Indukcyjne Metody Analizy Danych

laboratorium

Ćwiczenie 1. Klasyfikator oparty na twierdzeniu Bayesa przy naiwnym założeniu o wzajemnej niezależności atrybutów

opracował: P.B. Myszkowski * data aktualizacji: 25.02.2018

Cel ćwiczenia
Zapoznanie się z probabilistycznym klasyfikatorem opartym na twierdzeniu Bayesa przy samodzielnej implementacji w python

Realizacja ćwiczenia
•Zapoznanie się z metodą uczenia probabilistycznego klasyfikatora bazującego na twierdzeniu Bayesa i założeniu o wzajemnej niezależności atrybutów (*Naive Bayes classifier*).
•Własnoręczna implementacja, przy wykorzystaniu bibliotek języka python
•Wybór trzech zbiorów danych do przetestowania. Należy uwzględnić zbiór z wartościami ciągłymi.
•Przebadanie działania zaimplementowanego klasyfikatora na wyżej wymienionych zbiorach
•Porównanie zachowania algorytmu przy danych ciągłych na wybranym zbiorze, przy różnych podejściach (różne metody dyskretyzacji lub założenie o normalnym rozkładzie wartości)
•Sporządzenie sprawozdania (w LaTeX) z przeprowadzonego ćwiczenia

Informacje pomocnicze
Rozwiązywanie jest zadanie klasyfikacji przy pomocy klasyfikatora probabilistycznego, wychodzącym z założenia o wzajemnej niezależności atrybutów od siebie i opartym na twierdzenie Bayesa (ang. *Naive Bayes classifier*). Zadanie polega na implementacji algorytmu, który na podstawie danych uczących buduje bayesowski klasyfikator. Należy pamiętać o poradzeniu sobie z danymi ciągłymi, poprzez dyskretyzację lub założenie, że dane mają rozkład normalny, skąd można policzyć prawdopodobieństwo wystąpienia danej wartości. Jednym z celów zadania jest porównanie jakości klasyfikatora przy różnych podejściach do danych ciągłych.

Ocena jakości klasyfikatora – słowa kluczowe: Confusion matrix, Accuracy, Precision, Recall i Fscore. Warto pamiętać, że każda z miar ma inne zastosowanie/cechy ale też wady/zalety. Warto to przedyskutować we wnioskach.

Należy również zadbać o wygładzenie danych, aby uniknąć zerowych prawdopodobieństw. W ocenie prawdopodobieństwa tego, ze dany wektor danych należy do danej klasy, wymnażane są prawdopodobieństwa i pojedyncza wartość „zerowa” usunęłaby informacje pochodzące z innych atrybutów. W praktyce, jeśli dana kombinacja wartość atrybutu/klasa nie wystąpiła w danych uczących, to i tak nie możemy z góry zakładać zerowego prawdopodobieństwa. Najłatwiejszy sposób na poradzenie sobie z tym to zwiększenie o jeden częstości występowania wszystkich dyskretnych wartości atrybutu.

Do oceny skuteczności algorytmu zaleca się użycie kroswalidacji (walidacji krzyżowej). Może być zwykła, np. 10-fold. Należy zbadać wpływ rozmiaru kroswalidacji na skuteczność modelu klasyfikacji. Dodatkowo, można uwzględnić kroswalidacje stratyfikowaną.

Zasady oceny zadania

2pkt	Implementacja klasyfikatora Bayesa
1pkt	Implementacja i testowanie trzech różnych metod dyskretyzacji
1pkt	Krótki opis działania algorytmu Bayesa
2pkt	Zbadanie działania klasyfikatora na 3 wybranych zbiorach
1pkt	Porównanie działania algorytmu przy różnych podziałach danych – tabelki, wnioski
1pkt	Porównanie działania algorytmu – graficzne (wykresy) przestawienie uzyskanych wyników
2pkt	Porównanie działania algorytmu na wybranym zbiorze z wartościami ciągłymi, uwzględniając różne metody radzenia sobie z tymi danymi (różne sposoby dyskretyzacji i liczenie prawdopodobieństwa z założenia o rozkładzie normalnym wartości ciągłych atrybutów).

Uwaga! Przy testach proszę pamiętać o kroswalidacji

Literatura

1. Cichosz P. ”Systemy uczące się”, WNT Warszawa
2. Eksploracja danych (seria wykładów)
   http://wazniak.mimuw.edu.pl/index.php?title=Eksploracja_danych
   (o naiwnym klasyfikatorze Bayesa mowa jest na 9 wykładzie)
3. Zasoby Internetu: naive bayes classifier
4. Sugerowane zbiory do badań:
   https://archive.ics.uci.edu/ml/datasets.html?format=&task=cla&att=&area=&numAtt=&numIns=&type=&sort=dateUp&view=table
   (zbiór IRIS do testów/analizy)
   https://en.wikipedia.org/wiki/Iris_flower_data_set#/media/File:Iris_dataset_scatterplot.svg

Przydatne linki (python)

bayes sci-kit
http://scikit-learn.org/stable/modules/naive_bayes.html

classification metrics
http://scikit-learn.org/stable/modules/model_evaluation.html

crossvalidation
http://scikit-learn.org/stable/modules/cross_validation.html

stratified crossvalidation
http://scikit-learn.org/sdtable/modules/generated/sklearn.model_selection.StratifiedKFold.html

discretisation
https://docs.scipy.org/doc/numpy/reference/generated/numpy.digitize.html
https://stackoverflow.com/questions/6163334/binning-data-in-python-with-scipy-numpy

visualisation
https://blog.modeanalytics.com/python-data-visualization-libraries/ ←przegląd bibliotek do wizualizacji/wykresów