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?

   Answers:

   1. Shuffle the dataset, to make sure that the datapoints are not ordered based on some criteria.
   2. Split the dataset into training, validation and test sets. For example, 80%/10%/10% split. There must be no overlap of examples between the training, validation and test sets.
   3. Optimize your model parameters on the training set.
   4. Evaluate different hyperparameter options on the validation set. Can also use it for early stopping. Can evaluate on the validation set many times.
   5. Evaluate on the test set once the best model architecture has been chosen. This gives an indication of how well the model performs of held-out examples from the same distribution. Minimize the number of times you evaluate on the test set.

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

   Answer:

   1. Use cross-validation instead. After shuffling, split the dataset into N different parts (folds), for example N=10.
   2. Use one fold for testing, the others for training+validation.
   3. Rotate which fold you use for testing, until you've used all N parts.
   4. Average the performance across all the folds. This will be your overall performance on this dataset. You've essentially used the whole dataset both for training and testing, while not training and testing on the same examples in the same experiment.
   5. If you need to perform hyperparameter tuning, you can either
      i. Use some part of the training+validation data at each iteration for choosing the best hyperparameter values. For example, rotating the validation folds as well. This might not work well when the dataset is very small, as your validation set will be tiny and not very representative.
      ii. Perform an internal crossvalidation on the training+validation data at each iteration. Works better but requires quite a bit more computation.
   6. The performance given by cross-validation shows how well your overall machine learning approach works, but doesn't really show the performance of a single model.
   7. In order to actually produce a single good model in the end, you could use the best hyperparameters and train a new model on the whole dataset.

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

   Answers:

True/predicted	negative	neutral	positive
negative	1	1	0
neutral	3	4	0
positive	0	1	2

Accuracy	0.583333333333
$P_{negative}$:	0.25
$R_{negative}$:	0.5
$F1_{negative}$:	0.333333333333
$P_{neutral}$:	0.666666666667
$R_{neutral}$:	0.571428571429
$F1_{neutral}$:	0.615384615385
$P_{positive}$:	1.0
$R_{positive}$:	0.666666666667
$F1_{positive}$:	0.8
P_{macro}:	0.638888888889
R_{macro}:	0.579365079365
$F1_{macro}$:	0.582905982906

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.
      Answer: MSE or RMSE
   2. Detecting grammatical errors in a sentence.
      Answer: F-measure for the error class

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

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?

   Answer:

   Sounds like your model might be overfitting. Depending on the model type, you could try regularization, early stopping, dropout, getting more training data or reducing the complexity of the model.

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?

   Answer:

   You might want to check the code for bugs first.

   If everything is working properly, then sounds like your model might be underfitting. Depending on the model type, you could try increasing the capacity of the model, training for longer or reducing regularisation.

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?

   Answer:

   Best to check that there aren't any mistakes. This commonly happens when

   (a) Data isn't properly split and the examples in the validation/test split are still present in the training data.

   (b) You accidentally include the correct answer as an input feature, making the task really easy for the model.

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%.

   Answer:

   $error\_rate = 0.16 \pm 0.032$

   The question contains some irrelevant information.

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

   Answer:

   There is less than 5% chance that this performance difference is due to sampling noise and the systems are actually comparable.