Web Scale Data Mining homework 1

Karol Kubicki

8.12.2014

2.1) You are analysing a stream which contains s singletons (items occurring once) and d pairs of items, in some random order. You are sampling the stream using Algorithm 2, which uses a pseudorandom number generator and takes each item for further processing with probability 1/100 . What is the expected value of the fraction returned by the algorithm, as a function of s and d? Hint: follow the analysis in Section 4.2.1 of [10] (references in the textbook).

The correct answer would be: $\frac{d}{s+d}$. The answer from algorithm 2:

- number of obtained singletons: $\frac{s}{100}$,
- number of doubles obtained twice: $\frac{d}{100*100} = \frac{d}{10000}$,
- number of doubles obtained once: $d * 2 * \frac{1}{100}\frac{99}{100} = \frac{198d}{10000}$,
- expected value of the fraction returned by the algorithm: $\frac{d}{100s+199d}$.

2.2) Implement a naive Bayes classifier: • download from any source a small collection of documents annotated by several classes/topics; • split input set into training and test subsets; • implement a naïve Bayes classifier – do not use existing implementations; • check accuracy of predictions on a test set; • report results in the form of working, roughly documented code with the input set; • you are free to choose your favourite programming language.

I used documents from following database: http://ai.stanford.edu/~amaas/data/sentiment/ It provides you with documents divided into train and test sets. In each of them you have film reviews labeled as:

- positive - a review with rating equal or greater then 7 (maximum is 10),
- negative - a review with rating less or equal to 4.

From the aforementioned database I have randomly selected:

- 4000 training examples (2000 positive and 2000 negative),
- 2000 test examples (1000 positive and 1000 negative).

Nextly, I have trained naive Bayes classifier and checked its performance (see wsdm_homework2_naiveBayesCode_kubicki.pdf). I got 0.8065 accuracy.