Sprawozdanie IV

Reprezentacja niepewności

# Prawdopodobieństwo warunkowe – proste przeliczenie

Szacuje się, że 0,05% populacji USA ma HIV. Istnieje test na HIV: (a) jeśli badany ma HIV, test ma 98% szansy na pozytywny wynik; (b) jeżeli osoba nie ma HIV, test ma 3% szansy na pozytywny wynik. Tomek ma wynik pozytywny. Jakie jest prawdopodobieństwo, że ma HIV?

# Klasyfikator Naive Bayes - ręcznie

Rozważmy taki prosty zbiór treningowy, w którym każdy przykład ma cztery binarne atrybuty i przydzieloną jedną z dwóch klas (+/-):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Przykład** | **Atrybut\_1** | **Atrybut\_2** | **Atrybut\_3** | **Atrybut\_4** | **Klasa** |
| x1 | 1 | 1 | 1 | 1 | + |
| x2 | 1 | 1 | 0 | 1 | + |
| x3 | 0 | 1 | 1 | 0 | + |
| x4 | 1 | 0 | 0 | 1 | + |
| x5 | 1 | 0 | 0 | 0 | + |
| x6 | 1 | 0 | 1 | 0 | - |
| x7 | 0 | 1 | 0 | 0 | - |
| x8 | 0 | 0 | 1 | 0 | - |

W jaki sposób naiwny klasyfikator Bayesowski, wyuczony na powyższym zbiorze treningowym, zaklasyfikuje poniższy przykład? Policz ręcznie :)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Przykład** | **Atrybut\_1** | **Atrybut\_2** | **Atrybut\_3** | **Atrybut\_4** | **Klasa** |
| x9 | 1 | 1 | 0 | 0 | ? |

# Zastosowanie klasyfikatora Naive Bayes w Wece

1. Włącz Wekę i wczytaj plik weather.numeric.arff ze znanego Ci już zbioru danych: [data.tar.gz](https://ai.ia.agh.edu.pl/_media/pl:dydaktyka:psi:data.tar.gz)
2. Przejrzyj ten zbiór danych i przypomnij sobie czego on dotyczy.
3. Przejdź na zakładkę **Classify**. Weka udostępnia dwie wersje Naive Bayes. Zapoznaj się z ich opisami:
   * NaiveBayes
   * NaiveBayesUpdateable
4. Przetestuj obydwie wersje algorytmu? Jakie są różnice?
5. Następnie przetestuj te same algorytmy na pliku weather.nominal.arff. Jakie różnice występują teraz? Co jest przyczyną występowania różnic - odpowiedz korzystając z opisów algorytmów.

# Multinomial Naive Bayes w Wece

1. Wczytaj w Wece plik ReutersGrain-train.arff i zapoznaj się z jego budową. Jakie są atrybuty? Jakie przyjmują wartości?
2. Przejdź do zakładki **Classify**, z gałęzi meta wybierz **FilteredClassifier**:
   1. W ustawieniach wybierz classifier **NaiveBayes** oraz filter **StringToWordVector**.
   2. W polu Test options wybierz **Supplied test set** i wskaż plik ReutersGrain-test.arff.
   3. Uruchom klasyfikację wciskając **Start**. Zapoznaj się z otrzymanymi wynikami.
3. Wykonaj klasyfikację w analogiczny sposób korzystając z **NaiveBayesMultinomial** oraz algorytmu tworzenia drzewa decyzyjnego **J48**. Porównaj wyniki.
4. Zapoznaj się z opisem filtru **StringToWordVector**. Jak myślisz, które jego opcje mogłyby poprawić klasyfikację? Zwróć uwagę np. na opcje outputWordCounts, lowerCaseTokens, useStoplist. Przetestuj działanie wybranych opcji pojedynczo i w grupach korzystając z algorytmu **NaiveBayesMultinomial**. Jak wpłynęły na jego skuteczność?