Deep Learning for NLP 2020
Exercise 02 Solution

May 4, 2020

1 Pingo

- 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
  - x 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
  - x 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)
  - x Sigmoid
  - x tanh
  - □ ReLU
  - x Softplus
- Cross-entropy loss...
  - x ...is the natural choice when using softmax as the activation function
  - x ...is based on the distance between two probability distributions
  - □ ...is inferior to square loss for multi-class problems
- A perceptron can...
  - x ...separate data with a hyperplane
  - x ...solve the OR problem
  - x ...solve the AND problem
  - □ ...solve the XOR problem
  - x ...decide all linearly separable sets

2 Machine Learning Fundamentals

2.1 Datasets

State two benefits / useful applications of a development dataset.

- hyperparameter optimization
- early stopping
- (avoid overfitting the test set)

2.2 Evaluation Measures

1.

$$
\begin{aligned}
P_{\text{NN}} &= \frac{25}{25+3} && = 0.89 \\
R_{\text{NN}} &= \frac{25}{25+6} && = 0.81 \\
F1_{\text{NN}} &= \frac{2 \cdot 0.89 \cdot 0.81}{0.89+0.81} && = 0.85
\end{aligned}
$$

$$
\begin{aligned}
P_{\text{VB}} &= \frac{15}{15+11} && = 0.58 \\
R_{\text{VB}} &= \frac{15}{15+14} && = 0.52 \\
F1_{\text{VB}} &= \frac{2 \cdot 0.58 \cdot 0.52}{0.58+0.52} && = 0.55
\end{aligned}
$$

$$
\begin{aligned}
P_{\text{ADJ}} &= \frac{0}{0+13} && = 0.00 \\
R_{\text{ADJ}} &= \frac{0}{0+7} && = 0.00 \\
F1_{\text{ADJ}} &= \text{undefined}
\end{aligned}
$$

2.

$$
\begin{aligned}
P_{\text{micro}} &= \frac{25+15+0}{25+3+15+11+0+13} && = 0.60 \\
R_{\text{micro}} &= \frac{25+15+0}{25+6+15+14+0+7} && = 0.60 \\
F1_{\text{micro}} &= \frac{2 \cdot 0.60 \cdot 0.60}{0.60+0.60} && = 0.60
\end{aligned}
$$

$$
\begin{aligned}
P_{\text{macro}} &= \frac{1}{3} \cdot (0.89+0.58+0.00) && = 0.49 \\
R_{\text{macro}} &= \frac{1}{3} \cdot (0.81+0.52+0.00) && = 0.44 \\
F1_{\text{macro}} &= \frac{2 \cdot 0.49 \cdot 0.44}{0.49+0.44} && = 0.46
\end{aligned}
$$

3. **micro-averaging:** averages on the level of test instances; classes with large number of instance have strong influence on result

   macro-averaging: averages on the class level; classes with a small number of instances keep their influence

   Usage: It depends on what one wants to show. Use micro, when performance on largest classes is of importance. Use macro, when performance on small classes is of importance (which would go unnoticed with micro-averaging).

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.

- Issues:
  - Training neural networks is a non-deterministic process (batching of data, etc.), improvement could be the result of a lucky random seed
  - Independent of the RNG, improvements might not be significant
  - Repeatedly publishing papers to report improvements on the same test dataset is also a form of overfitting
- Recommendations:
  - When it matters, do multiple runs with different random seeds, then report a distribution
  - Your friend should test the model and the SVM on entirely new, unseen test data (if available)
  - If possible, a study should be conducted to evaluate if humans notice the model improvement in a usage scenario (admittedly, such a scenario is difficult to find for POS tagging)
  - Choose research questions with higher impact