**Definitions**

* Artificial Intelligence is a program that can act, sense, reason, act, and adapt (mimic cognitive functions that we humans associate with other human minds such as learning and problem solving).
* Machine Learning is a subset of AI whose performance improves as it is exposed to more data over time.
* Finally, Deep Learning is a subset of Machine Learning in which the performance of multi-layered neural networks, that learn from vast amounts of data, improves as it is exposed to more data over time.

**Diving Deeper: Machine Learning**

* Machine learning systems are not explicitly programmed, but rather learn patterns as they are exposed to more data over time. You have a problem P, and an algorithm that you build to solve P. You use a performance measure to judge whether or not the algorithm does a poor or a good job of solving P as it is exposed to more data (model becomes more experienced).
* One important limitation of Machine Learning can be seen in classification algorithms that use pixels to classify images. If each image has 256 by 256 pixels, and you have thousands of images, then you are working with a huge number of features.
* But a more important issue is that you lose the information by considering one pixel and its surrounding pixels. Human beings don’t examine one pixel at a time. Rather we look at a bunch of pixels at the same time to classify animals.

**Diving Deeper: Deep Learning**

* Neural Network receives the pixels of an image, learns how to extract these features that are meaningful from the image by combining them together in different combinations.

**History of AI, Machine Learning, and Deep Learning**

* Expert Systems, theory behind Neural Networks, AI Winter because of the failure to leverage the promises of expert systems and Neural Networks. Machine Learning was successful in speech recognition, and Google algos. Finally, Deep Learning overcame began to outperform machine learning problems.
* 1950: Alan Turing developed the Turing test.
* 1956: AI was accepted as a field at the Dartmouth Conference.
* 1957: Frank Rosenblatt invented the perceptron algorithm.
* 1959: Arthur Samuel published an algorithm for a checkers program using Machine Learning.
* Expert systems mimic human experts that run on powerful computers and follow programmed rules.
* 1986: Backpropagation algorithm is able to train multi-layer perceptron’s.
* 1990s – 2000s: Deep Blue chess system beating the world chess champion Garry Kasparov, Google’s search engine launched using AI technology.

**Modern AI**

Two spaces where we see drastic growth are computer vision (self-driving cars, diagnosing illnesses better than experts) and NLP (sentiment analysis, better language translation, clustering articles).

**Machine Learning Workflow**

Problem Statement: What are you trying to solve?

Data Collection: What data you need to solve it? Where do you get it from?

Data Exploration and Pre-processing: Clean the data so you model can use it.

Modeling: Build the model.

Validation: Test the models on hold-out sets.

Decision-making and deployment: Communicate to stakeholders or put into production.

**Machine Learning Vocabulary**

Target value is the value you are trying to predict. Such as species, prices, or classes.

Features are what is used for prediction.

Example is a single data point within your data (a single row).

Label is the target value for a single point such as versicolor, or $560899.

**Exploratory Data Analysis**

Approach to analyzing data sets to summarize main characteristics, and it gives us an initial feel for the data. Doing so will help you figure out if you need more data, or does your current data require more data cleaning, or does your data make sense.

* There are different approaches for conducting EDA.
* Different EDA Techniques.
  + Summary Statistics such as median, min, max, correlations, etc. (Pandas)
  + Visualizations such as histograms, scatterplots, boxplots (Matplotlib, Seaborn).
* Sampling to take a peek into our data.
  + Random samples, for larger datasets, make computation easier (sometimes you want to stratify so that your sample data’s proportion is representative of the original data or population).
  + A very common reason for sampling is splitting the data into training and testing when building and evaluating an ML model.
  + Other sampling technique is either sampling with replacement, or without replacement.
* Producing EDA Visualizations.
  + Matplotlib and Seaborn are very good libraries.
  + EDA – Part 2 Video, and the Lab.

**Using Residuals to find outliers**

* This approach assumes that you have a model as residuals are the differences between the actual and predicted values of the outcome variable.
* However, a major problem with ordinary residuals is that their magnitude depends on the units of measurement. So, it is difficult to detect unusual y values.
* You can eliminate the units of measurement by dividing the residuals by an estimate of their standard deviation.

