CS 419M Introduction to Machine Learning **Spring 2021-22**

Lecture 1: Review of Probability and Linear Algebra

Lecturer: Abir De *Scribe: Ipsit Mantri*

1.1 Probability

The following topics are important for machine learning:

1. Conditional Independence or probabilities, Bayesian relationship
2. Introducing Random Variables: Bernoulli, Normal, Poisson
3. Expectation, Variance, Covariance
4. Central Limit Theorem, Law of large numbers

1.1.1 Why Probability is important for Machine Learning?

Suppose you have a set of **Data**. This can be

- A set of images $\xrightarrow{m_\theta}$ identify the objects
- A set of paragraphs $\xrightarrow{m_\theta}$ identify the topics

where m_θ denotes a model with parameters θ. The objective is to devise a model which learns to do a task. The following question can be posed:

Q. **What is the underlying thesis/concept behind the hope that it will be possible to train the model m_θ to get an accurate output?**
Ans. **The law of large numbers**

1.1.2 The Law of Large Numbers and Central Limit Theorem

Suppose that we have N samples $X_i \sim f(\cdot), \quad i = 1, \ldots, N$ drawn from an unknown distribution. Most of the times, the goal of machine learning is to estimate the distribution f or its properties. For example, in the task of generating images (GAN, VAE etc), the objective of the machine learning model is to identify the underlying distribution of the data samples. Being able to learn the underlying distribution is important even for tasks like housing price prediction, image classification etc because if a test sample is sampled from a different distribution is given to the model, the model will likely fail and give incorrect results.

Q. **What is the least that can be found about f using $X_1, \ldots, X_N$?**
Ans. Central Limit Theorem states that as long as N is large, the

$$\mathbb{E}_{X \sim f(\cdot)}[X] = \frac{\sum X_i}{N}$$

The Law of Large Numbers further states that

$$\mathbb{E}_{X \sim f(\cdot)}[X] \to \frac{\sum X_i}{N}$$

as $N \to \infty$. Moreover, the random variable $Z_N = \frac{\sum X_i}{N}$ follows the normal distribution $\mathcal{N}(\mathbb{E}[X], \sigma^2)$ with $\sigma^2 \propto \frac{1}{N}$ with

$$\lim_{N \to \infty} \mathbb{E}[Z_N] = \mathbb{E}[X]$$

Q. **What else do we need to completely characterize f?**
Ans. **We need the higher order moments!** Using these the moment generating function can be obtained.

Definition 1.1. *The moment generating function (MGF) $F(s)$ for a random variable X is defined as the Laplace Transform of the probability density function (PDF) $f(x)$. The inverse Laplace Transform of MGF will give the PDF.*

$$F(s) = \int_{-\infty}^{\infty} e^{-sx} f(x)\, dx$$

$$f(x) = \int_{-\infty}^{\infty} e^{sx} F(s)\, ds$$

Proposition 1.2. *Using only the moments, the MGF can be determined.*

Proof.

$$\begin{aligned} F(s) &= \int_{-\infty}^{\infty} e^{-sx} f(x)\, dx \\ &= \int_{-\infty}^{\infty} \sum_{k=0}^{\infty} \frac{(-s)^k x^k}{k!} f(x)\, dx \\ &= \sum_{k=0}^{\infty} \frac{(-s)^k}{k!} \mathbb{E}[X^k] \end{aligned}$$

Note: The discussion about the region of convergence of the Laplace transform is out of the scope of this course.

□

Proposition 1.3. *Using just the N samples $X_i \sim f(\cdot), \;\; i = 1, \ldots, N$, the PDF f can be estimated using the Law of Large Numbers.*

Proof. This can be proved as follows:

1. All moments $\mathbb{E}[X^k]$ can be estimated by invoking the Law of Large Numbers
2. The MGF can be found by invoking the Claim 1.2 above.
3. f can be estimated by taking the inverse Laplace Transform of the MGF

□

1.1.3 Practice Problems

1. A fair coin is tossed 5 times. Find $\mathbb{P}(\#H > \#T)$
   It is easy to see that as 5 is odd, there is not possibility of $\#H = \#T$. Hence $\mathbb{P}(\#H > \#T) = 1/2$
2. $X \sim f(\cdot), Y \sim g(\cdot)$. Then $Z = X + Y \sim ?$

$$\begin{aligned}
\mathbb{P}(Z \leq z) &= \mathbb{P}(X + Y \leq z) \\
&= \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{z-x} f(x)\, g(y)\, dx\, dy \\
&= \int\limits_{-\infty}^{\infty} f(x)\, G(z-x)\, dx
\end{aligned}$$

On differentiating, we get

$$f_Z(z) = \int\limits_{-\infty}^{\infty} f(x) g(z-x)\, dx = (f * g)(z)$$

1.2 Linear Algebra

Solve the following problems:

1. For two square matrices $A, B$of size $n \times n$ if $AB = BA$ for all B, then show that $A = cI_n$ for some $c \in \mathbb{R}$
   Pick B to be a diagonal matrix with pair-wise distinct elements. Then it can be shown that A is also a diagonal matrix. Now pick B to be a matrix with all ones, i.e. $B = [1]_{ij}$. As A is diagonal, $AB = BA$ implies all diagonal entries of A are equal i.e., $A = cI_n$ for some $c \in \mathbb{R}$

2. If $x^\top Ax = 0 \ \forall x \in \mathbb{R}^n$ then show that A is skew-symmetric.
   On differentiating the above equation we get

$$(A + A^\top)x = 0 \ \forall x \in \mathbb{R}^n$$

   This implies $A = -A^\top$

3. Show that $\text{rank}(AB) \leq \text{rank}(A)$
   Each column of AB can be viewed as a linear combination of columns of A. Hence, if the dimension of column space of A is r, the dimension of column space of AB cannot be more than r. In other words, $\text{rank}(AB) \leq \text{rank}(A)$

4. Suppose you have a uniform sampler which samples uniformly from $[0, 1]$. Propose an algorithm which uses this uniform sampler to generate samples from any given distribution.
   Suppose we have a uniform random variable $U \sim \text{Uniform}([0, 1])$. We need to find a function g such that the PDF of the random variable $g(U)$ will be same as that of the given distribution, say f. That is $g(U) \sim f$. Which is same as

$$\begin{aligned} \mathbb{P}(g(U) \leq x) &= \mathbb{P}(U \leq g^{-1}(x)) \\ \implies \int\limits_{-\infty}^{x} f(x)\,dx &= \int\limits_{-\infty}^{g^{-1}(x)} 1\,du \\ \implies F(x) &= g^{-1}(x) \\ \implies g &= F^{-1} \end{aligned}$$

1.3 Group Details and Individual Contribution