Lecture 1: Introduction to Generative Models

Lakshita Mohanty (190454) | Harsh Kumar (190360)

January 5, 2022

1 Introduction

Before we begin our exploration, we must analyse and understand the answers to some basic questions like -

- What is **Intelligence** ?
- What is **Artificial Intelligence** ?
- What is **Machine Learning** ?

These fundamental questions will keep popping up while we go through the topics. But we will more formally base our discussions on the last question.

2 What is Machine Learning ?

Machine Learning is a field of study wherein intelligent models are designed which implement algorithms to solve a problem through the use of data.
Some examples of ML models are -

- Recommendation Systems
- Classification Systems like Facial Recognition or Image Classification
- Regression Systems like Price Prediction Models

3 Representation of Data

For simplicity, we represent the input matrix (containing all the data points) to a model as **X** and the output (label, number, etc.) as **Y**. We represent a model using a **joint probability density function (pdf)** of **X** and **Y**. So, the most general form of representing the probabilistic distribution of the whole data, is **P(X,Y)**.

4 Discriminative Model v/s Generative Model

Implementing the Chain Rule on the joint pdf of **X** and **Y**, we get -

$$\mathrm{P(X,Y) = P(X|Y)\cdot\ P(Y) = P(Y|X)\cdot\ P(X)}$$

These two definitions of **P(X,Y)** gives rise to two different types of models - **Discriminative Models** and **Generative Models**.

4.1 Discriminative Model

P(Y|X) - viz. probability of Y given X; given an input matrix, it estimates the label/output vector.

$$\text{Y*} = \underset{\text{Y}}{\text{argmax}}\ \text{P(Y|X)}$$

Here, Y* is the actual label. A Discriminative Model can essentially differentiate between labels but it **cannot** generate new data. Examples of Discriminative Models can be - Support Vector Machines (SVMs), Logistic Regression Models, Neural Networks, etc.

4.2 Generative Model

P(X|Y) - viz. probability of X given Y; given the output vector (or any data point), it estimates the input matrix (or generates new data).

$$\text{X'} \sim \text{P(X|Y)}$$

By sampling, we mean to obtain some X' from the distribution given by P(X|Y). A Generative Model essentially generates more data from a given data point. Examples of Generative Models can be - Generative Adversarial Networks (GANs), Gaussian Mixture Modelling (GMM), etc.

5 How to make models?

The above discussed models can be made using two different approaches -

- Parametric
- Non-Parametric

5.1 Parametric Models

A parametric model, as the name suggests, has parameters which majorly define the model . Such models are limited by form. However, they can scale efficiently with the size of the data. Parametric models have a finite dimensional space. Below are listed some parametric models along with the parameters involved -

- Neural Networks - Weights (and Biases)
- Logistic Regression Model - Coefficients
- HMM (Hidden Markov Model) - $\begin{cases} \text{transition - exponential} \\ \text{emmision - GMM} \end{cases}$
- GMM (Gaussian Mixture Modelling) - μ_i (mean), σ_i^2 (variances), ω_i (weights)

5.2 Non-Parametric Models

In contrast to parametric models, non-parametric models do not have any parameters. They aren't limited by form but they cannot scale efficiently with the size of the data. Such models require a huge amount of memory to function and have massive runtime complexities as well. Non-parametric models have a infinite dimensional space. Examples of non parametric models are -

- K Nearest Neighbours
- Decision Trees
- SVMs (Support Vector Machines)
- Histogram

6 Generative Modelling (Parametric Models)

Let us have a probability distribution $\mathbf{P(X)}$ and the model has parameters θ. Then the model is represented as $\mathbf{P}_\theta\mathbf{(X)}$ or $\mathbf{P(X;}\ \theta\mathbf{)}$. Given the model, we need to learn the parameters θ in the following way -

$$\theta* = \underset{\theta}{\text{argmin}}\ (\text{P}_\theta,\ \text{P}_{\text{data}})$$

where, $\mathbf{P}_\theta$ is our parametric model. Let's see what each part of the above expression mean.

6.1 $\mathbf{P}_{\mathbf{data}}$

$\mathbf{P}_{\mathbf{data}}$ could be the sample of data available or the true distribution of our model (which is known somehow).

- Often samples are available for images, audio, etc. For instance, MNIST data, IRIS data, IMAGENET. These samples can be used to define an estimate of the data distribution.
- P* (true pdf) is also available/known in some cases. This case is popular in fields like statistical mechanics, where we use Boltzmann Distribution.

6.2 Distance Metric (d)

It is the notion of distance defined between 2 probability distributions. Examples of the same are - KL divergence, Earth Mover's Distance (EMD), and Cross Entropy. It can also be defined as the distance between P_θ and the samples (data). This is quantified using the concept of negative likelihood.

$$\text{Negative Likelihood} = \text{-}\ \text{P(X;}\ \theta)|_{\text{X=samples}}$$

6.3 Learning Algorithms

Various learning algorithms can be used for training, such as

- Gradient Descent
- EM (Expectation Maximization) Algorithm
- Viterbi Algorithm
- MLE (Maximum Likelihood Estimation), MAP (Maximum A Posteriori)

6.4 The Learnt Model

The $\theta*$ obtained may be a single fixed value (scalr/vector) or it could be a distribution p(θ). The later are termed as Bayesian Models where even the parameters have a distribution.

7 What can we do with the model?

After we estimate our model using the methodology described above, what can we do with it? Let's see some uses of such probabilistic models -

- Prediction : As discussed earlier, we can use the probability distribution $\mathbf{P(Y|X)}$ to make predictions.

$$\text{Y*} = \underset{\text{Y}}{\text{argmax}}\ \text{P(X|Y;}\theta)\cdot\text{P(Y)}$$

- Sampling : The probability distribution $\mathbf{P(X|Y)}$ can be used to sample or generate more data **X'** from given data **X**.

$$X' \sim P(X|Y)$$

- Density Estimation : For instance, in a model dealing with image classification, we can estimate the probability density $\mathbf{P(X;}\ \theta\mathbf{)}$ of the images such that it is high for the right label and negligible for the wrong ones.
- Representation Learning : This helps us understand how our data is structured by representing it appropriately. For instance, unsupervised learning like clustering, etc., segments the data on the basis of a particular metric and the cluster label is used to represent the corresponding data.