**Estimation and Inference – Part 1**

*Statistical estimation and inference*

* An estimate is just going to give us an estimate of a certain parameter such as an average.
* Machine learning and statistical inference are similar because we are using our sample data to infer qualities of the actual population distribution and models that would have generated that data.
* Machine learning applications that focus on understanding parameters and individuals’ effects involve more tools from statistical inference (statistical significance).
* On the other hand, some models just focus on making predictions rather than worrying too much on feature effects.
  + For example, say you want to predict if a customer will leave the company or not (churn) using features such as the length of time as a customer, type and amount purchased, avg customer satisfaction survey scores, and other customer characteristics.
  + In this classification problem, you want a model that estimates the probability that a customer will leave.
  + You can estimate the impact of each factor in predicting churn and determining whether the measured impact is statistically significant.

*Parametric v. Non-Parametric Model*

* A Parametric model is a statistical model that they make strict assumptions about the distribution from which the data is pulled. Will also have a finite number of parameters.
* Non-Parametric models are model that don’t make those assumptions.
  + Customer lifetime value is an estimate of the customer’s value to the company.
  + To estimate lifetime value, we make assumptions about the data.
  + We can assume that the relationship between the response variable and features is linear.
  + The most common way of estimating parameters in a parametric model is a MLE.

*Commonly Used Distributions*

* Uniform distribution is where there is an equal chance that you’ll get any value within our range.
* In the normal distribution, the idea is that the most likely value is going to be those values that are closest to the means and those that are further out are rare. It has a given mean and a standard deviation that determines how spread out is the shape of the distribution.
* Normal distribution is popular due to the Central limit theorem, the distribution of your sample means, or the sampling distribution will be normal if your sample size is large.
* Exponential curve, you’ll have most values to the left, and it will be often used to say what is going to be the amount of time before the next event.
* Poisson distribution: number of events that happened during a certain amount of time. Lambda is both the average value and the variance.
* How many people are going to watch a video in the next ten minutes? If Lambda was 1, we say that most of the time there’s only one person that watches every 10 minutes and it’s tight around that one. But if it’s something like 10 people who watch it every 10 minutes, then you would probably have more of a spread of your standard deviation.

**Hypothesis Testing**

* A hypothesis is a statement about a population parameter. Create a null and an alternative hypothesis.
* Suppose you have two coins:
  + Coin 1 has a 70% probability of coming up heads.
  + Coin 2 has a 50% probability of coming up heads.
* You can do an experiment where pick one of two coins randomly, toss it 10 times, and record the number of heads. Given the heads you see, you can calculate the likelihood ratio.
  + Pr (Tossing 3 heads | Coin 1) = 0.117
  + Pr (Tossing 3 heads | Coin 2) = 0.009.
  + Likelihood ratio = P1(3)/P2(3) = 0.117/0.009 = 13. Thus, coin 1 was 13 times more likely to give us the output (3 heads) than coin 2.

**Type I and Type II Error**

* Table

  Description automatically generated
  + Null Hypothesis: We’re tossing a fair coin.
  + Alternative hypothesis: We are tossing an unfair coin.
  + Type I error will be that we’re working with a fair coin, but we reject the null hypothesis.
  + Type II error will be that we are tossing an unfair coin, but we think that we were tossing a fair coin.
* The power of a test is 1-Pr (Type II Error). This is the probability that we correctly reject the null hypothesis.
  + Going back to our customer churn example. Suppose our null hypothesis is that customers churn is just due to chance, and our alternative hypothesis is that customer that had been with the company for more than two years will not churn.
  + Thus, a type I error will be that the effect is due to chance, but we determine that customer who have been with company for more than two years are less likely to churn.
  + A type II error will be that there is a significant effect, but we ascribe the effect to chance.
  + The likelihood ratio is called a test statistic we use it to decide whether to accept/reject null.

**Significance Level and P-Values**

* To get the rejection region, we calculate the test statistic, and choose the significant level beforehand. It is a probability threshold below which the null hypothesis will be rejected.
* Common alphas are often 0.01 or 0.05.
* The p-value is the smallest significance at which the null hypothesis will be rejected.