Deep Learning Book Notes

Chapter 3: Probability and Information Theory

Compiled By: Dakshit Agrawal

Discussion held on 10th August 2018.

Section 3.1: Why Probability?

- Three possible sources of uncertainty:
    - **Inherent stochasticity in the system being modeled**, e.g., most interpretations of quantum mechanics describe the dynamics of subatomic particles as being probabilistic. Another example may be of a number written as 4 and a 9 (occurs when the top loop is incomplete).
    - **Incomplete observability** occurs when all the factors that drive the behaviour of the deterministic system cannot be observed, e.g. Monty Hall problem.
    - **Incomplete modeling** occurs when we use a model that must discard some of the information we have observed, which results in uncertainty in the model's predictions.
- There are two kinds of probability:
    - **Frequentist probability** is related to the rates at which the events occur, ie if an event 'A' has probability 'p' of occuring, it means that a proportion 'p' of infinite many repetitions of the experiment will result in event 'A'.
    - **Bayesian probability** is the usage of probability to represent a degree of belief. A '1' represents absolute certainty of the event, whereas '0' indicates absolute certainty of the event not happening.
- To satisfy several properties that are expected for uncertainty to have by common sense reasoning, it is necessary to treat bayesian probabilities as behaving exactly the same as frequentist probability.
- **Probability** can be seen as the extension of logic to deal with uncertainty.

- **Logic** provides a set of formal rules for determining what propositions are implied to be true or false given the assumption that some other set of propositions is true or false.
- **Probability theory** provides a set of formal rules for determining the likelihood of a proposition being true given the likelihood of other propositions.

Section 3.2: Random Variables

- A variable that can take on different values randomly.
- It is a description of the states that are possible, and must be coupled with a probability distribution that specifies how likely each state is.
- The variable itself is denoted with a lowercase letter in plain typeface. The values it takes are denoted with lowercase script letters, e.g. x_1 and x_2 are possible values of random variable x.
- This variable may be discrete (finite or countably infinite number of states) or continuous (a real value).

Section 3.3: Probability Distributions

- A description of how likely a random variable or set of random variables is to take on each of its possible states.

Section 3.3.1: Discrete Variables and Probability Mass Functions

- Probability distribution over discrete variables is described using a **Probability Mass Function (PMF)**. It is usually denoted by '*P*'.
- PMF that act on many random variables at the same time are known as joint probability distribution. $P(\mathrm{x} = x, \mathrm{y} = y)$ denotes the probability that $\mathrm{x} = x$ and $\mathrm{y} = y$ simultaneously.
- A PMF must satisfy the following properties:
  - Domain of *P* must be the set of all possible states of x.
  - $\forall x \in \mathrm{x}, 0 \leq \mathrm{P}(x) \leq 1$.
  - $\Sigma_{x \in \mathrm{x}} P(x) = 1$. This property is known as being **normalized**. Without this property, we could obtain probabilities greater than one.

Section 3.3.2: Continuous Variables and Probability Density Functions

- Probability distribution over continuous variables is described by a **Probability Distribution Function (PDF)**. It is denoted by 'p'.
- A PDF must satisfy the following properties:
  - The domain of p must be the set of all possible states of x.
  - $\forall x \in \mathrm{x}$, $0 \le P(x)$. Note that it is not required to be less than 1.
  - $\int p(x)\, dx = 1$.
- The probability density function $p(x)$ does not give the probability of a state directly. Instead the probability of landing inside an infinitesimal region with volume δx is given by $p(x)\delta x$. This can be integrated to find the actual probability mass of a set of points.
- Uniform distribution function is denoted by $u(x; a,b)$. Within $[a, b]$, $u(x; a, b) = 1/(\mathrm{b} - \mathrm{a})$. $u(x; a, b) = 0$ for all $x \notin [a, b]$.

Section 3.4: Marginal Probability

- The probability distribution over a subset of random variables given is known as the **marginal probability distribution**.
- For discrete random variables: $\forall x \in \mathrm{x}$, $P(\mathrm{x} = x) = \Sigma_y P(\mathrm{x}= x, \mathrm{y} = y)$.
- For continuous random variables: $p(x) = \int p(x, y)\, dy$.

Section 3.5: Conditional Probability

- The probability of an event, given that some other event has happened is given by **conditional probability**. The formula is as follows:

$$P(y = y \mid x = x) = \frac{P(y = y , x = x)}{P(x = x)}$$

- Conditional probability is defined only when $P(\mathrm{x} = x) > 0$, i.e. we cannot compute the conditional probability conditioned on an event that never happens.

