Deep Learning for NLP 2020
Exercise 02

April 28, 2020

1 Pingo

Try to find the right answer(s) to each question on your own or in a group with your colleagues. The interactive survey will be conducted near the end of the practice class.

- Which of the following statements about precision/recall are correct?
  - ☐ A model which always predicts class A has a precision of 100% for class A
  - ☐ A model which always predicts class A has a precision of 0% for class A
  - ☐ A model which always predicts class A has a recall of 100% for class A
  - ☐ A model which always predicts class A has a recall of 0% for class A
  - ☐ F1 is a combination of precision and recall
  - ☐ F1 is a combination of precision, recall and accuracy
- Which of the following activation functions are continuously differentiable?
  - ☐ Unit Step (Threshold)
  - ☐ Sigmoid
  - ☐ tanh
  - ☐ ReLU
  - ☐ Softplus
- Cross-entropy loss...
  - ☐ ...is the natural choice when using softmax as the activation function
  - ☐ ~~...is based on the distance between two probability distributions~~
  - ☐ ...is inferior to square loss for multi-class problems
- A perceptron can...
  - ☐ ~~...separate data with a hyperplane~~
  - ☐ ...solve the OR problem
  - ☐ ...solve the AND problem
  - ☐ ...solve the XOR problem
  - ☐ ...decide all linearly separable sets

2 Machine Learning Fundamentals

2.1 Datasets

State two benefits / useful applications of a development dataset.

2.2 Evaluation Measures

Precision, recall and F1 measure are typical measures for evaluating the results of machine learning systems. Assume you built a simple POS-tagger which only operates on the three tags `NN`, `VB` and `ADJ`.

1. Compute precision, recall and F1 measure for each individual class based on the following confusion matrix:

		predicted class		
		`NN`	`VB`	`ADJ`
true class	`NN`	25	5	1
	`VB`	2	15	12
	`ADJ`	1	6	0

   Hint: For n classes and a confusion matrix $C \in R^{n \times n}$, the evaluation measures are defined for class i by:

$$P_i = \frac{\text{TP}}{\text{TP+FP}} = \frac{C_{i,i}}{\sum_{j=1}^{n} C_{j,i}}$$
$$R_i = \frac{\text{TP}}{\text{TP+FN}} = \frac{C_{i,i}}{\sum_{j=1}^{n} C_{i,j}}$$

   and

$$F1 = \frac{2 \cdot P \cdot R}{P + R}$$

2. Compute the micro-/macro-averaged variant of precision, recall and F1 across all classes.

   Hint:

$$P_{\text{micro}} = \frac{\sum_{i=1}^{n} \text{TP}_i}{\sum_{i=1}^{n} \text{TP}_i + \text{FP}_i} \widehat{=} \frac{\texttt{sum(diag(C))}}{\texttt{sum(C)}}$$
$$R_{\text{micro}} = P_{\text{micro}}$$
$$P_{\text{macro}} = \frac{1}{n} \cdot \sum_{i}^{n} P_i$$
$$R_{\text{macro}} = \frac{1}{n} \cdot \sum_{i}^{n} R_i$$

3. Explain the difference between the micro- and macro-averaged variants. Which variant is better suited for which occasion?

2.3 Meaningful Research

> It is the year 2015. A friend of yours played around with a Bidirectional Long-Short Term Memory Conditional Random Field Model (BiLSTM-CRF) for part-of-speech (POS) tagging. A single test run with his model on the Penn Treebank corpus (which was created in 1992) yields 97.55% accuracy. This marks a 0.39% improvement over the previous state of the art, a Support Vector Machine (SVM) baseline from the year 2004. Given that this is a new state-of-the-art result, your friend plans to submit a paper on the model to ACL 2015.

Which issues do you spot in your friend's research approach? What would you recommend?