

Exercise Sheet 11

Submit until Tuesday, January 26 at 2:00pm

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 v_1, \dots, v_m be the distinct words from all the text documents. Assume the two classes are called A and B . Let $p_A = \Pr(C = A)$, $p_B = \Pr(C = B)$, $p_{iA} = \Pr(W = v_i | C = A)$, and $p_{iB} = \Pr(W = v_i | C = B)$ 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 $w_i = \log_2(p_{iA}/p_{iB})$ and $b = -\log_2(p_A/p_B)$. Then prove that Naive Bayes classifies an object x as A if and only if x lies on one particular side of H . The proof is easier than it might first seem from this hint.

Exercise 2 (10 points)

Implement a perceptron classifier, as explained in the lecture. You find Python code for reading the input files + a class skeleton on the Wiki, similar as for the last exercise sheet. Note: learning iteratively considers each training example - there is no linear algebra solution that avoids this.

Evaluate your code on the two new sets of train+test data provided on the Wiki. Post your results on the table on the Wiki. Write the top-20 features (with largest absolute weight) and their weights to files *features-genre.txt* and *features-ratings.txt*, respectively. Briefly discuss your results and the top features, similar as for the last exercise sheet.

Also run Naive Bayes again (using your own solution or the master solution) on the two new sets of train+test data. Compare the precisions and the words with the largest absolute weights (positive or negative). Briefly discuss whether you think the quality differences are significant.

Add your solution to Exercise 1 (as a PDF) and your code for Exercise 2 to a new sub-directory *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. Express your further deepening love and appreciation for linear algebra.