An Introduction to Machine Learning for Materials Science: A Basic Workflow for Predicting Materials Properties

Ben Afflerbach
University of Wisconsin – Madison

5/12/2021
Summary

What is Machine Learning?
• Machine learning is a tool that finds patterns in large datasets that might be hard to discover otherwise

How can we use it for Materials Science?
• It can be included in existing materials science workflows to accelerate research, materials design, and materials discovery
An Application: Predict a Materials Property

Bandgaps in Semiconductors

Machine learning prediction here is obtained from only properties of the elements in the material!

A Basic Materials Design Workflow

- Identify Materials Properties
- Train Model of Properties
- Predict Properties For New Chemical Compositions
- Synthesize and Verify Predictions

**Training Details**

- Generate Training Data
- Data Cleaning
- Feature Generation and Engineering
- Model Assessment
- Model Optimization
- Predictions
Machine Learning is Pattern Matching

https://chem.libretexts.org/Bookshelves/Introductory_Chemistry (accessed May, 2020)
Key Distinction in ML

**Supervised Learning**
- Find a **Function** that represents the data

**Unsupervised Learning**
- Find **Structure** in the data
  - No Labels
Key Distinction in ML

**Regression**
- Bandgap
- Fatigue Strength
- Transformation Temp.

**Classification**
- Insulator
- Conductor
- Brittle Failure
- Ductile Failure
- Is Shape Memory Alloy
- Not Shape Memory Alloy
Model Types

- Linear Models
- Kernel Ridge
- Support Vector Machines
- Nearest Neighbors
- Gaussian Processes
  - **Decision Trees**
  - **Random Forests**
- Neural Networks

We’ll focus on just one type that is easier to understand conceptually and doesn’t require advanced math

For a more complete list of models
Decision Trees: Structure

- **Root Node**: Starting point which contains all data.
- **Decision Node**: Contains a single splitting criteria based on one feature.
- **Leaf Node**: Final Node in a branch where prediction is made.
- **Split**: A single division in the dataset based on the values of a single feature.
- **Branch**: Refers to a subset of data that is present after a series of splits.
Decision Trees: Inputs

Input Data

\( A_N > 10 \)  
Leaf

\( T_m < 1000 \)  
Leaf

Leaf

Features (descriptors)

<table>
<thead>
<tr>
<th>Index</th>
<th>( A_N )</th>
<th>( T_m ) (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fe</td>
<td>26</td>
<td>1800</td>
</tr>
<tr>
<td>Al</td>
<td>13</td>
<td>930</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>3800</td>
</tr>
</tbody>
</table>

Individual Data points
Decision Trees: Outputs

<table>
<thead>
<tr>
<th>Index</th>
<th>$A_N$</th>
<th>$T_m$ (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fe</td>
<td>26</td>
<td>1800</td>
</tr>
<tr>
<td>Al</td>
<td>13</td>
<td>930</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>3800</td>
</tr>
</tbody>
</table>

Features (descriptors):

- $A_N > 10$
- $T_m < 1000$

Individual Data points:

- $\text{Radius} = 67 \text{ pm}$
- $\text{Radius} = 118 \text{ pm}$
- $\text{Radius} = 156 \text{ pm}$
Summary

What is Machine Learning?
• Machine learning is a tool that finds patterns in large datasets that might be hard to discover otherwise

How can we use it for Materials Science?
• It can be included in existing materials science workflows to accelerate research, materials design, and materials discovery