Section 3.6: The Chain Rule of Conditional Probabilities

- Any joint probability distribution over many random variables may be decomposed into conditional distributions over only one variable, known as the **chain rule**:

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \prod_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)})$$

Section 3.7: Independence and Conditional Independence

- Two random variables x and y are **independent** if their probability distribution can be expressed as a product of two factors, one involving only x and one involving only y:

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y)$$

- Two random variables x and y are conditionally independent given a random variable z if the conditional probability distribution over x and y factorizes in this way for every value of z:

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z}, p(\mathrm{x} = x, \mathrm{y} = y \mid \mathrm{z} = z) = p(\mathrm{x} = x \mid \mathrm{z} = z)p(\mathrm{y} = y \mid \mathrm{z} = z)$$

- An example of conditional independence is the case of taking tests in an examination hall. If two students plan to cheat, their probability of passing the test depends on the presence of an invigilator. If the invigilator is not present, then they can cheat, and their probabilities of passing are dependent on each other. However, if the invigilator is present, they cannot cheat, and their probabilities of passing are not dependent of each other.

Section 3.8: Expectation, Variance and Covariance

- The **expectation**, or **expected value**, of some function *f(x)* with respect to a probability distribution *P(x)* is the average, or mean value, that *f* takes on when *x* is drawn from *P*. For discrete variables, this can be computed as follows:

$$\mathbb{E}_{\mathrm{x} \sim P}[f(x)] = \Sigma_x P(x)f(x)$$

  For continuous variables, is is computed as follows:

$$\mathbb{E}_{x \sim P}[f(x)] = \int p(x)\, f(x)\, dx$$

- Expectation follows linear property.

$$\mathbb{E}_x\ [\alpha f(x) + \beta g(x)] = \alpha\mathbb{E}_x\ [f(x)] + \beta\mathbb{E}_x\ [g(x)]$$

- **Variance** gives a measure of how much the values of a function of a random variable x vary as we sample different values of *x* from its probability distribution. If it is low, the values of *f(x)* cluster near their expected value.

$$\mathrm{Var}(f(x)) = \mathbb{E}[\ (\ f(x) - \mathbb{E}[f(x)]\ )^2\ ]$$

- The square root of the variance is known as the **standard deviation**.
- The **covariance** gives a sense of how much two values are **linearly** related to each other

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}[\ (f(x) - \mathbb{E}[f(x)])\ (g(y) - \mathbb{E}[g(y)])\ ]$$

- If covariance is positive, then both variables tend to take up values on the same side of the mean. If covariance is negative, then both variables tend to take up values on the opposite side of the mean.
- The magnitude of covariance does not give a good idea of how related the two values are. For this, we use **correlation**, which is Covariance divided by the standard deviation of the two values or distributions. In other words, correlation is the covariance of normalized distributions.
- If two variables are independent to each other, then covariance is 0. However, the converse is not true. Covariance equal to 0 ensures only **linear independence**. E.g., *f(x)* = x, and *g(y)* = y = x^2 have covariance 0, but are strongly dependent on each other.
- For a more detailed explanation of covariance, its properties, correlation and examples, refer to the following video:

  https://www.youtube.com/watch?v=IujCYxtpszU&index=21&list=PL2SOU6wwxB0uwwH80KTQ6ht66KWxbzTIo

Section 3.9: Common Probability Distributions

Section 3.9.1: Bernoulli Distribution

Short description given in book.

Section 3.9.2: Multinoulli Distribution

Short description given in book.

Section 3.9.3: Gaussian Distribution

- Two parameters are sufficient to describe a Gaussian Distribution
    - The parameter $\mu \in \mathbb{R}$ gives the coordinate of the central peak, as well as the mean of the distribution: $\mathbb{E}[x] = \mu$.
    - $\sigma \in (0, \infty)$ gives the standard deviation of the distribution. The variance is given by σ^2.
- In the absence of prior knowledge about what form a distribution over the real numbers should take, the normal distribution is a good default choice for two major reasons.
    - Many distributions we wish to model are close to being normal distributions. The **central limit theorem** shows that the sum of many independent random variables is approximately normally distributed. For more information, refer to the following video by Khan Academy: https://www.khanacademy.org/math/ap-statistics/sampling-distribution-ap/sampling-distribution-mean/v/sampling-distribution-of-the-sample-mean
    - Out of all possible probability distributions with the same variance, the normal distribution encodes the maximum amount of uncertainty over the real numbers. Full justification is given in section 19.4.2 (complicated maths).

Section 3.9.4: Exponential and Laplace Distribution

Short description given in book.

