Organize the project files into folders.

I put the csv files into the Datasets folder.

I began working on the project report: starting with the introduction section.

Defines overfitting in the report:

V1

Overfitting occurs when a model can't identify the testing data correctly despite fitting the training data accurately.

V2

Overfitting is a problem caused by the selection of training data: while the model would successfully fit the training data, it would be susceptible to the testing data's unexpected changes.

SELECTED THE VERSION V2 FOR THE PROJECT DESCRIPTION

1. Project Description

Abstract: Over-fitting is a common problem for machine learning applications

Over-fitting is a problem caused by the selection of training data. Although the model would successfully fit the training data, the testing data would be susceptible to unexpected changes.

Continuous integration helps us to check for over-fitting problems. When the training data aren't representative of the dataset, the data will cause the model to overfit.

This project aims to create a platform for taking ML code and ML data files as inputs for running automated training. Building a continuous Machine Learning program can help to improve existing ML models. In this report, the goal is to apply the concepts of ML to building a continuous ML integration system that integrates with GitHub to run user-defined tests.

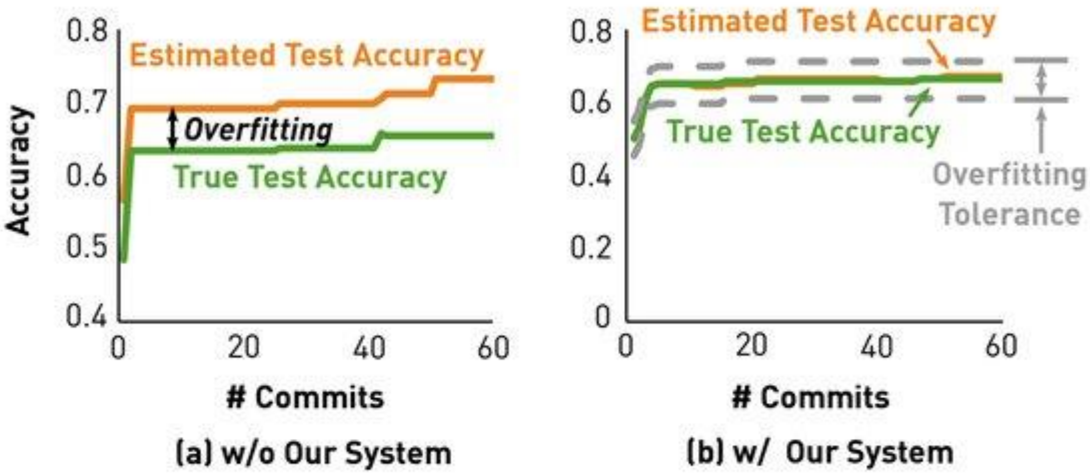
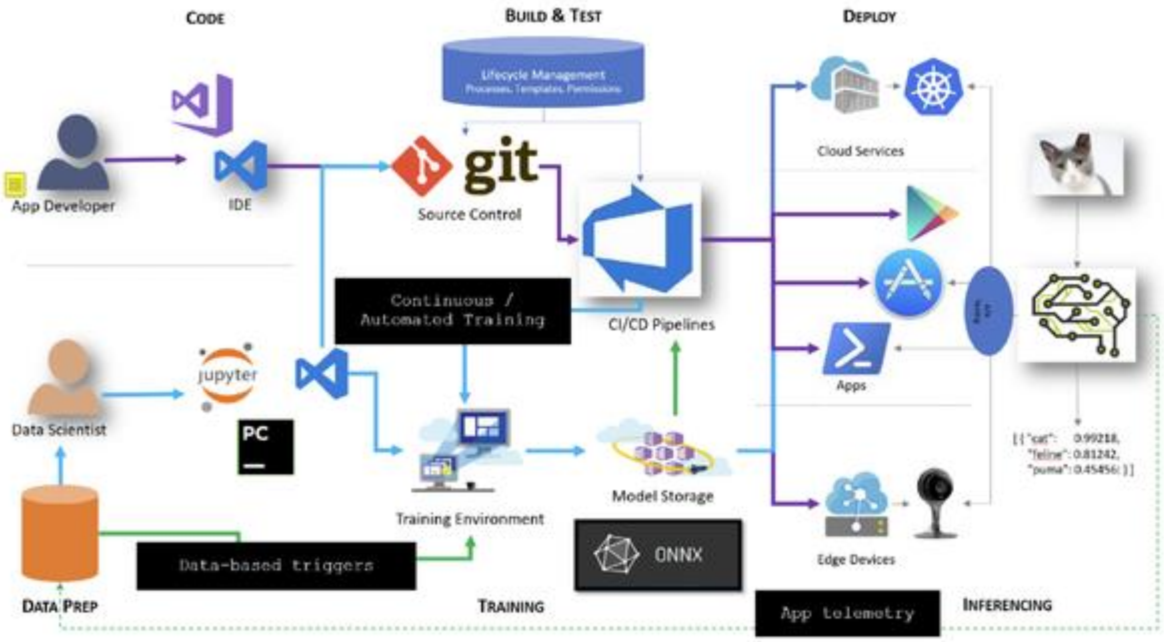


Figure 1: (a) The challenge of building a CI system for ML is that, if not being careful, one might overfit the test set when committing multiple models during the CI process; (b) The goal of our system is to provide rigorous guarantees on the overfitting behavior by, intuitively, measuring the "information leakage" from the test set during the CI process.

Continuous integration helps with improving the machine learning models. ML models are vulnerable to over-fitting. There's currently limited research into building new continuous ML integration systems. Some papers [1] are great for identifying system components, design, and project context.

Why is continuous integration required for machine learning?

Continuous integration helps us to check for over-fitting problems. When the training data isn't representative of the dataset, the data will cause the model to overfit. The new system will help to improve the model's score by using version control to compare new code with the previous versions. The diagram below helps to illustrate how continuous integration will support machine learning programming.



Prerequisite information

1. The user must learn about the machine learning life cycle for the context of the model quality.
2. GitHub Actions can verify new integrations into the system.
3. While GitHub can monitor the code changes, it is a more challenging task to track the Machine learning data files. Learn how to trace the changes in the ML data files.

Sources

1. Karlaš, B., Interlandi, M., Renggli, C., Wu, W., Zhang, C., Mukunthu Iyappan Babu, D., Edwards, J., Lauren, C., Xu, A., &amp; Weimer, M. (2020). Building Continuous Integration Services for machine learning. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery &amp; Data Mining. Building Continuous Integration Services for Machine Learning | Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining