

Selekcja cech

Zadanie 1. Celem laboratorium jest zapoznanie się z algorytmami selekcji cech. Wykorzystamy zbiór `leukemia`¹. Zbiór ten służy do klasyfikacji binarnej: przewidywania, czy dana osoba cierpi na białaczkę ostrą. Zbiór zawiera dane dotyczące 72 osób (instancji). Każda instancja opisywana jest przez 7070 cech.

1. Wczytaj zbiór danych. Przydatna będzie funkcja `scipy.io.loadmat`.
2. Usuń cechy charakteryzujące się niską wariancją, tj. takie, których wariancja jest niższa niż pewien ustalony próg. Porównaj skuteczność predykcji na oryginalnym zbiorze i po usunięciu cech o niskiej wariancji.
3. Z powstałego zbioru cech wybierz m najlepszych cech, stosując algorytm rekursywnej eliminacji cech (ang. Recursive Feature Elimination, RFE). Przydatna będzie funkcja `sklearn.feature_selection.RFE` lub `sklearn.feature_selection.RFECV`.

Liczba pozostawionych cech nie może być zbyt duża w stosunku do liczby instancji, gdyż inaczej klasyfikator nie będzie zbyt dobrze generalizował. Można przyjąć np., że $m < n/3$, gdzie n to liczba instancji w zbiorze treningowym.

Jako klasyfikatorów dostarczających informacji o ważności cech użyj regresji logistycznej oraz lasów losowych.

4. Zbadaj wpływ dwóch metryk: dokładności oraz metryki AUC na dokładność klasyfikacji przy użyciu regresji logistycznej oraz lasów losowych. Zastosuj cztero- lub sześciokrotną walidację krzyżową.
5. Porównaj skuteczność (tzn. dokładność) zastosowanego podejścia z wbudowanymi metodami selekcji cech:
 - (a) metodą regularyzacji L1 dla regresji logistycznej,
 - (b) metodą opartą na ważności cech (ang. feature importances) dla lasów losowych. W tym podejściu należy użyć klasy `SelectFromModel`.

¹<https://jundongl.github.io/scikit-feature/files/datasets/leukemia.mat>