

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

opracował: P.B. Myszkowski

Cel ćwiczenia

Zapoznanie się z probabilistycznym klasyfikatorem opartym na twierdzeniu Bayesa przy samodzielnej implementacji (np. Java/C++/C#).

Realizacja ćwiczenia

- Zapoznanie się z metodą uczenia probabilistycznego klasyfikatora bazującego na twierdzeniu Bayesa i założeniu o wzajemnej niezależności atrybutów (*Naive Bayes classifier*).
- Własnoręczna implementacja w wybranym obiektowym języku programowania.
- 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 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 twierdzeniu Bayesa (ang. *Naive Bayes classifier*). Zadanie polega na implementacji algorytmu, który na podstawie danych uczących buduje bayesowski klasyfikator. Należy pamiętać o poradeniu 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, że 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 wartości atrybutu/klasa nie wystąpiła w danych uczących, to i tak nie możemy z góry zakładać zerowego prawdopodobieństwa. Najłatwiejszy sposób na poradenie 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.

2pkt	Implementacja klasyfikatora Bayesa
1pkt	Implementacja i testowanie 2 różnych metod <u>dyskretyzacji</u>
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) przedstawienie 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 krosvalidacji

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 mówi rozdział 9
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>