Project 5: Making an AI

1. Table of Display

Topic	Method
What is ML	Summary
Intro to Scikit-Learn	Summary w/Code
Hyperparameters and Model Validation	?
Feature Engineering	?
DTrees and Random Forests	?
Face Detection Pipeline	Final Project

2. Table Of Contents

3. Background

3.1. What is Machine Learning?

Machine Learning (ML) is the art of designing mathmatical models of data which can then be taught via tunable parameters. Because these algorithms are designed to asist in understanding data, there can be some debaste of whether ML could be considered a branch of Aritifical Intelegence (AI) anymore.

Because ML works with big data which vary greatly in both size, complexity there must also be various methods of analizing this data. The two general categories of ML are **Supervised** and **Unsupervised** learning. We will explorw these methods futher in this paper.

3.1.1. Supervised learning

Supervised learning takes data and labels associated with the data to model their relationship. This ML model is used to apply labels to novel data. Supervised learning is commonly subdivided into **classification** and **regression**.

3.1.1.1. Classification

Classification models use descrite categories such as a status. This type of model may be used to identify objects in an image or seek for when a part might be damaged within a piece of machinery.

This model will require the devleoper to fist make a labeled dataset. Then you must design model inputs and the general assumptions your can provide. After this you can provide a set of model paramters which can be adjusted by the model during the training stage. The end requlst is that when introduced to novel data the classical model will provide a predictive label.

3.1.1.2. Regression

Regression models place their categories in a continuous spectrums. This can be used to seek the pobablity of damage to an piece of machinery or seek more complex relationships such as genetics to regions.

Regression models could be treadted as the oposite of Demensionality Reduction models. They will extract a new unknown relationship and create a new dimesion of labels.

3.1.2. Unspervised Learning

Unsuprvised learning results from not having a labeled dataset. This type of learning is used to find relationships within a data set. Unspervised Learning can be further subdivided into **Clustering** and **Demensitionaly Reduction** models.

3.1.2.1. Clustering

Clustering models act simlarly to classification models exceept they are seeking the distinct groups with the dataset. This can be used to seekout groups within massive datasets that humans may never see.

One method to do clustersing is through the *k*-means fits model. This method finds the center of data clusters, it must be profivided with a value *k* or this value may be tunable. This model seeks the postion of centers that has the minimum distance between all points in the dataset.

3.1.2.2. Dimensionality Reduction

Demensionality Reduction as its name suggests is designed to simpfly a data set into the smalled scturcure possible. This type of model could be used to seek important parameters to watch in larger systems which may take years for a human to analyze. This method allows us to infer new strutures that may not be labels (or may not exist).

A demnstionality reduction model typcially will remove one or more of the layers of data. This could be used to review a large dataset and place the data neatly into graphs that huamns and easily understand, interpret and apply.

3.2. Scikit-Learn

While developing a ML algrithem from scratch is possible it could be cumbersom and likly has already been done. Thus to prevent reinventing the wheel, you can use the modeule **Scikit-Learn**. Scikit-Learn contains many of the popular algorithems in one consitent API.

3.2.1. Data Represtientation

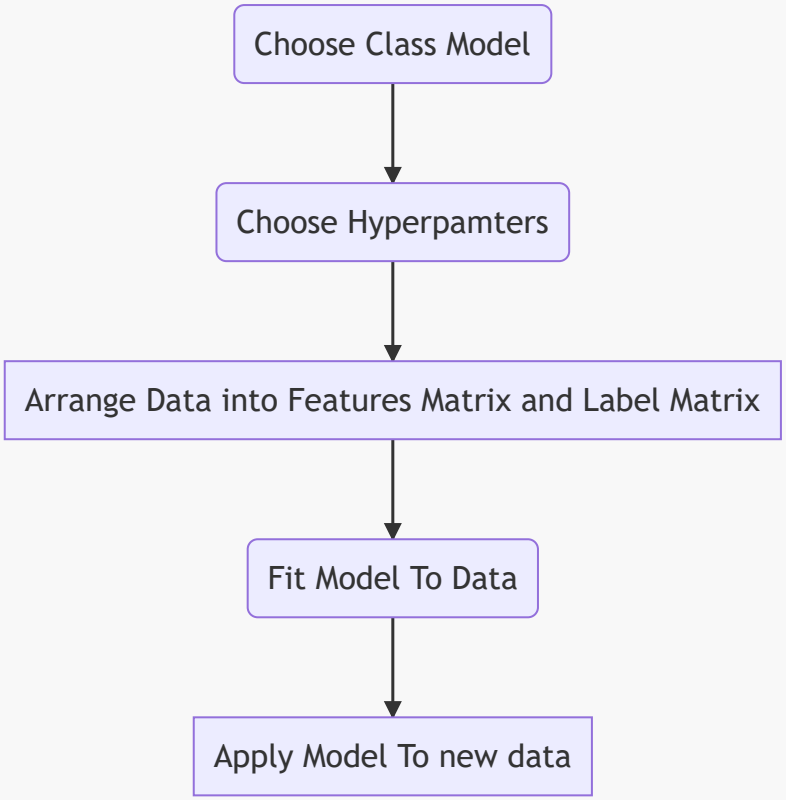
Scikit-Learn descibes data with a *features matrix* this mastrix can be stored in a NumPy array or Pandas DataFrame. The rows of the features matrix are often called sameles and the columns are considered features. Thus the features matrix is considered to have the shape `[n_samples, n_features]`.

The features matrix is then coupled with a *label array*. This array is typically is a Numpy array or Pandas Series of n_samples, but it can be two-dimesional with the shapre [n_smaple, n_targets]. The label array could be considered the dependent variable the model is attempting to predict.

3.2.2. Essentials of Scikit

Scikit is designed to a *consitant* API used for *inspection* of data. It also procvides a simplistic *composition* through the simplification of complex ML algorithms into their foundational parts. It also uses common data structures and provides excellent user-oriented defaults.

This results in the following workflow for algorithm design.



3.2.2.1. Using Scikit-Learn

Each model class is stored as a class within Scikit-Learn and are typically imported as a single object. For instance to import the *LinearRegression* modle you would use

```
from sklearn.linear_model import LinearRegression
```

3.3. Model Validation

After selecting a model and selecting its hyperparameters you must then *validate* the model. The two methods of model validation we will discus are **holdout sets** and **cross validation**. Once you have validated a model you must then select the best model by balancing the bias and variance of that model for the data you are using.

First let us setup a model to validate:

```

```

Model Selection

Hyperparmeters

Holdout Sets

3.5. Feature Engineering