Probability and Information Theory

Nishanth Nair

May 19, 2020

Overview

- Probability theory allows us to make uncertain statements and to reason in the presence of uncertainty.
- Information Theory enables us to quantify the amount of uncertainty in a probability distribution.
- Machine Learning must always deal with uncertain and sometimes stochastic (non deterministic) quantities.
- Possible sources of uncertainty:
  - Inherent stochasticity
  - Incomplete observability
  - Incomplete modelling
- In many cases,it is more practical to use a simple but uncertain rule rather than a complex but certain one (even if its true and deterministic).
  - For example "Most birds fly" is more useful and cheap to develop than a complex rule of the form "birds fly except for very young that are learning... sick or injured etc..etc.
- **Two views of probability**
  - **Frequentist Probability**:
    * We assume we havve a set of possible events each we assume occurs some number of times.
    * If we have N distinct events $x_1, x_2 ... x_N$, no two occuring simultaneously, and the events occur with frequencies $n_1, n_2 ... n_N$, we say, the probability of event x_i is given by:

$$P(x_i) = \frac{n_i}{\sum j = 1^N n_j}$$

  - **Bayesian Probability**
    * Probability represents a degree of belief.
    * It is possible that two different observers may assign different probabilities to the same event.
      · Also, the probability of an event is likely to change as we learn more about the event or the context of the event.
    * It can be thought of as an approximation of the frequentist version.

- All the same rules and formulae apply.
- We can view new knowledge as providing a better estimate of the relative frequencies.

Probability Review

- **Random variable**: some aspect of the world about which we may have uncertainty.It is a variable that can take on different values randomly.
  - Notation:
    - X - Random Variable
    - x - event, an observation or set of outcomes.
  - A random variable can be discrete or continuous.
  - **Discrete random variable**:Has a finite or countably infinite number of states.
  - **Continuous random variable**:Associated to a real value.
- **Probability distribution**:
  - It is a description of how likely a random variable is to take on one of its possible states.
  - A probability distribution over discrete variables is described using a **probability mass function(pmf)**
    - To be a pmf on a random variable X, a function P must satisfy the following conditions:
      - The domain of P must be the set of all possible states of X
      - $\forall x \in X, 0 \leq P(x) \leq 1$
      - $\sum_{x \in X} P(x) = 1$
  - A probability distribution over continuous variables is described using a **probability density function(pdf)**
    - To be a pdf on a random variable X, a function P must satisfy the following conditions:
      - The domain of P must be the set of all possible states of X
      - $\forall x \in X, P(x) \geq 0$
      - $\int P(x)dx = 1$
    - A pdf does not give the probability of a state directly, instead,it gives the probability of landing in an infinitesimal region.
    - The probability that x lies in a set S is given by the integral of P(x) over that set.
- **Joint Distributions**:
  - Defined over a set of random variables, instead of a single random variable
  - A joint distribution is a probabilistic model.
- **Uniform distribution**:
  - where all probabilities are equal.

- Describes the state where we know the least about the possible outcomes.
- **Discrete case**: Assume a single random variable X with k different states, then, the uniform distribution on X is defined as:

$$P(X = x_i) = \frac{1}{k}$$

- **Continuous case**: Uniform distribution of X in the interval [a,b] where b > a is defined as:

$$U(X : a, b) = \begin{cases} 0 & X \notin [a, b] \\ \frac{1}{b-a} & X \in [a, b] \end{cases}$$

- **Marginal Probability Distribution**:
  - Sum over one (or more) variables (marginalization)
  - If we know the probability distribution over a set of variables, the probability distribution over a subset of the variables is called the marginal probability distribution.