Section 3.9.5: The Dirac Distribution and Empirical Distribution

Short description given in book. Last paragraph last line is explained in section 5.5 (we will take up the topic there).

Section 3.9.6: Mixture of Distributions

- Sometimes probability distributions may be combinations of various probability distributions:

  $$P(x) = \Sigma_i P(c = i)P(x \mid c = i)$$

  where *P(c)* is the multinoulli distribution over component identities.

- Empirical distribution over real-valued variables is a mixture distribution with one Dirac component for each training example.
- A **latent variable** is a random variable that we cannot observe directly. They may be related to x through the joint distribution like *P(x, c) = P(x | c)P(c)*. The distribution *P(c)* over the latent variable and the distribution *P(x |c)* relating the latent variables to the visible variables determines the shape of the distribution *P(x)*, even though it is possible to describe *P(x)* without reference to the latent variable.
- An example of latent variables: Suppose the experiment is throwing 2 dice and calculating the probability of a particular sum. The distribution of the sum is across 11 values of a random variable x. However, we can also relate this visible random variable to a latent variable c, that can take the values 1-6, corresponding to the possible outputs of a single dice.
- **Gaussian Mixture Models (GMM)** are a mixture of Gaussian distributions that can be used to model a probabilistic distribution. The fitting of a GMM to a particular distribution can be done with the help of Expectation-Maximization Algorithm. For more info, refer to the following site:

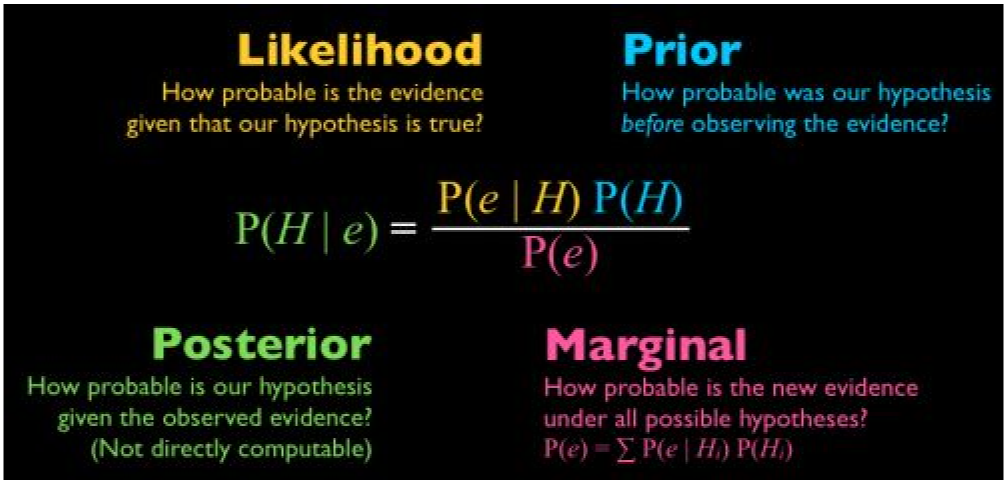
  http://www.aishack.in/tutorials/expectation-maximization-gaussian-mixture-model-mixtures/

Section 3.10: Useful Properties of Common Functions

- Two functions are explained in the book, **logistic sigmoid** and **softplus** functions.
- Formula for logistic sigmoid is given by Eq. 3.30. Its plot is shown in Figure 3.3.
- Formula for softplus is given by Eq. 3.31. Its plot is shown in Figure 3.4.
- Some properties of the two functions are mentioned in Eq. 3.33 - Eq. 3.41.
- The values before passing through a non-linearity are generally called **logits**.

Section 3.11: Bayes' Rule

- If we know *P(*y | x*)* and *P*(x), we can calculate *P*(x | y) using **Bayes' rule**:



Likelihood

How probable is the evidence given that our hypothesis is true?

Prior

How probable was our hypothesis *before* observing the evidence?

$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

Posterior

How probable is our hypothesis given the observed evidence? (Not directly computable)

Marginal

How probable is the new evidence under all possible hypotheses?

$P(e) = \sum P(e \mid H_i)\, P(H_i)$

Section 3.12: Technical Details of Continuous Variables

- A proper formal understanding of continuous random variables and probability density functions requires developing probability theory in terms of a branch of mathematics known as measure theory.
- The probability of a continuous vector lying in some set $\mathbb{S}$ is given by the integral of *p(x)* over the set $\mathbb{S}$. It is possible to construct two sets $\mathbb{S}_1$ and $\mathbb{S}_2$ such that $p(\mathbf{x} \in \mathbb{S}_1) + p(\mathbf{x} \in \mathbb{S}_2) > 1$ but $\mathbb{S}_1 \cap \mathbb{S}_2 = \phi$. These are constructed using heavy use of infinite precision of real numbers, e.g. fractal-shaped sets, and are beyond the scope of undergraduates.

