**Introduction to Machine Learning for Materials Science Unit**

Goals:

* Understand core ideas of how machine learning models solve problems
* Describe and perform key steps in a machine learning workflow for materials design
* Identify strengths and weaknesses of machine learning

Target Audience and Background:

* Upper level undergraduate students in materials science and engineering
* Current materials targeted towards an introduction to computational materials course in which students will be learning about other tools and resources.

Materials Overview (See details section for more info):

* **Lecture Materials** - Divided up into **5 modules** which each have the three components below. Each module is estimated to take ~20 minutes of class time to complete.
  + **Reading materials** - to be completed before class.  
    Assignments are found in “Lecture Materials / Reading List” and are freely available
  + **Powerpoint Slides** - to be discussed during class  
    All slide decks are found in “Lecture Materials” and named by module number and title
  + **Quiz Questions** - either taken at end of class or as homework  
    Can be found in “Lecture Materials / Quiz Questions”. An answer document is also provided.
* **Jupyter Notebook Lab** - The lab jupyter notebook covers a specific example of designing and building a decision tree / random forest model to predict the band gaps of elementary and binary semiconductors and insulators. Activities and questions are provided throughout the notebook. The notebook and all necessary materials are found in the “Lab Materials” folder and relies on the existing file structure. A version is also available online at: <https://nanohub.org/tools/intromllab/>

Assessment:

* Assessment for lectures is provided by brief quiz questions that can be given at the end of lecture, or optionally as homework.
* Assessment questions for the lab activity are given throughout the Jupyter Notebook. Students submit their answers by completing cells within the notebook. Optionally Students can also copy paste their answers into a separate document which could then be submitted for grading.

Proposed Unit Structure:

|  |  |  |
| --- | --- | --- |
| Duration | Activity | Resources |
| 1 hr | Pre-class Reading | Reading 1-3 |
| 1 hr | In Class Lecture 1 | Powerpoint 1-3  Quiz 1-3 |
| 1 hr | Pre-class Reading | Reading 4-5 |
| 1 hr | In Class Lecture 2 | Powerpoint 4-5  Quiz 4-5 |
| 2 hrs | Jupyter Notebook Lab | Lab Notebook |

**Note: numbers next to lectures, quizzes, readings correspond directly to module numbers.**

Adapting Content of the Lab:

If lab time is restricted and you are looking to simplify or cut sections we recommend looking to the following first

1. Leave out the section on visualization of decision trees. This section reinforces existing evidence of overfitting, but isn’t required.
2. Skip exercises in sections 1-3. Students will still need to run these cells to make the notebook function, but these sections revolve around setting up the data to be ready for machine learning. Students can pick up in Section 4 where they start building models and setting up train and test splits.

Lecture Module Details:

The lecture materials have been divided into 5 modules which will familiarize students with the core concepts of machine learning and how it can be used in materials science. The overarching theme of the materials is to highlight how machine learning can potentially be used in a materials design workflow, and throughout the modules this objective is used as motivation and as a reference to call back to.

**Module 1**: Machine Learning Basics in Materials Science

Learning Goals

* Students remember machine learning is pattern matching
* Students can identify types of machine learning problems
  + Explain Regression vs Classification
  + Explain Supervised and Unsupervised Learning
* Students can describe structure of a decision tree model

**Module 2**: Data Cleaning and Featurizing

Learning Goals

* Students can describe key steps in a general machine learning workflow
  + Explain the purpose of data cleaning
  + Describe common methods for generating materials science specific features
  + Define several basic feature engineering steps (normalization, Removing correlated features, and constant features)
* Students can analyze the variables, inputs, outputs and aim of an ML problem
  + Describe how ML modeling efforts can fit into a larger materials research workflow. Specifically materials design.
  + Identify types of inputs and outputs for a model
  + Explain the benefits of feature normalization

**Module 3**: Model Assessment, Optimization, Predictions

Learning Goals

* Students can identify key features of a parity plot
* Students can explain the difference between training, testing, and validation data
* Students can explain the use for various error metrics (RMSE, MSE, MAE)
* Students can explain the difference between parameters and hyperparameters

**Module 4**: Challenges in Materials Science Machine Learning

Learning Goals

* Students can explain three common challenges and potential solutions

**Module 5**: Machine Learning Tools for Materials Science

Learning Goals

* Students can identify 2 tools for performing machine learning workflows
* Students can identify several resources for learning and performing machine learning on their own