- **Conditional Probability**:
  - $P(A|B) = \frac{P(A,B)}{P(B)}$

- **Chain Rule**:
  - $P(A, B) = P(A|B).P(B)$
  - Generally, $P(X_1, X_2 ... X_n) = P(X_1|X_2..X_n).P(X_2...X_n)$
  - Example,

$$\begin{aligned} P(A, B, C) &= P(A|B, C).P(B, C) \\ &= P(A|B, C).P(B|C).P(C) \end{aligned}$$

  - Alternatively

$$\begin{aligned} P(A, B, C) &= P(B, A, C) \\ &= P(B|A, C).P(A, C) \\ &= P(B|A, C).P(A|C).P(C) \end{aligned}$$

- **Independence**: Two random variables X and Y are independent if:

$$\forall x \in X, y \in Y, P(X = x, Y = y) = P(X = x).P(Y = y)$$

- **Conditional Independence**: Two random variables X and Y are conditionally independent given Z if

$$\forall x \in X, y \in Y, z \in Z, P(X = x, Y = y|Z = z) = P(X = x|Z = z).P(Y = y|Z = z)$$

- **Bayes's Rule**:
  - Update our belief about X, given evidence E.
  - Posterior = $\frac{\text{Prior . Likelihood}}{\text{Normalizer}}$

$$
\begin{aligned}
P(X|E) &= \frac{P(X,E)}{P(E)} \\
&= \frac{P(E,X)}{P(E)} \\
&= \frac{P(X).P(E|X)}{P(E)}
\end{aligned}
$$

  - P(X) - Prior
  - P(X|E) - Posterior
  - P(E|X) - Likelihood
  - P(E) - Normalizer
- **Alternate form**:

$$
P(X|E) = \frac{P(X).P(E|X)}{P(E|X).P(X) + P(E|\bar{X}).P(\bar{X})}
$$

  - Posterior, P(X|E): Probability of X after taking into account E, for and against X
  - Prior, P(X)
  - Prior belief against X,$P(\bar{X})$ = 1-P(X)
  - Likelihood, $P(E|X)$: Belief in E, given that X is true
  - Likelihood, $P(E|\bar{X})$: Belief in E, given that X is false
- **General form**:

$$
P(X = x_k|E = e_j) = \frac{P(X = x_k).P(E = e_j|X = x_k)}{\sum_i P(E = e_j|X = x_i).P(X = x_i)}
$$

Log rules review

- $\log_a(bc) = \log_a(b) + \log_a(c)$
- $\log_a(b^c) = c\log_a(b)$
- $\log_a(\frac{1}{b}) = -1\log_a(b)$
- $\log_a(1) = 0$
- $\log_a(a) = 1$
- $\log_a(a^r) = r$
- $\log_{\frac{1}{a}}(b) = -1\log_a(b)$
- $\log_{a^m}(b^n) = \frac{n}{m}\log_a(b), m \neq 0$
- $\log_a(b).\log_b(c) = \log_a(c)$
- $\log_b(a) = \frac{1}{\log_a(b)}$

Information Theory

- Information theory attempts to quantify how much information is present in a signal.
- Ignoring any particular feature of an event, we can develop a usable measure of the information we get using the probability of occurance of the event.
- The basic intiution is that learning an unlikely event has occured is more informative than learning a likely evvent has occured.
- The measure of information of an event, say $I(x)$ for an event x having probability of occurence $P(x)$, should havve the following properties:
  - Information should be non-negative. $I(x) \geq 0$
  - If probability of occurence is 1, we get no information, $I(x) = 0$, when $P(x) = 1$
  - If two independent events occur,Information should be the sum of individual information.
  $$I(x_1, x_2) = I(x_1) + I(x_2)$$
  - Information measure should be a continuos and monotonic function of the probability.
- Based on the above, we can derive the **self information** of an event X=x to be: $I(x) = -log(P(x))$
  - With log_e, the unit of information is **nats**.(from natural)
    * One nat is the information gained by observing an event with probability 1/e
  - With log_2, the unit of information is **bits** or **shannons** (from binary)
  - log_3 are trits (from trinary)
  - log_{10} are Hartleys
- Units can be changed by changing the base as follows:

$$log_{b_2}(x) = log_{b2}(b_1).log_{b_1}(x), \text{using the formula } \log_a(b).\log_b(c) = \log_a(c)$$

Pointwise Mutual Information (PMI)

- PMI is a measure of how much knowing one outcome tells you about another.

$$\text{PMI}(x, y) = \log_2 \frac{p(x, y)}{p(x)\ p(y)}$$

- PMI measures the chance two outcomes tend to co-occur (the numerator) relative to the chance they would co-occur if they were independent events (the denominator).
- The $\log_2$ makes it easier to reason about very large or very small values of this ratio - and let's us give it a unit: bits
- If X and Y are independent, then $\text{PMI}(x, y) = 0$ for all values of x and y.
- "Point-wise" refers to the fact that we're picking single outcomes for "x" and "y" (i.e. x = "raining", y = "cloudy").

