Homework 1

CS 6678
Advanced Machine Learning
Instructor: Dr. Leslie Kerby

George Lake

January 19 2026

Part A - Core Content Engagement (The AI School)

This homework covers the Data Science & Non-Neural ML - Module 1 from the AI School online resource. Module 1 of the course covers the Foundations of Data Science and has the following sections:

- **A** - What is data science?
- **B** - The data science lifecycle
- **C** - Machine learning categories
- **D** - Ethics and responsibility
- **E** - Tooling and environment setup

I completed all of the sections in my own Jupyter Notebook and saved it to my GitHub repo.

My favorite part of Module 1 was setting up the scikit-learn pipeline and integrating the various packages (NumPy, pandas, matplotlib). I have used these packages before, but it has been piecemeal usage as I needed them. I think getting some organized structure to their usage will be very beneficial to me.

My least favorite part of Module 1 was me not having a good understanding of some of the programming they used when showing examples. However, I do not think this is a bad thing. I definitely have resources available, and have been using them, to learn more about what was being done.

Part B - Conceptual Questions

1. Prediction vs Decision-Making

In machine learning, there are fundamental concepts of prediction vs decision-making. A predictive model, or task, is focused on estimating a value or a state. A prediction is answering a question, what is the truth? (or what will happen?). A decision-making model, or task, is focused on choosing an action based on the prediction and consequences of that action. While the tasks are linked, decision-making is built on prediction, keeping them as separate ideas or functions gives flexibility in implementation.

For most scenario's we strive for high-accuracy models. However, this does not mean that they are perfect. You can still arrive at poor decisions even with these models. For example, we can look at the medical industry.

Imagine that we have a dataset that we are using for screening patients for a very rare disease affecting only a small percentage of the population, like 1%. Even with a highly accurate model (99%), the decision would be most likely to decide that everyone is healthy and send them home. The outcome is that everyone with the disease would be sent home without treatment. In this case, a lower accuracy model would most likely catch the sick people. It would lead to more false positives though.

2. When Machine Learning is the Wrong Tool

Machine learning is fundamentally probabilistic, it generates an approximation based on patterns. It is not a good tool to use when the rules of a system are fully known and require 100% accuracy. It is also not always the best tool to use when we already have efficient algorithms that are 100% accurate. For example, a company payroll system. This needs to be not just highly accurate, but 100%. Further, if we employed a ML model and an employee asked a question about why their pay was a certain way, we would not be able to answer. The decision to pay a certain amount would be determined by probability. We also have efficient tools for payroll that are 100% accurate. We can also state that machine learning uses inductive reasoning (based on what I have seen do this). It becomes less useful when we have a problem suited to deductive reasoning (starting with a universal truth).

3. Hidden Assumptions

I will explain three different assumptions that are commonly made when applying machine learning to real-world data.

1. **Independent and Identically Distributed** - This assumes that each data point in the dataset is generated independently of the others, and that the training data and the test data are drawn from the exact same underlying probability distribution.
   An example of a dataset that can violate this assumption is weather data. In this case, the data points are highly correlated with previous time steps. This is independence violation, which occurs when the outcome of one data point influences another.

   Another example of a violation is if a model is trained on medical data from hospital A, but tested on patients from Hospital B. The demographics (distribution) may differ. This is identical distribution violation, which occurs when the training data looks fundamentally different from the real-world data.
   Medium article

2. **Smoothness** - This assumes that if two input points ‘x1‘ and ‘x2‘ are close to each other in the input space, their corresponding outputs ‘y1‘ and ‘y2‘ should also be close (small changes in the input = small changes in the output). This allows a model to generalize unseen data by interpolating between known points.

   An example of a dataset that can violate this assumption is discontinuous or chaotic systems. For example, in a cryptographic hashing algorithm, changing a single bit of the input results in a completely different output. A standard ML model cannot learn a hashing function because the smoothness assumption breaks down. Small changes in the input lead to large changes in the output.
   Science Direct article

3. **Manifold** - The assumption that real-world, high-dimensional data (like images, audio, or text) is not randomly scattered across all dimensions. Instead, it lies on a lower-dimensional structure (a manifold) embedded within that high-dimensional space. An example, using a small 100x100 pixel image. To an algorithm, this is a 10,000-dimensional space, every single pixel is a separate coordinate. However, zoomed out, the picture represents something that is bound by lower dimensional space. The pixels in an image of a ball can only be changed so much before it does not look like a ball anymore.

   An example of a dataset that violates this assumption is encrypted network traffic. The manifold assumption relies on data having redundancy and patterns (in English, the letter Q is almost always followed by U). When analyzing encrypted data, this assumption does not hold.
   Medium article