Exploration: The Machine Learning Process (w1)

Introduction

Now that we have defined what machine learning is, a natural question to ask is, "How do we *'do'* machine learning?" In this Exploration, we will unpack the machine learning process, and dig deeper into training, validating and applying machine learning models. We'll also see a brief example of "where the math is" within the machine learning process.

The Machine Learning Process

Machine learning involves data, code, mathematics and deployment within software systems. While we do write plenty of code, the code we write does not explicitly model the specific problem at hand. Rather, the code we write establishes a means of reading data, implementing the mathematics within a trainable model, validating the quality of that model, and then using that model to make new predictions.

Let's take a look at the general steps involved for any machine learning project, no matter how small or large it may be.

The 7 steps of machine learning

Machine learning begins with a problem statement, such as being able to predict what kind of beverage a drink is based on certain attributes, such as color and alcohol content. To apply a machine learning solution, we must:

1. Gather data suitable for our problem
2. Prepare the data, such as cleaning it, removing anomalies, scaling, and normalizing attribute values
3. Select a model (a chunk of code that uses particular mathematics and algorithms) appropriate for our problem and available data
4. Train our model using the data to incrementally improve our model's ability to make a high-quality prediction
5. Evaluate or "validate" our model, by testing it out with predictions that we know the answer to, and comparing its results with the "correct answers"
6. Tune model parameters to further improve our model, such as increasing training repetitions, or adjusting model "hyperparameters"
7. Employ our model in a larger software system, to make predictions on new, previously unseen data

Note that these steps are not gospel, and they represent a convenient way of thinking about the machine learning process within the scope of active practitioners. Since practitioners operate within larger organizations, such as businesses, product teams, government entities, or research teams, machine learning processes exist within a larger, project-based scope.

Let's listen to machine learning expert Dr. Andrew Ng about his perspective on the machine learning project lifecycle.

#2 Machine Learning Engineering for Production (MLOps) Specialization [Co…

Here we see similarities between the "project level view" of machine learning, and the previous, "practitioner level view" of machine learning. We must have a project goal, that ideally should start small. We must acquire and prepare data, select a model, train it, evaluate our model's performance, and fine-tune our model. Lastly, once we are satisfied with the performance of our model, we must deploy it within the context of a larger software system.

For example, consider this contrived scenario of predicting the quality of wine. A winery would like to know where they might invest the most money in promoting the top three wines they will produce this year, and they need to know this in advance of the wine being bottled, tasted and rated. The winery has a historical record of many attributes of the grapes that they have grown over time, and what grapes ended up in their most successful wines. ***[Accessible version for screen-reading software] (https://canvas.oregonstate.edu/courses/2025514/pages/exploration-questions-and-answers)***

1. ▶ What problem statement are we defining?
2. ▶ What data do we gather?
3. ▶ How do we prepare the data?
4. ▶ What model might we select?
5. ▶ How might we train the model?

6. ▶ How can we test the quality of our model?
7. ▶ How can we improve the quality of our model?
8. ▶ How can we deploy and use our model?
9. ▶ How must the system be maintained?

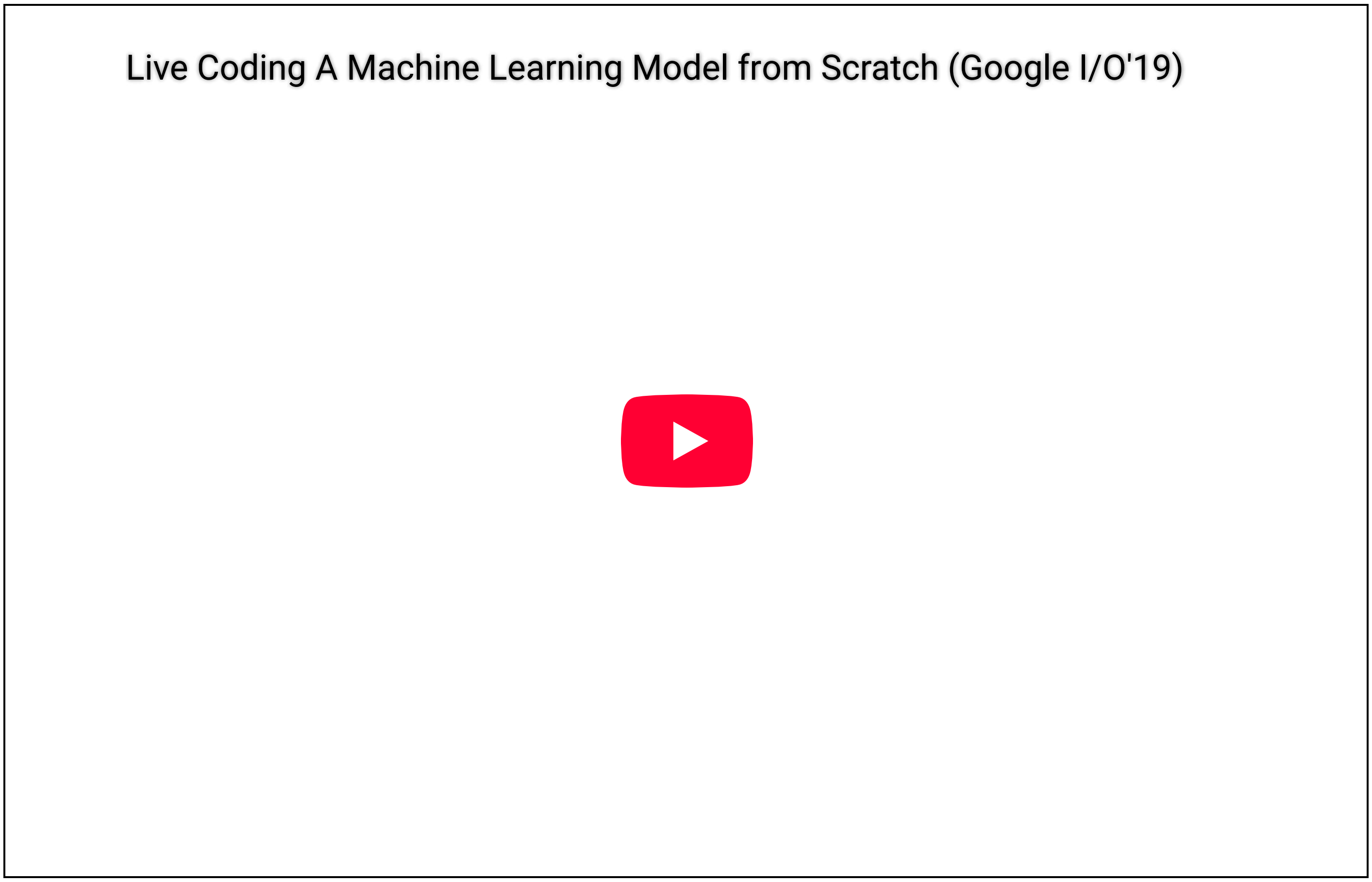
This is just one illustration of the machine learning process - your descriptions can certainly be different, and the provided examples are not the only approach to the problem. What is most important is seeing that machine learning is a process, and what the general steps of that process are.

