Machine Learning

ML Quiz #2

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)

Solutions Submissions

- Only submit BEFORE you listen to the answers during the lecture
- Submission form

Questions A

1. Sara wants to build cat/dog images dataset. She invited 10 friends with overall 7 cats and 5 dogs. She took 1000 images of each pet. What do you think of this dataset for machine learning?
2. In a volleyball classification task, 200 olympiad videos were downloaded. For every video, 100 1-min short video clips were extracted for winning and losing points. Then the 20000 were shuffled and split to train/test/val. What do you think of this dataset for machine learning?

Questions B: True or False

1. If the model performs well on the test set, it suggests that it has captured the *underlying patterns* in the training data and can make accurate predictions on *similar instances*
2. There is always a chance of encountering data in real-world that is significantly different from the training and test sets
3. If the training data contains *biases* that are not representative of the target population, the model may not be able to generalize well
4. The distribution of the data *may change over time*, rendering the model's learned *patterns outdated* and affecting its ability to generalize to new data

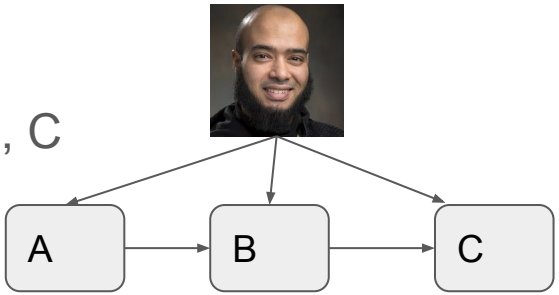
Questions C: True or False

1. The purpose of train-test splitting in machine learning is to assess the model's ability to generalize to unseen data
2. The training set is used to optimize the model's parameters, while the test set is used to tune hyperparameters
3. The purpose of using a separate test set is to provide an **unbiased** evaluation of the **final** chosen model's performance
4. Randomly **shuffling** the data **before splitting** it into train, validation, and test sets helps the independency of samples (as in IID)
5. If a model performs well on both the validation and test sets, it is guaranteed to have good performance on unseen data

Questions D: True or False

1. The quality and representativeness of the training data are crucial for generalization
2. Sometimes, a smaller but high-quality training set may outperform a larger but lower-quality training set.
3. A larger training set will always result in better generalization performance for a machine learning model

Question E



- In a computer vision pipeline, we have 3 models A, B, C
  - Input to A is an image of a person
  - Input to B is an image of a person + output of A
  - Input to C is an image of a person + output of B
  - **Final system output is output of C**
- Data Collection team collected 300 persons data.
- The data is provided for model's owner to use for his train/val/test purposes
- The performance of each model can be reported separately
- A more important performance is for the whole system (the pipeline)
  - Find a possible mistake in computing the pipeline whole performance