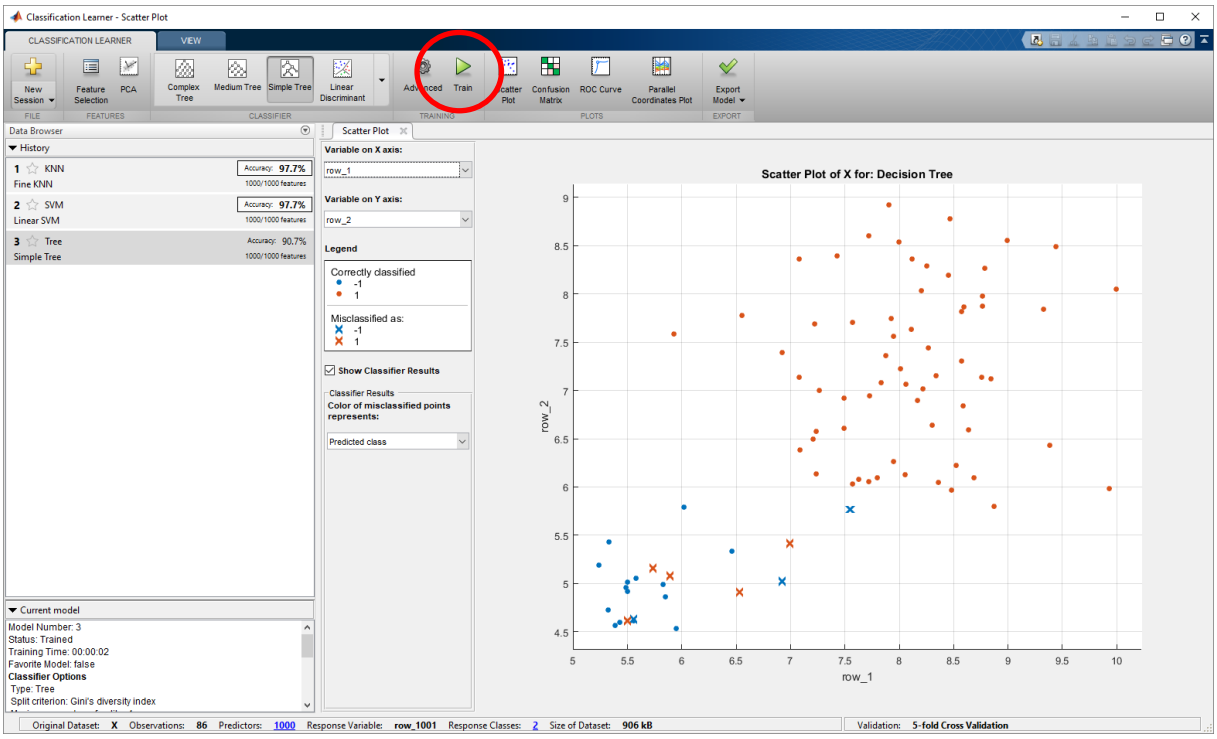
7. Po ustaleniu opcji predictor/response, w kroku 3 należy wybrać metodę walidacji klasyfikatora i kliknąć Start Session
8. Następnie należy wybrać geny do dalszej analizy. Variable on X axis, Variable on y axis. Każda z sekcji powinna zgłosić się do prowadzącego po własne numery genów!



9. W polu Classifiers należy wybrać odpowiedni klasyfikator



10. Po wybraniu odpowiedniego klasyfikatora kliknąć Train



11. Zapoznaj się z dostępnymi opcjami: Scatter Plot, Confusion Matrix, ROC Curve, Export Model
12. Wykonaj zadania, a następnie umieść wyniki w sprawozdaniu.

Zadania do wykonania:

1. Dla 10 krotnej walidacji krzyżowej oraz klasyfikatora Simple Tree:
   a) Podać nazwy genów i określić ich funkcje
   b) Zamieścić wykres Scatterplot dla wybranych genów, określić czy dane są separowalne liniowo
   c) Zamieścić tablicę pomyłek (confusion matrix). Napisać gdzie znajduje się TP, TN, FP, FN. Podać wartości czułości, specyficzności i skuteczności. Dodatkowo obliczyć błąd i dokładność, wyniki dla błędu i dokładności porównać z wynikami z aplikacji Classification Learner.
   d) Napisać co określają TPR, FNR, PPV oraz FDR
   e) Czym charakteryzują się błędy pierwszego i drugiego rodzaju. Kiedy ważniejsza jest czułość, a kiedy specyficzność w kontekście zdrowych i chorych pacjentów.
   f) Podać wartości AUC. Napisać co ilustruje krzywa ROC, jakie informacje można z niej odczytać. Czy AUC jest dobrym wyznacznikiem. Jaka wartość AUC jest wystarczająco dobra dla klasyfikatora. Czy porównywanie samych wartości AUC jest dobrym podejściem.
   g) W wykresie ROC zmienić klasę pozytywną na przeciwną, dlaczego wyniki AUC są inne? Dlaczego krzywa ROC się zmienia?

h) Wygenerować kod Matlaba

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 brak walidacji. W sprawozdaniu porównać i zamieścić wyniki dla 5 krotnej walidacji krzyżowej, 10 krotnej walidacji krzyżowej, HoldOut 25,50 i 75% oraz opcji bez walidacji. Wyniki zamieścić w tabeli, która powinna zawierać:
   a) Dokładność klasyfikatora (modelu)
   b) Błąd
   c) Czułość
   d) Skuteczność
   e) Specyficzność
   f) 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:
      a. Dokładność klasyfikatora (modelu)
      b. Błąd
      c. Czułość
      d. Skuteczność
      e. Specyficzność
      f. AUC

   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?

Materiały:

SVM

Laboratorium sieci neuronowe maszyny wektorów podpierających, Danuta Gaweł

KNN

http://home.agh.edu.pl/~horzyk/lectures/miw/KNN.pdf

https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Drzewa decyzyjne

http://zsi.tech.us.edu.pl/~nowak/odzw/PED_w3.pdf

Krzywa ROC

https://media.statsoft.pl/_old_dnn/downloads/krzywe_roc_czyli_ocena_jakosci.pdf

Czułość/Specyficzność

https://pl.qwe.wiki/wiki/Sensitivity_and_specificity

http://zsi.tech.us.edu.pl/~nowak/odzw/PED_w3.pdf