– Without the point-wise (i.e. just "mutual information") refers to the average (expected value) point-wise mutual information between all possible assignments to x and y.

Entropy

- Entropy is the notion of how uncertain the outcome of some experiment is.
- The more uncertain or the more spread out the distribution, the higher the entropy.
- Self information only deals with a single outcome. We can quantify the amount of uncertainty in an entire probability distribution using Entropy.
- Mathematically, for the discrete case, if we have a probability distribution, $P(X) = \{P(x_1), P(x_2)..., P(x_n)\}$, **Shannons Entropy**, H(x) is defined as:

  $H(X) = -\sum_i P(x_i).log_2P(x_i) = -E[log_2P(x_i)] = E[I(x)]$, where E is the expected value

- In the continuous case,H(X) is called the **Differential entropy**

$$H(X) = -\int P(x)log_2(P(x))dx$$

- Entropy, is thus the expected value of the information of the distribution.
- This gives us a **lower bound** on the number of bits (or nats etc) needed on average to encode symbols drawn from a distribution with probability P.
- Note: Entropy can be greater than or equal to 0.It does not have an upper bound.
- Distributions that are nearly deterministic have low entropy.
- Distributions close to a uniform distribution have high entropy.

Interpreting Entropy

- Imagine you want to send one of two messages to your friend, message A or message B.
- Imagine sending A and B were equally likely: $P(A) = P(B) = 0.5$, so you decide on the following code:

$$A \to 0$$
$$B \to 1$$

- Since there are only two options, a single bit will suffice.
- Note that 1 bit is equal to:

$$\begin{aligned} 1bit &= -log_2(1/2) \\ &= -(1/2)log_2(1/2) - (1/2)log_2(1/2) \\ &= -P(A)logP(A) - P(B)logP(B) \\ &= H(X), \text{ where } P(x) = 0.5 \end{aligned}$$

Example 2: - imagine you want to send one of three messages $m \sim M$:

$$P(A) = 0.5$$
$$P(B) = 0.25$$
$$P(C) = 0.25$$

- Since A is sent more often, we might want to give it a shorter code to save bandwidth. So we could try:

$$A \to 1$$
$$B \to 01$$
$$C \to 11$$

- Number of bits this code uses on average:

$$0.5 \times 1 \text{ bit} + 0.25 \times 2 \text{ bits} + 0.25 \times 2 \text{ bits} = 1.5 \text{ bits}$$

- Entropy of the distribution:

$$H(M) = -0.5log_2(0.5) - 0.25log_2(0.25) - 0.25log_2(0.25) = 1.5 \text{ bits}$$

- It turns out that this code is optimal, and in general the entropy $H(M)$ is the fewest number of bits on average that any code can use to send messages from the distribution M.
  - If we take bits to mean information, then the entropy is the minimum amount of information needed (on average) to uniquely encode messages $m \sim M$.

