5 Selekcja zmiennych za pomocą przyrostu informacji

Cel ćwiczenia

Celem ćwiczenia jest analiza jednej z podstawowych metod selekcji zmiennych dyskretnych opartej na przyroście informacji. W zadaniu wykorzystywane będą pakiety **numpy** i **scipy.sparse**. Do wczytywania danych, można posłużyć się pakietami **pandas** (metoda **read_csv**) oraz **sklearn**.

Zadania

1. Używając standardowego słownika języka Python napisać funkcję `[xi, ni]=freq(x, prob=True)`, która dla zadanej kolumny danych x dyskretnych zwróci: unikalne wartości xi, ich estymowane prawdopodobieństwa pi lub częstości ni.

2. Napisać funkcję `[xi, yi, ni] = freq2(x,y, prob=True)`, która dla zadanych kolumn danych x i y zwróci: unikalne wartości atrybutów xi, yi oraz łączny rozkład częstości lub liczności ni (w zależności od prarametru `prob`).

3. Wykorzystując powyższe funkcje, napisać funkcje, które wyliczą: entropię `h=entropy(x)` oraz przyrost informacji `i=infogain(x,y)` zgodnie ze wzorami:

$$I_H(X) = -\sum_{i=1}^{n} \Pr\{X = x_i\} \log_2(\Pr\{X = x_i\})$$

   Informacja wzajemna i przyrost informacji:

$$I_(X,Y) = H(X) + H(Y) - H(XY)$$
$$I_(Y,X) = H(Y) - H(Y|X)$$

   gdzie:

$$H(Y|X) = \sum_{i=1}^{k} \Pr\{X = x_i\} H(Y|X = x_i)$$

4. Wczytać dane testowe `zoo.csv` oraz dokonać selekcji/stopniowania atrybutów z wykorzystaniem kryterium przyrostu informacji.

5. Sprawdzić czy funkcje `freq`, `freq2` działają dla atrybutów rzadkich (pakiet **scipy.sparse**). Przerobić funkcje tak aby działały dla atrybutów rzadkich.

6. Wykonać eksperyment podsumowujący:

   (a) Wczytać bazę `Reuters Corpus Volume I`, informacje na temat zbioru danych dostępne są pod adresem: `http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm`,

   (b) Dane `data` zawierają liczności wystąpień słów w streszczeniach artykułów prasowych z korpusu agencji **Reuters**, tagi opisujące atrykuły zawarte są w tabeli `targets`. Przed przystąpieniem analizy należy zbinaryzować dane tak aby określały one nie liczność wystąpnienia, a tylko sam fakt wystąpienia.

   (c) Wybrać jeden atrybut decyzyjny z grupy tagów `target` (na przykład mkcmdGSPO oznacza artykuły dotyczące sportu).

   (d) Za pomocą funkcji `infogain` wyznaczyć dla każdego słowa przyrost inforamcji

   (e) Wypisać 50 zmiennych (słów) dostarczających najwięcej informacji nt. wybranej zmiennej decyzyjnej.

Przykładowy kod wczytują baze **Reuters** korzystając z API pakietu `sklearn`, jako zmienną decyzujną wybrano tag `GSPO` (atrybut nr 87)) oznaczajacy sport:

```
from sklearn.datasets import fetch_rcv1
rcv1 = fetch_rcv1()
X = rcv1["data"]         # lub: rcv1.data
Y = rcv1.target[:,87]  # tag GSPO - Sport
```