Języki i Biblioteki Analizy Danych

Laboratorium 9.: Klasyfikacja

mgr inż. Zbigniew Kaleta

Klasyfikacja statystyczna – rodzaj algorytmu statystycznego, który przydziela obserwacje statystyczne do klas, bazując na atrybutach (cechach) tych obserwacji.
(https://pl.wikipedia.org/wiki/Klasyfikacja_statystyczna)

Obserwacja statystyczna – pojedyncza realizacja zmiennej losowej. W praktyce zwykle jest to wielowymiarowa zmienna losowa, wówczas obserwacją statystyczną jest wektor realizacji składowych zmiennych losowych dotyczących tego samego badanego elementu populacji (jednostki statystycznej).
(https://pl.wikipedia.org/wiki/Obserwacja_statystyczna)

Przykłady obserwacji:

- e-mail
- zdjęcie
- wyniki badań krwi pacjenta

W uczeniu maszynowym najchętniej reprezentujemy obserwacje jako wektory liczb zmiennoprzecinkowych, ewentualnie całkowitych.

Klasyfikacja binarna polega na przypisaniu obserwacji do jednej z dwóch klas (tak/nie, chory/zdrowy, spam/nie-spam...).

Klasyfikacja wieloklasowa polega na przypisaniu obserwacji do jednej z wielu klas (czy zdjęcie przedstawia psa, kota, chomika czy świnkę morską? czy artykuł należy do kategorii ekonomia, polityka, sport czy inne?).

Klasyfikacja wieloetykietowa polega na przypisaniu obserwacji do żadnej, jednej lub więcej z wielu klas (j.w. ale na zdjęciu może być i pies, i kot).

Ogólna procedura w uczeniu maszynowym:

- zdefiniuj dobrze swój problem
- przygotuj odpowiednio duży i dobry zbiór danych
- podziel (losowo lub metodycznie) zbiór na dane treningowe i testowe, względnie: treningowe, walidacyjne i testowe
- wybierz algorytm i parametry
- wyucz model
- przetestuj model

Ocena klasyfikacji binarnej:

$$Accuracy = \frac{\text{przypadki poprawnie zaklasyfikowane}}{\text{wszystkie przypadki}} = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}$$

- klasyfikator dał odpowiedź 1, poprawna odpowiedź 1 - True Positive
- klasyfikator dał odpowiedź 0, poprawna odpowiedź 0 - True Negative
- klasyfikator dał odpowiedź 1, poprawna odpowiedź 0 - False Positive (błąd typu I)
- klasyfikator dał odpowiedź 0, poprawna odpowiedź 1 - False Negative (błąd typu II)

$$Precision = \frac{|TP|}{|TP| + |FP|}$$

$$Recall = \frac{|TP|}{|TP| + |FN|}$$

F1 - średnia harmoniczna precyzji i recall

$$\frac{2}{f_1} = \frac{1}{prec} + \frac{1}{rec} \Rightarrow f_1 = \frac{2 \cdot prec \cdot rec}{prec + rec}$$

Uogólniając (na średnią ważoną):

$$F_ß = \frac{(1 + ß^2) \cdot prec \cdot rec}{ß^2 \cdot prec + rec}$$

(im **większa** ß tym **mniejsze** znaczenie ma precyzja)

Ocena klasyfikacji wieloklasowej:

Macro-average precision - obliczamy precyzję dla poszczególnych klas i liczymy średnią arytmetyczną

Ocena klasyfikacji wieloetykietowej:

Hamming loss - liczba odpowiedzi błędnych (każda etykieta liczona jako osobna odpowiedź) podzielona przez liczbę wszystkich odpowiedzi. Im mniej tym lepiej.

Algorytmy klasyfikacji:

- drzewa decyzyjne
- perceptron
- kNN
- Naiwny Klasyfikator Bayesa
- SVM
- sieci neuronowe
- ...

Lektura dodatkowa:

- https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f
- https://www.edureka.co/blog/classification-in-machine-learning/
- https://machinelearningmastery.com/types-of-classification-in-machine-learning/
- https://pythongeeks.org/classification-in-machine-learning/