

Assignment 4

Information Retrieval and Text Mining 18/19

2018-12-17; to be submitted 2019-01-10

Roman Klinger, Johannes Erwerle

- **Groups:** Working in groups of up to three people is encouraged, up to four people is allowed. More people are not allowed. Copying results from one group to another is not allowed. Changing groups during the term is allowed.
- **Grading:** Passing the assignments is a requirement for participation in the exam in all modules IRTM can be part of. Altogether 80 points need to be reached. There are five assignments with 20 pen & paper points and 10 programming points each. That means, altogether, 150 points can be reached. Explain all solutions.
- **Submission:** First make a group in Ilias, then submit the PDF. Write all group members on each page of the PDF. Only submit *one* PDF file. If you are technically not able to make a group (it seems that happens on Ilias from time to time), do not submit a PDF multiple times by multiple people – only submit it once. Submission for the programming tasks should also be in the same PDF.

Task 1 (Naïve Bayes) 6 points

Train a Naïve Bayes (the version of the model which we discussed in class) given the following documents annotated with classes c_1 and c_2 . Use Add-One-Smoothing. Provide all parameters for a full model specification.

c_1 “happy new year”

c_1 “happy holiday”

c_2 “study again”

c_2 “study year”

Given the document

- “happy work”

Which class is assigned by the model?

Task 2 (Maximum Entropy Classification) 6 points

Given the following features (without making a difference between upper and lower case) and documents:

weight	feature	Class y	document
$\lambda_1 = 0.2$	$f_1(y,x) = 1$ if “\$” in x and y = SPAM	SPAM	$x_1 = \$1$ million from Nigerian defense minister
$\lambda_2 = -0.1$	$f_2(y,x) = 1$ if “\$” in x and y = HAM	SPAM	$x_2 =$ Please contact Nigerian finance minister
$\lambda_3 = 0.5$	$f_3(y,x) = 1$ if “Nigerian” in x and y = SPAM	SPAM	$x_3 =$ You won \$30,000!
$\lambda_4 = -0.2$	$f_4(y,x) = 1$ if “Nigerian” in x and y = HAM	SPAM	$x_4 =$ Buy these Ginsu knives now.
$\lambda_5 = -0.1$	$f_5(y,x) = 1$ if “you” in x and y = SPAM	HAM	$x_5 =$ You should send the Nigerian wildlife report.
$\lambda_6 = 0.4$	$f_6(y,x) = 1$ if “you” in x and y = HAM	HAM	$x_6 =$ Thanks for great dinner. I owe you \$20.
$\lambda_7 = 0.1$	$f_7(y,x) = 1$ if y = SPAM		
$\lambda_8 = 0.0$	$f_7(y,x) = 1$ if y = HAM		

Subtask 2.1, 4 points

What is $p(\text{SPAM}|x_1)$ given the maximum entropy classifier with the specified features and weights?

Subtask 2.2, 4 points

Calculate the partial derivative of the log-likelihood of all documents with respect to λ_6 !

Task 3 (Maximum Entropy Training) 4 points

Given the following instances with labels:

- Class 1: TermA TermB
- Class 1: TermA TermB
- Class 2: TermD
- Class 2: TermE

List all relevant features (f_i , crossproduct of terms with classes) with weights (λ_i) which you expect to be optimal in a maximum entropy classifier trained on these examples (there are multiple possible solutions which would lead to the same classification result). Explain your solution.

Task 4 (Evaluation) 4 points

Explain in your own words what the difference between macro and micro averaging is, when calculating the F measure!

Please make an example with 10 instances and four different classes which shows a lower micro average F score than macro average F score.

Please list ten combinations of gold and predicted classes. Explain why your solution leads to the proposed relationship between micro and macro F score.

Programming Task 4 (10 points)

The assignment data contains two files `games-train.csv` and `games-test.csv`. These are German app reviews for games (a subset of the data described in http://www.lrec-conf.org/proceedings/lrec2016/pdf/59_Paper.pdf).

The files are formatted as follows:

- Column 1: Title of game
- Column 2: Class of review (good or bad)
- Column 3: Title of review
- Column 4: Review text

Title and review texts can be empty.

Subtask 1, 6 points

Implement a naive Bayes classifier from scratch which predicts the class (good, bad) stated in column 2. You can use all information from the training file to build your classifier. You are free in choosing the meta-parameters (smoothing, stop-word deletion, stemming, preprocessing) or optimizing those on validation data/via cross validation.

Which terms have the highest importance, according to the model?

Subtask 2, 4 points

Implement an evaluation system to Subtask 1. What is your precision, recall, and F to predict the class good and what is your precision, recall, and F to predict the class bad? Also report the numbers of TP, FP, FN. Discuss your results.