Luleå University of Technology

D7047E - Advanced Deep learning

03 April 2023

Group 1 | Ex.1 | Berezin Ilya

Contents

Task 1.1. Loss Functions

The following formulas are useful for doing the exercise, where n denotes the length of both the prediction vector y and the ground truth vector g.

Cross-Entropy Loss (or Logistic Loss)

$$H(y,g) = -\sum_{i}^{n} g_i \log(y_i)$$

Mean Squared Error-Loss

$$MSE(y,g) = \frac{1}{n}\sum_{i}^{n}(y_i - g_i)^2$$

Hinge Loss (or SVM Loss)

$$SVM(y,j) = \sum_{i|i\neq j}^{n} \max(0, y_i - y_j + 1)$$

where j is the index of the correct label for the sample.

Consider the following two vectors:
Prediction $g = [0, 1, 0]$
Ground truth $y = [0.25, 0.6, 0.15]$

Cross-Entropy Loss

Cross-Entropy Loss calculated using the formula:

H(y, g) $= -(g_1 * \log(y_1) + g_2 * \log(y_2) + g_3 * \log(y_3)) = -(0 * \log(0.25) + 1 * \log(0.6) + 0 * \log(0.15)) = -\log(0.6) =$ **0.5108**

Mean Squared Error Loss

Mean Squared Error Loss calculated using the formula:

MSE(y, g) $= (1/3) * ((0.25 - 0)^2 + (0.6 - 1)^2 + (0.15 - 0)^2) = (1/3) * (0.0625 + 0.16 + 0.0225) =$ **0.0817**

Hinge Loss

Hinge Loss calculated using the formula:
The correct class index $j = 1$ ($g = [0, 1, 0]$)

SVM(y, 1) $= \max(0, y_1 - y_2 + 1) + \max(0, y_3 - y_2 + 1) = \max(0, 0.25 - 0.6 + 1) + \max(0, 0.15 - 0.6 + 1) = \max(0, 0.65) + \max(0, 0.55) =$ **1.2**

Hinge Loss between g and y = 1.2

Task A. Theoretical Foundations

Typically, people refer to accuracy as THE evaluation metric, but there are a lot of evaluation metrics which can be better suited than accuracy depending on the task/dataset.

1. In which situation using accuracy is not necessarily a good idea?

   Using accuracy could be not a good idea in case of imbalanced dataset, where model can achieve high enough accuracy predicting main class, even though it may perform poorly on a minor class. This could be an issue in prediction of medical diagnosis, where minor class could be important.

2. What part of the formula for computing the accuracy makes it less desirable than the Jaccard Index (Intersection Over Union) in a multi-class setting?

   The accuracy metric computes the proportion of a correct predictions out of total predictions made. In case of a multi-class modelling, it could be less desired than the Jaccard Index, since the accuracy takes in consideration both TP and TN. If number of TN is much higher (imbalanced datasets), the accuracy can be misleadingly high. Jaccard Index considers only TP, FP and FN, making itself more robust to a class imbalance and focused on performance of classifier in terms of correct predictions.

3. What is the difference between Jaccard Index (Intersection Over Union) and F1-Measure? Which one is more suited to measure performance of neural networks?

   Jaccard Index is the ratio of TP/(TP+FP+FN) and focuses on the overlap between the predicted and ground truth data.

   F1 measure is the harmonic mean of precision and recall focused on a balance between those metrics.

   Metric should be chosen based on specific problem and dataset characteristics. F1 is more suitable for classification tasks, providing balance between precision and recall, while JI can be more suitable in case of multi-level classification.

Task B. Practice

In this part of the exercise, we want to compute some common alternatives which can be used instead of accuracy. We'll take an example from a real case scenario of layout analysis at pixels level of historical documents.

Given the following prediction and ground truth (note: this is a multi-class and multi-label scenario), where B stays for background, T for text, D for decoration and C for comment.

```
       |1  2  3  4  5  6  7  8 |

GT     |B  B  B  B  TD TD TD TD|

P      |B  T  TD BD BC TC T  TD|
```

Class frequencies and the metrics per class:

Class frequencies for B, T, D, and C are 1, 3, 3, 1.

- Jaccard Index

  B = $\frac{1}{2}$ = 0.5

  T = $\frac{1}{3}$ = 0.3333

  D = $\frac{2}{4}$ = 0.5

  C = $\frac{0}{2}$ = 0

- Precision
  B = $\frac{1}{1}$ = 1

  T = $\frac{1}{2}$ = 0.5

  D = $\frac{2}{3}$ = 0.6667

  C = $\frac{0}{1}$ = 0

- Recall
  B = $\frac{1}{1}$ = 1

  T = $\frac{1}{3}$ = 0.3333

  D = $\frac{2}{3}$ = 0.6667

  C = $\frac{0}{1}$ = 0

- F1-measure
  B = $\frac{2*1*1}{1+1}$ = 1

  T = $\frac{0.5*0.3333}{0.5+0.3333}$ = 0.4

  D = $\frac{2*(0.6667*0.6667)}{0.6667+0.6667}$ = 0.6667

  C = $\frac{0}{0}$ = na

Then compute their mean in two different ways: once with class balance (sum of per class values divided by number of classes) and once with the class frequencies (per class values times the class frequency).

Mean with class balance:

- JI: $\frac{0.5*1+0.3333*3+0.5*3+0*1}{8}$ = 0.375
- P: $\frac{1*1+0.5*3+0.6667*3+0*1}{8}$ = 0.5625
- R: $\frac{1*1+0.3333*3+0.6667*3+0*1}{8}$ = 0.5
- F1: $\frac{1*1+0.4*3+0.6667*3}{8}$ = 0.5333

Mean with class frequencies:

- JI: $\frac{0.5*1+0.3333*3+0.5*3+0*1}{8}$ = 0.375
- P: $\frac{1*1+0.5*3+0.6667*3+0*1}{8}$ = 0.5625
- R: $\frac{1*1+0.3333*3+0.6667*3+0*1}{8}$ = 0.5
- F1: $\frac{1*1+0.4*3+0.6667*3}{8}$ = 0.5333

Compute the global Exact Match metric.

Exact Match is 1/8, as only one prediction matches the ground truth perfectly.

1. Intuitively the Exact Match would be the strictest metric possible. However, it might not be the lowest number: how come?

   Exact Match metric is the strictest one since it counts only the instances where the predicted label matches the ground truth exactly. It can result in not the lowest number, and JI, Precision, Recall or F1 might be lower in some cases.

2. Can you compute the global accuracy on this example? Justify your answer.

   Global accuracy is not well suited for multi-label classification. Accuracy measures the rate of correct predictions out of the total number, however in multi-label dataset it becomes ambiguous. One single instance may have multiple correct labels and accuracy doesn't count for partial matches.

3. What is the main issue going from multi-class to multi-label setting?

   The main issue is that in multi-class an instance can belong to only one class, while in multi-label to multiple. It complicates the evaluation of a model performance as accuracy metric becomes less meaningful. In a multi-label scenario, it's crucial to consider overlap between the predicted and ground truth sets, so JI, Precision, Recall and F1 are more suitable.