Cross Entropy

- Suppose we have a finite sample of messages (introducing some variance), and we train a machine learning model (introducing some bias) to estimate the true probabilities.
- Let the predicted distribution be $Q(X)$ and the true distribution be $P(X)$.
- Now we generate a code based on $Q(X)$, and use it to encode real messages (which come from $P(X)$).
- How many bits do we use, on average?
- If we design an optimal code for Q, we use $-\log_2 Q(x)$ bits for message x.
- Then we average this over $x \sim P$ to get:

$$\mathrm{CE}(P, Q) = \sum_x -P(x) \log_2 Q(x) = \mathrm{E}_{x \sim P(x)} [-\log_2 Q(x)]$$

- Since we "crossed" the code from Q and used it on P, this is known as the **cross-entropy**.
- The code trained on Q can't possibly be better than the optimal code on P itself. This gives us:

$$\mathrm{CE}(P, Q) \geq H(P)$$

- **ML Context**
  - Cross entropy is the most commonly used loss function in machine learning.
  - In unsupervised learning (density estimation), we use it exactly as-is, with x as the data.
  - In supervised learning,
    - We take the random variable to be the label y,
    - and take our distributions to be conditional ones: $P(y \mid x)$ and $Q(y \mid x)$:
    - This gives us:

$$\text{CE}(P, Q)(x) = \sum_{y'} -P(y' \mid x) \log_2 Q(y' \mid x)$$

- It's common to average over x and to approximate $P(y \mid x)$ with discrete samples (x, y) from a test set T, in which case we get:

$$\text{CE}(P, Q) \approx \frac{1}{|T|} \sum_{(x,y) \in T} \sum_{y'} -\mathbb{1}[y = y'] \log_2 Q(y' \mid x) = \frac{1}{|T|} \sum_{(x,y) \in T} -\log_2 Q(y \mid x)$$

- We'll commonly also write this using natural logarithms, but you can always convert between the two by the formula:

$$\log_2(x) = \log_2(e) \cdot \ln(x)$$

Kullback-Leibler divergence(KL Divergence)

- Entropy is the average number of bits needed if we design the code using the true probability distribution, P(x)
- Cross entropy is the average number of bits needed if we design the code with Q(x) (i.e a trained model or predicted probability) but end up sending them with probability, P(X) (i.e test set or true probability)
- KL Divergence is the difference between these quantities.
  - It is a measure of how different two probability distributions are.
  - The more Q differs from P, the worse the penalty would be, and thus the higher the KL divergence.
- Formally, if we have 2 seperate probability distributions P(X) and Q(X) over the same random variable X, we can measure how different these two distributions are using KL Divergence.

$$\begin{aligned}
D_{KL}(P \parallel Q) &= \mathrm{E}_{x \sim P(x)} \left[-\log_2 \frac{Q(x)}{P(x)} \right] \\
&= \mathrm{E}_{x \sim P(x)} \left[-\log_2 Q(x) - \log_2 P(x) \right] \\
&= \mathrm{E}_{x \sim P(x)} \left[-\log_2 Q(x) \right] - \mathrm{E}_{x \sim P(x)} \left[\log_2 P(x) \right] \\
&= CE(P, Q) - H(P)
\end{aligned}$$

- KL divergence is 0 iff P and Q are the same distribution in the discrete case and equal almost everywhere in the continuous case. - It is not symmetric: $D_{KL}(P \mid\mid Q) \neq D_{KL}(Q \mid\mid P)$ - Since it is not symmetric, it cannot be used as a distance measure for optimization. - KL divergence is a useful measure of similarity. - KL divergence is sometimes called relative entropy. $-D_{KL}(P \mid\mid Q)$ is relative entropy.

- **ML Context**
  - KL divergence measures the "avoidable" error.
    - When our model is perfect,the KL divergence goes to zero. (Q = P)
  - In general, the cross-entropy loss - and prediction accuracy - will not be zero, but will be equal to the entropy $H(P)$.
  - This "unavoidable" error is the **Bayes error rate** for the underlying task.
    - Bayes error rate is the lowest possible error rate for any classifier of a random outcome.

Structured Probablistic Models

- When we represent the factorization of a probability distribution with a graph,$\mathcal{G}$, we call it a structured probablistic model or graphical model.
- This factorization greatly reduces the number of parameters needed to describe the distribution.
- In the graph, each node corresponds to a random variable and an edge means the probability distribution is able to represent direct interactions between the two random variables.
- **Directed Graphical models**: Represent factorizations into conditional probability distributions.

