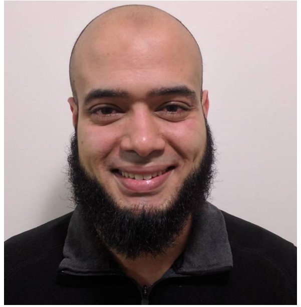
Machine Learning

Modeling Pipelines

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)

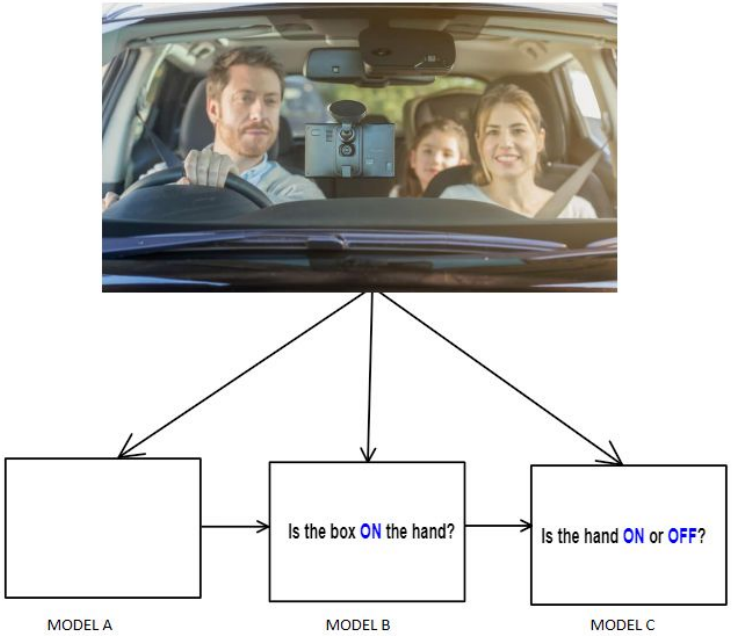


Note

- Brainstormed and summarized by students

Problems using Pipeline Modelling

Having a pipeline in Machine Learning may result in final performance degradation.



Is the box **ON** the hand?

Is the hand **ON** or **OFF**?

MODEL A

MODEL B

MODEL C

Train/Validation Environment

MODEL B: Is the box ON or OFF the hand.

Input: Cropped image on hand

Output: ON, probability like 0.9

MODEL C: Is the hand ON or OFF the wheel.

Input: Accuracy result from MODEL B **(0.9)**

Output: ON

Test Environment

MODEL B: Is the box ON or OFF the hand.

Input: Cropped image on hand

Output: ON, probability like 0.4

MODEL C: Is the hand ON or OFF the wheel.

Input: Accuracy result from MODEL B **(0.4)**

Output: OFF

Summary

The training and the testing should have the same distribution (Same I.I.D).

	MODEL B	**MODEL C**
Train/Val	The box fits perfectly on the hand (No noise). Output: Accuracy **(0.9)**	Good input accuracy from MODEL B (The box is on the whole hand). Output: ON
Test	The box may not fit the hand. ● The box is bit bigger ● Or Not fully on the hand ● The box is not even on the hand Output: Accuracy **(0.68)**	Semi accurate/noisy data. (The box does not fit the hand). Output: OFF (Even if the hand is really on)

Solutions

There are three ways to solve this problem.

- Divide the train data into 2 blocks.
- Add noise on the training data (Data Augmentation).
- Get rid of the pipeline.

Solution1: Divide the train data into 2 blocks

Motivation: Give each algorithm separate part of the data

If we ,for example, have 1000 images. We divide the data as follows.

	MODEL B	**MODEL C**
Data 1 (450 images)	Train	Do NOT use
Data 2 (450 images)	Val	Train (on Val from B)
Test Data (100 image)	Test	Test

CONS:

- the data should be big enough

Solution2: Add noise on the training data

- Motivation: we need our algorithm to be less sensitive to previous models mistake
- In our case: create scale invariant algorithm (hence less affected with previous model mistakes)
- Augment each picture to generate new versions with different types of boxes on the hand.
    - Ex: bigger , shifted right/left, smaller.. Etc. And train MODEL C with it.

PROS:

- We got bigger amount of data
- Scale invariant data (the differences do not affect the result)
- Same I.I.D for training and testing.

Solution3: Get rid of the pipeline

- Reduce the depth of the pipeline or get rid of it.
  - Think of a way to judge a given input image whether the hand is ON/OFF without using several models.
  - We need something like (one model: input => output).

“Acquire knowledge and impart it to the people.”

“Seek knowledge from the Cradle to the Grave.”