Praktyczne zastosowanie metod SI – Laboratorium nr 4

Znają już Państwo jeden sposób ograniczenia wymiarowości zbioru danych – PCA. W pewnym sensie "wadą" tego rozwiązania jest to, że powstają nowe zmienne. A co, jeżeli np. mamy zbiór danych, któe nadal zbieramy, i chcemy ustalić które atrybuty są najbardziej przydatne? Do rozwiązania tego problem służy **selekcja atrybutów** (co ciekawe, przy okazji można w ten sposób podnieść jakość klasyfikacji).
Do selekcji atrybutów wykorzystamy pakiet sklearn.feature_selection.

ZADANIE NR 1:

1. Proszę wczytać zbiór leukemia z pliki arff (z użyciem funkcji **loadarff** z pakietu **scipy)**. Zbiór ten zawiera zaledwie 72 próbki, opisane przy pomocy aż 7129 atrybutów (oznaczają one poziomy ekspresji genów u pacjentów z białaczką).
   Adres pliku: http://wikizmsi.zut.edu.pl/uploads/3/32/Leukemia2.zip

   Przy wczytywaniu przydadzą się polecenia:

```
data, meta = arff.loadarff('Leukemia2.arff')
data = np.array(data.tolist())
```

   Następnie proszę podzielić dane na wartości atrybutów oraz klasę (klasa to ostatnia kolumna macierzy data).

2. Proszę wykonać selekcję atrybutów przy pomocy metody VarianceThreshold, która odrzuca atrybuty o wariancji poniżej argumentu podanego w konstruktorze. Na zbiorze Leukemia wyraźne efekty mogą dać dopiero wysokie wartości progu, takie jak 1e5.
3. Następnie proszę przetestować naiwny klasyfikator Bayesa przy użyciu funkcji cross_val_score z k=7 (oraz obliczyć średnią z uzyskanych 7 wyników) w dwóch wariantach:
   a. X oryginalny
   b. X_new – X po selekcji atrybutów
4. Proszę poeksperymentować z wpływem wartości parametru threshold dla metody VarianceThreshold – zmieniać próg w zakresie od 1e2 do 1e7 i narysować efekty na wykresie oraz zaznaczyć przy pomocy poziomej linii jakość klasyfikacji uzyskaną bez selekcji.

ZADANIE NR 2:

Proszę posłużyć się klasą SelectKBest z pakietu sklearn.feature_selection (ważny argument konstruktora to k – liczba atrybutów do wybrania; przydatna metoda: fit_transform, która od razu zwróci zbiór danych po selekcji). Proszę wykorzystać tą klasę w połączeniu z następującymi metodami selekcji atrybutów:

1. mutual_info_classif – mierzy zależność między atrybutami.
2. chi2 – test chi kwadrat, sprawdza powiązanie atrybutów z klasą.

Następnie proszę przygotować eksperyment: zmiana liczby atrybutów co 50 – 50, 100, 200,.. ,1000 oraz mierzenie jakości klasyfikacji z wykorzystaniem cross_val_score (k=7) z metryką accuracy i naiwnym klasyfikatorze Bayesa. Wyniki proszę przedstawić na wykresie: na osi OX liczba atrybutów, na osi OY średnie wyniki accuracy dla klasyfikatora.

ZADANIE NR 3:

Proszę przetestować działanie powyższych metod selekcji na zbiorze digits (wybór 5 dowolnie wybranych liczb atrybutów) i wyrysować uzyskane efekty (jako obrazki 8x8 pikseli, gdzie zamalowane na czarno są piksele odpowiadające wybranym atrybutom – przyda się polecenie reshape()).