Prof. Swapnil More

Introduction to Machine Learning

Data Science, Artificial Intelligence and Machine Learning

- Data Science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.
- Data Science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems to perform tasks that ordinarily require human intelligence.

Why learn and what is learning

- Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
- **Machine learning focuses on the development of computer programs** that can access data and use it learn for themselves.

What is Machine Learning ?

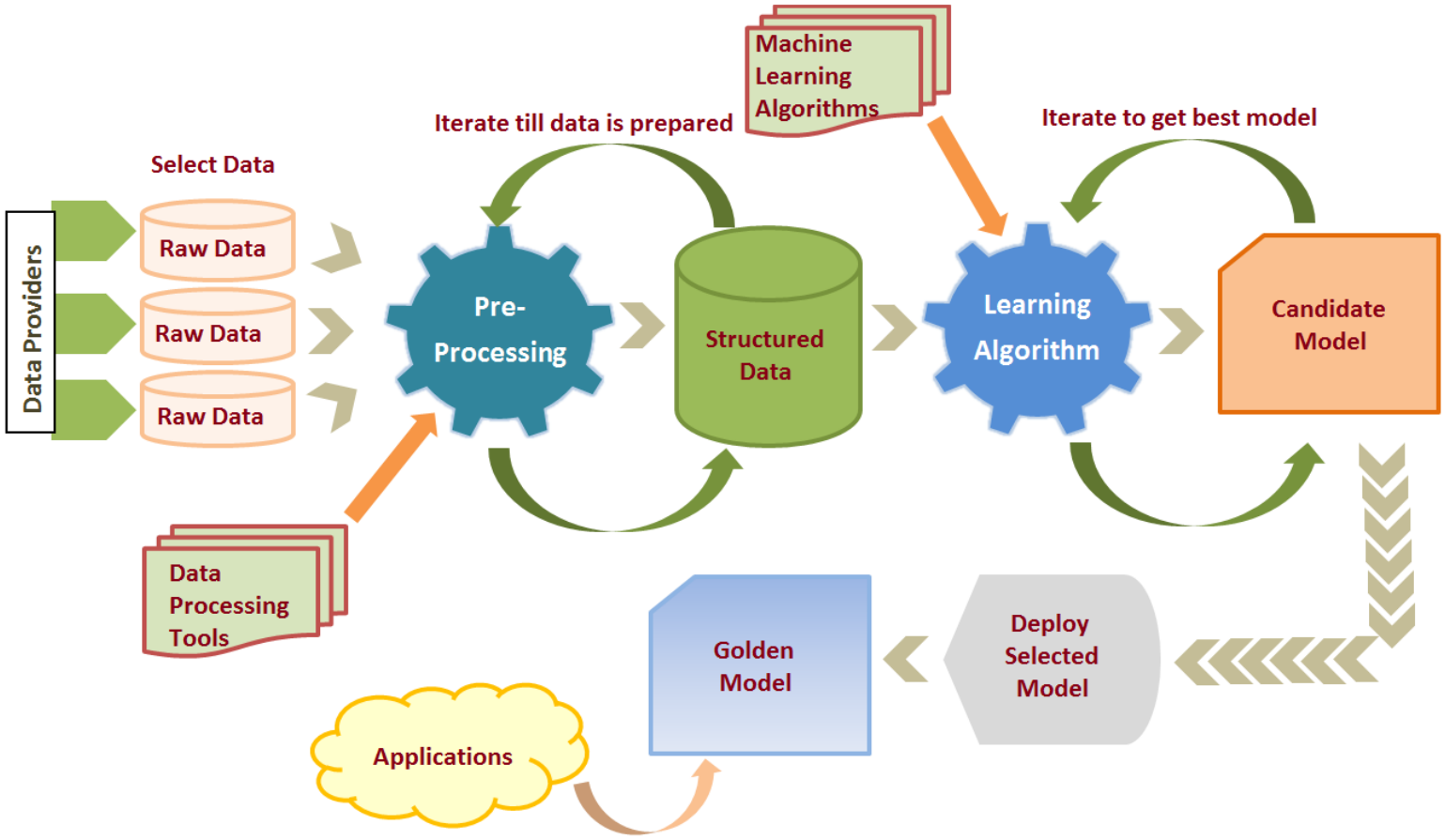
- **Machine learning** (**ML**) is the study of computer algorithms that improve automatically through experience.
- Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so.

Traditional Programming Vs. Machine Learning

Traditional programming is a manual process—meaning a person (programmer) creates the program. But without anyone programming the logic, one has to manually formulate or code rules.

In machine learning, on the other hand, the algorithm automatically formulates the rules from the data.

Machine Learning Process



Machine Learning Process

1 - Data Collection

- The quantity & quality of your data dictate how accurate our model is
- The outcome of this step is generally a representation of data (Guo simplifies to specifying a table) which we will use for training
- Using pre-collected data, by way of datasets from Kaggle, UCI, etc., still fits into this step

2 - Data Preparation

- Wrangle data and prepare it for training
- Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)
- Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data
- Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis
- Split into training and evaluation sets

Machine Learning Process

3 - Choose a Model

- Different algorithms are for different tasks; choose the right one

4 - Train the Model

- The goal of training is to answer a question or make a prediction correctly as often as possible
- Linear regression example: algorithm would need to learn values for m (or W) and b (x is input, y is output)
- Each iteration of process is a training step

5 - Evaluate the Model

- Uses some metric or combination of metrics to "measure" objective performance of model
- Test the model against previously unseen data
- This unseen data is meant to be somewhat representative of model performance in the real world, but still helps tune the model (as opposed to test data, which does not)
- Good train/eval split? 80/20, 70/30, or similar, depending on domain, data availability, dataset particulars, etc.

Machine Learning Process

6 - Parameter Tuning

- This step refers to *hyper parameter* tuning, which is an "art form" as opposed to a science
- Tune model parameters for improved performance
- Simple model hyper parameters may include: number of training steps, learning rate, initialization values and distribution, etc.

7 - Make Predictions

- Using further (test set) data which have, until this point, been withheld from the model (and for which class labels are known), are used to test the model; a better approximation of how the model will perform in the real world

Key Elements of Machine Learning

- There are tens of thousands of machine learning algorithms and hundreds of new algorithms are developed every year.
- Every machine learning algorithm has three components:
- **Representation**: how to represent knowledge. Examples include decision trees, sets of rules, instances, graphical models, neural networks, support vector machines, model ensembles and others.
- **Evaluation**: the way to evaluate candidate programs (hypotheses). Examples include accuracy, prediction and recall, squared error, likelihood, posterior probability, cost, margin, entropy k-L divergence and others.
- **Optimization**: the way candidate programs are generated known as the search process. For example combinatorial optimization, convex optimization, constrained optimization.
- All machine learning algorithms are combinations of these three components. A framework for understanding all algorithms.

Dimensionality Reduction (Feature Reduction)

- The number of input variables or features for a dataset is referred to as its dimensionality.
- Dimensionality reduction refers to techniques that reduce the number of input variables in a dataset.
- More input features often make a predictive modeling task more challenging to model, more generally referred to as the curse of dimensionality.
- High-dimensionality statistics and dimensionality reduction techniques are often used for data visualization. Nevertheless these techniques can be used in applied machine learning to simplify a classification or regression dataset in order to better fit a predictive model.

Do you Know ?

- Descriptive and Inferential Statistics:
- Probability,
- Distribution,
- Distance Measures (Euclidean and Manhattan),
- Correlation and Regression,
- Hypothesis Testing.

DATASET

- Creating our own dataset
- Importing the dataset
- Handling Missing Data