Preprocesing danych

Zadanie 1. Celem zadania jest przeanalizowanie technik wstępnego przetwarzania danych i ich wpływu na wyniki klasyfikacji. Pracować będziemy na zbiorze `Credit approval dataset`[1]. Jako klasyfikatorów użyj metod: regresji logistycznej, naiwnego Bayesa (NB), najbliższych sąsiadów (klasyfikatora k-NN), metody wektorów nośnych (Support Vector Machines, SVM) oraz lasów losowych (ang. *random forests*) z ich domyślnymi parametrami.

1. Zidentyfikuj w zbiorze `Credit approval dataset` cechy (tj. kolumny), dla których brakuje podanych wartości. Uzupełnij brakujące wartości. Do wyboru jest wiele możliwości, przykładowe z nich to:

   - usunięcie kolumn (cech) lub wierszy (przykładów), w których występują brakujące dane
   - uzupełnienie brakujących danych średnią, medianą lub modą wartości cechy
   - uzupełnienie brakujących danych najczęstszą wartością cech, wartością losową lub wartością zerową
   - uzupełnienie brakujących danych metodą nabliższych sąsiadów

   Zastanów się, jaki sposób imputacji brakujących danych będzie najbardziej odpowiedni. Jak będzie wyglądało uzupełnianie brakujących wartości dla danych numerycznych, a jak dla danych nominalnych?

2. Zwizualizuj rozkład wartości każdej cechy oraz zależności od innych cech w tzw. macierzy rozrzutu (ang. *scatter matrix*).

3. Kodowanie wartości nominalnych. Przeanalizuj, jak klasyfikatory radzą sobie z reprezentacją cech w postaci nominalnej oraz w kodowaniu *one hot encoding*. Które reprezentacje cech są dopuszczalne dla każdego z klasyfikatorów? Jeśli obie reprezentacje są dopuszczalne, która reprezentacja będzie bardziej efektywna?

4. Przeprowadź skalowanie cech. Najczęściej stosuje się jedno z dwóch podejść:

   - normalizacja (ang. *min-max scaling*):
     $x \leftarrow \frac{x - x_{\min}}{x_{\max} - x_{\min}}$

[1] `https://archive.ics.uci.edu/ml/datasets/credit+approval`

- standaryzacja (ang. *standarization*):
  $x \leftarrow \frac{x-\mu}{\sigma}$

  Zbadaj, jak skalowanie cech wpływa na dokładność klasyfikatorów k-NN oraz lasów losowych.

5. Porównaj wyniki otrzymane przez klasyfikatory. Wyniki przedstaw jako dokładność razem z przedziałem ufności.

   Ponieważ zbiór danych jest nieduży, pomiary wykonaj przy pomocy 5-krotnej walidacji krzyżowej.

6. Dla wybranego klasyfikatora przedstaw wykres precyzji w funkcji pełności (ang. *precision-recall curve*) oraz wykres charakterystyki roboczej odbiornika (ang. *receiver operating characteristic, ROC*).

Wykonując preprocesing danych, pamiętaj, aby wykonać go najpierw na aktualnym zbiorze treningowym, a następnie na zbiorze testowym, korzystając z informacji uzyskanych ze zbioru treningowego. Inaczej będziemy mieli tzw. wyciek informacji ze zbioru testowego do zbioru treningowego.

O ile to możliwe, przetwarzanie danych zorganizuj w tzw. potok (ang. *pipeline*)

Literatura

[1] Sebastian Raschka, Model Evaluation, Model Selection and Algorithm Selection in Machine Learning, 2018.