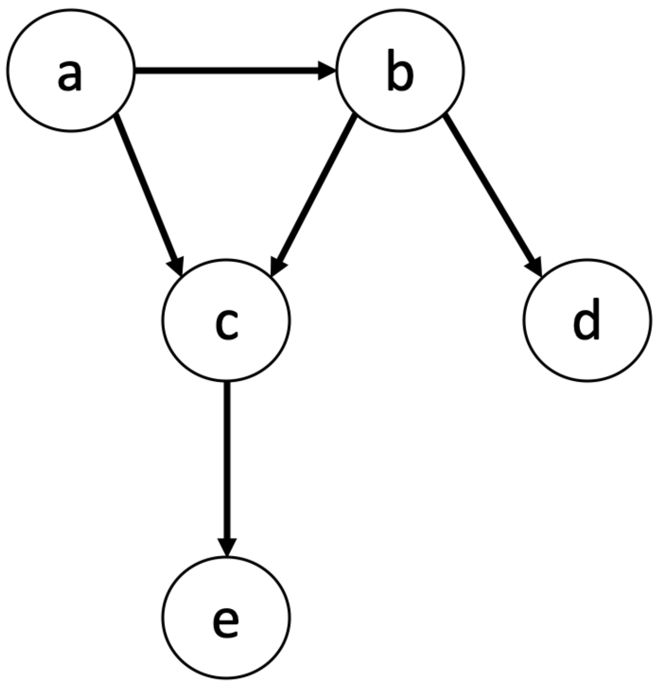

Figure 1: Directed graphical model

- The above corresponds to a probability distribution that can be factored as follows:

$$P(a, b, c, d, e) = P(a).P(b \mid a).P(c \mid a, b).P(d \mid b).P(e \mid c)$$

- **Undirected Graphical models**:
  - Represent factorizations into a set of functions.
  - Unlike the directed case, these functions are not probability distributions.
  - Any set of nodes that are all connected to each other in $\mathcal{G}$ is called a **Clique**,$\mathcal{C}$.
  - Each Clique, $\mathcal{C}^{(i)}$ in an undirected model is associated to a factor $\phi_i(\mathcal{C}^{(i)})$
  - $\phi_i(\mathcal{C}^{(i)})$ must be non-negative. There is no constraint that it shoulld sum to 1 like a probability function.

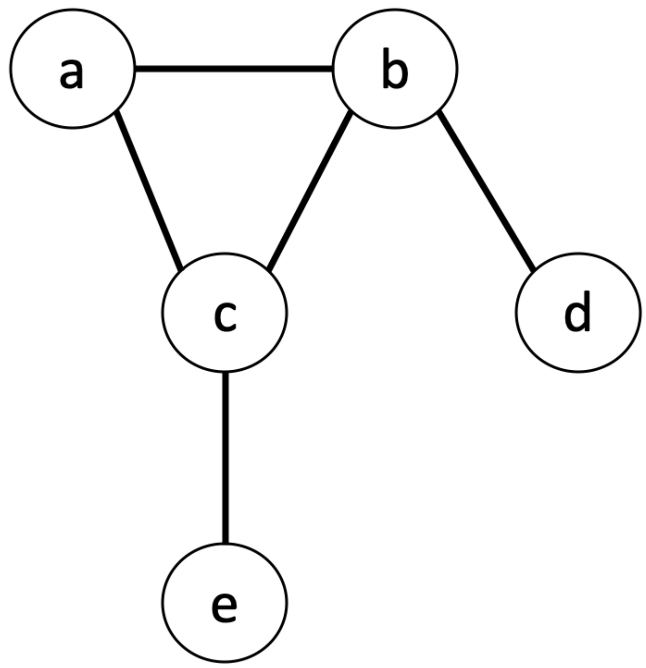

Figure 2: Undirected graphical model

- The above corresponds to a probability distribution that can be factored as follows:

$$P(a, b, c, d, e) = \frac{1}{Z}.\phi_1(a, b, c).\phi_2(b, d).\phi_3(c, e)$$

- Z is a normalizing constant defined to be the sum or integral over all states of the product of the ϕ functions.