CO553 - Introduction to Machine Learning: Evaluation

Prepared by Marek Rei

Autumn 2020/2021

1 Questions

Here is a set of various questions to improve your understanding of machine learning evaluations.

1. You are given a dataset of 10,000 ECG recordings, together with corresponding labels that indicate whether the patient had ventricular fibrillation (a type of cardiac arrhythmia). You need to develop a classifier to assign these labels automatically. How do you set up and use the dataset?

2. Instead of 10,000 examples, you get 200 examples. How does that change your setup?

3. You have built a model to predict the sentiment of a tweet: whether the tweet is positive, negative of neutral. Given 12 examples, this is the output you get:

Datapoint ID	True sentiment	Predicted sentiment
1	neutral	neutral
2	neutral	negative
3	negative	negative
4	positive	neutral
5	neutral	negative
6	neutral	negative
7	neutral	neutral
8	negative	neutral
9	neutral	neutral
10	positive	positive
11	positive	positive
12	neutral	neutral

   (a) Construct the confusion matrix.

   (b) Calculate accuracy.

   (c) Calculate precision, recall and F1 for each class.

   (d) Calculate macro-precision, macro-recall and macro-F1

4. Which evaluation metric would you want to observe most closely for the following tasks? Note: this will not be a comprehensive list of possible valid evaluation metrics for each of these tasks. Just the most likely candidates.

   1. Predict the amount of rain for tomorrow.

   2. Detecting grammatical errors in a sentence.

   3. Identifying the type of land in an aerial photo (e.g., crops, forest, buildings, meadow, etc).

5. You've trained a model. It gets very good performance on the training set but bad performance on the validation set. What is happening and what can you do?

6. You've trained another model. Now it gets bad performance both on the training and validation set. What is happening and what can you do?

7. You've trained one more model. This time you get unexpectedly good performance on the validation set. Much better than you would have expected. Time to celebrate?

8. You use a neural network classifier to detect whether a photo contains a stop sign or not. The model takes 200x300 pixel images as input. You train on 5000 images and test on 500 images. 40% of the images in either dataset contain the stop sign. The model accuracy is reported as 84%. Calculate the error rate and its confidence interval at 95%.

9. What does it mean when the paper reports that the performance difference between system A and system B is statistically significant?