

## IN4080, 2020, Exercise set 4: 10 Nov.

Review of some ethics papers:

- Ziyuan Zhong, "A tutorial on Fairness in Machine Learning", Towards Data science. NB: you can skip Section 5 of the text.

<https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb>

- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144). NB: you can skip Section 4 of the paper, as well as the details of the experimental design and evaluation results.

<https://sameersingh.org/files/papers/lime-kdd16.pdf>

### Exercise 1

Assume we are classifying sentences into two classes, sentiment-bearing and non-sentiment-bearing and get the following confusion matrix.

		Gold	
		sentiment	non-sentiment
Predicted	sentiment	50	50
	Non-sentiment	150	750

- What is the accuracy of this classifier?
- What is the recall, precision and f-score for the sentiment bearing class?
- What is the recall, precision and f-score for the non-sentiment bearing class?

### Exercise 2

J&M, Ch. 4, sec 4.7 *Evaluation: Precision, Recall, F-measure* shows an example with 3 classes and calculates P, R and F for each class and micro- and macro-averaged P(recision) across the 3 classes. Calculate micro- and macro-averaged R(ecall) and F(-measure) across the 3 classes.

### Exercise 3

When applying P, R and F, we should keep in mind that there are two different kinds of scenario. They differ in particular when it comes to micro averaging.

- In the general case, as we see e.g., in information retrieval and span-evaluation for chunking or NER, the total number of predicted items and the total number of gold items may differ.
- In the special case of classification, where a classifier puts each item into exactly one class, there are equally many predicted items as gold items. This was the case in exercise 1 and 2 above.

To see the difference, consider the following. We are evaluating a NER, which considers only three classes, and we count the following numbers.

	True pos.	False pos.	False neg.
Person	720	180	80
Organization	180	20	60
Location	60	0	20

- a) Calculate P, R and F for each class
- b) Calculate macro- and micro-averaged P, R and F