Now, we have mentioned that machine learning involves code. Let's take a look at the machine learning process again, and see how these steps might manifest as code.

Intro to Machine Learning Lesson 1: How Models Work | Kaggle

Notice how, in this example, we see a classification problem, and the type of model is a supervised classification tree. While the presenter uses the term "fitting the model," we prefer the term "training the model" - you will see both terms used in the field of machine learning. More importantly, notice how the presenter has tried to illustrate these steps with code, such as `melbourne_housing_model.predict(housing_data)`. The general idea is authentic, but as practitioners we must write a lot more code! If only things were as easy as `figure_out_my_life(past_experiences)`. This is where you come in: by learning about machine learning first principles in this course, we hope to equip you with the knowledge and skills to define, create, validate and refine your machine learning models for successful deployment in solving real-world problems.

But wait! Doesn't machine learning involve mathematics and statistics? So far, we have abstracted the details of preparing data, using a model, and training a model - and this is where the mathematics lies. Take a look at the following presentation up to the 3m42s mark.



Notice how one of the presenter's problems involves image analysis and classification, such as labeling a photo of a cat as "cat." We do not merely feed a model visual pictures and expect it to know what to magically learn. Rather, when preparing the image data, we transform the pixel data of those images into numerical data contained in matrices. During training, we use some basic linear algebra to process the data, and a linear model of bias and weights (not magic!) to make predictions.

If the mention of linear algebra makes your face look like 🙀□😬🥴 , then do not worry. Most of the mathematics in machine learning is straightforward, and leverages statistics, some calculus and some basic linear algebra. If you are new to machine learning, you are in the right place. In this course, we will work together to combine data, code, mathematics, algorithms, and an experimental mindset to create predictive machine learning solutions.

Key Points

- Machine learning is a process involving data, code, experimentation, and deployment/use
- The steps of a machine learning process include acquiring data, preparing data, selecting a model, training a model, testing the model, and deploying the model

- This course enables us to practice the steps of the machine learning process
- The mathematics in machine learning mostly exists within data preparation, model training, and model prediction

Check Your Understanding

Before you continue, please respond to the following questions and prompts to verify your understanding. Click any ▶ disclosure arrow to view a possible answer. ***[Accessible version for screen-reading software] (https://canvas.oregonstate.edu/courses/2025514/pages/exploration-questions-and-answers)***

1. ▶ Do I need to have a really big problem to solve in order to try out machine learning?
2. ▶ When preparing our collected data, why is it important to set aside 20% of it, and not use it for training a model?
3. ▶ Is there usually one best model to use for a given problem or training data set?
4. Define the following terms.
   1. ▶ Hyperparameter
   2. ▶ Model
   3. ▶ Evaluation / validation
   4. ▶ Deployment

Additional Resources

Consider these curated resources an *essential*, if not mandatory, starting point for a deeper investigation into the topics within this Exploration.

- Reading: **How it Works (https://www.wolfram.com/language/introduction-machine-learning/how-it-works/)** Introduction to Machine Learning by Etienne Bernard, free from Wolfram
- Courseware: **Mathematics for Machine Learning and Data Science (https://www.deeplearning.ai/courses/mathematics-for-machine-learning-and-data-science-specialization/)** by DeepLearning.AI
- Tutorial: **Intro to Machine Learning (https://www.kaggle.com/learn/intro-to-machine-learning)** by Kaggle