Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Björn Buchhold

Information Retrieval
WS 2013/2014

http://ad-wiki.informatik.uni-freiburg.de/teaching

UNI FREIBURG

Exercise Sheet 11

Submit until Tuesday, January 28 at 4:00pm

Exercise 1 (10 points)

Prove that Naive Bayes for text documents, as explained in the last lecture and as implemented for the last exercise sheet, is a linear classifier when there are exactly two classes.

Hint: Proceed as follows (and make sure you understand the argument). Let $w_1, ..., w_m$ be the distinct words from all the text documents. Assume the two classes are called A and B. Let $\Pr(A)$, $\Pr(B)$, $\Pr(w_i|A)$, and $\Pr(w_i|B)$ be the prior probabilities computed in the training phase of Naive Bayes. Let $H = \{x : w \cdot x = b\}$ be a hyperplane in $\mathbb{R}^m$, where w is an m-dimensional vector with ith component equal to $\log_2(\Pr(w_i|A)/\Pr(w_i|B))$ and $b = -\log_2(Pr(A)/Pr(B))$. Then prove that Naive Bayes classifies an object x as A if and only if x lies on one particular side of H.

Exercise 2 (10 points)

Copy your code from the last exercise sheet and extend it such that it outputs two additional pieces of information. First, output the number of outliers in the training set. Second, output the width of the maximal margin around the hyperplane defined in Exercise 1.

Hint: Let H be the hyperplane as defined in Exercise 1 (you can use that definition, even if you did not do the proof). For each document, compute on which side of H it lies. From this, you can easily compute the number of outliers = documents on the "wrong" side of the hyperplane. Also compute the minimal distance of a non-outlier document from each class to H. The sum of these two distances gives you the margin size.

Run your code on the two data sets linked on the Wiki. Enter your results in the table linked on the Wiki. In your *experiences.txt*, briefly discuss how they compare with our SVM results from the lecture (first row in the result table linked on the Wiki).

Add your solution to Exercise 1 (PDF) and your code for Exercise 2 to a new sub-directory *exercise-sheet-11* of your folder in the course SVN, and commit it. Make sure that *compile*, *test*, and *checkstyle* run through without errors on Jenkins. As usual, also commit a text file *experiences.txt* where you briefly describe your experiences with this exercise sheet and the corresponding lecture. As a minimum, say how much time you invested and if you had major problems, and if yes, where.