- A set of points is said to have **measure zero**. It intuitively means that a set of measure zero occupies zero volume in space.
- A property that holds throughout all space except for on a set of measure zero is told to hold **almost everywhere**.
- For continuous variables that are deterministic functions of one another, e.g. random variables, x and y, such that y = *g(*x), where *g* is an invertible, continuous, differentiable transformation, the following holds:

$$p_x(x) = p_y(g(x))\ |\partial g(x) / \partial x|$$

Section 3.13: Information Theory

- Information theory is a branch of applied mathematics that evolves around quantifying how much information is present in a signal.
- A measure to quantize information should have the following properties:
  - Likely events should have low information content
  - Less likely events should have higher information content
  - Independent information should have additive information
- **Self-information** of an event x = *x* is defined to be

$$I\ (x) = -\log P(x)$$

  Here *P(x)* is PMF in the case of discrete random variables, and PDF in the case of continuous random variables. For continuous random variables, an event with unit density still has zero information, even though that event is not guaranteed to occur.

- If log is natural logarithm, the units of self-information are nats. If log is base 2, then the units are called bits or shannons.
- **Shannon Entropy** of a distribution is the expected amount of information in an event drawn from that distribution. It can also be interpreted as the amount of uncertainty in an entire probability distribution. For continuous random variables, it is called **differential entropy**.

$$H(x) = \mathbb{E}_{\mathrm{x} \sim P}[I(x)] = -\mathbb{E}_{\mathrm{x} \sim P}[\log P(x)]$$

- To measure the difference between two distributions, the **Kullback-Leibler (KL) divergence** is generally used:

  $$D_{KL}(P \parallel Q) = \mathbb{E}_{x \sim P} [\log \frac{P(x)}{Q(x)}] = \mathbb{E}_{x \sim P} [\log P(x) - \log Q(x)]$$

  It is the extra amount of information needed to send a message containing symbols drawn from probability distribution *P*, when we use a code that was designed to minimize the length of messages drawn from probability distribution *Q*.

- The KL Divergence is always non-negative and asymmetric. For proof, refer to the following video: https://www.coursera.org/lecture/bayesian-methods-in-machine-learning/jensens-inequality-kullback-leibler-divergence-E8CFE
- For explanation of Figure 3.6, refer to the following blog: https://wiseodd.github.io/techblog/2016/12/21/forward-reverse-kl/
- For further read on KL Divergence, refer to the following blog: https://medium.com/@cotra.marko/making-sense-of-the-kullback-leibler-kl-divergence-b0d57ee10e0a
- The **cross entropy** $H(P, Q) = H(P) + D_{KL}(P \parallel Q)$ is the expected amount of information that is sent in a message containing symbols drawn from probability distribution *Q*, when we use a code that was designed to minimize the length of messages drawn from probability distribution *Q*.

  $$H(P, Q) = - \mathbb{E}_{x \sim P}[\log Q(x)]$$

Section 3.14: Structured Probabilistic Models

- Instead of using a single function to represent a probability distribution, we can split a probability distribution into many factors that we multiply together.
- These factorizations greatly reduce the number of parameters needed to describe the distribution.

- The factorization of a probability distribution can be represented with a graph, called a **structured probabilistic model** or **graphical model**, which are of two types, **directed** and **undirected**.
- **Directed models** represent factorizations into conditional probability distributions. It contains one factor for every random variable x_i in the distribution, and that factor consists of the conditional distribution over x_i given the parents of x_i, denoted by $Pa_G(x_i)$:

$$p(\mathrm{x}) = \prod_i p(\mathrm{x}_i \,|\, Pa_G(x_i))$$

- **Undirected models** represent factorizations into a set of functions. Any set of nodes that are connected to each other in the graph are called a **clique**. Each clique is associated with a factor $\phi^{(i)}$ $(C^{(i)})$, which are functions and not probability distributions of any kind.

$$p(\mathrm{x}) = \frac{1}{Z} \prod_i \phi^{(i)} \, (C^{(i)})$$

  Here Z is the normalizing constant, defined to be the sum or integral over all states of the product of the ϕ functions, so that the distribution may be normalized.

- Being directed or undirected in not a property of a probability distribution, but rather a property of a particular description of a probability distribution.