1. What is Entropy, Cross-Entropy, Relative-Entropy and KL-divergence?

Those terminologies are from information theory. To better understand the basic concepts of information theory, let’s start from an example. Suppose there are many balls with four different colors in a box. The distinct colors are red, green, blue and black. One ball has been randomly selected from the box. Now you need to guess what the color is of that selected ball. You are allowed to ask questions multiple times until you get the selected color. According to the above description, what is the best strategy to get the selected color with the least guess times?

One of the best strategy is asking following questions: Is the color either red or green? If yes, then ask if the color is red? If No, then ask if the color is blue? No matter the answer of first question is, you will get the selected color correctly with only two questions. Do you know why this is the best strategy and why two questions are necessary to make sure you can have the color? That is due to Entropy!

Notice that this strategy is based on an assumption that we do not have any extra information about the ratio of different colors in the box, so we assume each ratio is ¼. Therefore, the best strategy tell us that the number of questions you need to get the right color is

More general, assume the ratio of a specific color is p, then the least number of questions to get the right color is . This value is called self-information which represents how many information a single event can provide. Generally, an event unlikely to happen will provide more information than an event likely to happen. In this game, this value means the number of questions you need to get a specific selected color.

Furthermore, the expected number of questions for all colors is . This value is called information entropy or entropy which represents the uncertainty of a system. The bigger the value, the more uncertain the system is. From the other perspective, it also represents how many efforts you need to eliminate the uncertainty of a system. In this game, the effort means the number of questions you need to have correct guess.

Based on the above description, we know the self-information for each color is .

The information entropy is .

Suppose you are provided an extra information about the color distribution: red:green:blue:black = 1/2:1/4:1/8:1/8. What is the best strategy then? Suppose we still use the previous strategy then the expected number of questions to get the selected color is .

However, if you use following strategy q, the expected number of questions is

Apparently, if you use the second strategy, you need less questions. The strategy distribution q can be arbitrary. However, the best strategy distribution is just the same as the real distribution.

When you involve a strategy to eliminate the uncertainty of a system, the cross entropy illustrates how many efforts you need to eliminate the uncertainty. The smaller the cross entropy the better your strategy is. The smallest cross entropy is approached when your strategy distribution is exactly the same as the real distribution.

How to compare the difference of two strategy then? The relative entropy is what we need in such case. The math formula of relative entropy is , which means the difference the cross entropy H(p, q) and entropy H(p) assuming p is the real distribution. The relative entropy has another name called KL-divergence. Notice that

The definition of KL(q||P) is the difference the cross entropy H(q, p) and entropy H(q) assuming q is the real distribution.

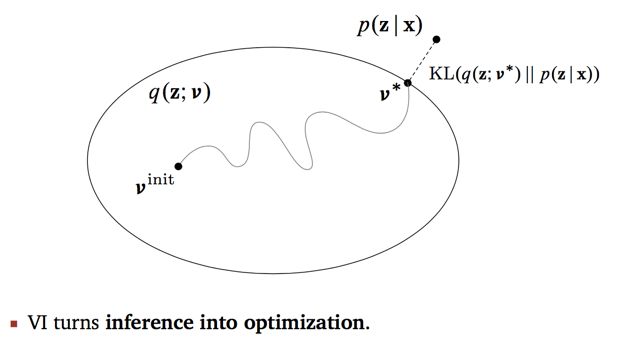
1. What is VI?

The inference in probabilistic models is often intractable. We can use a family of tractable distribution to approach the intractable distribution. By searching the distribution that is most similar or closest to the intractable distribution, we convert the inference problem to an optimization problem. This process is called variational inference. The metric to represent the distance or dissimilarity between two distributions is KL-divergence.

Usually, in a Bayesian model we have the observed variable X, the latent variable Z. The posterior is as following:

However, the marginalization over Z in the dominator is typically intractable. Therefore, the posterior is also intractable. VI is trying to find a distribution q(Z;v) to replace P(Z|X). The process of optimization is as follows. The optimization formula is:

Find the best parameter v of q to minimize the KL-divergence.



1. How to find the best distribution to replace the posterior?

Apparently, based on the above chart, the optimization problem is just finding a solution with the smallest KL-divergence. However, the issue is based on the above formula of KL-divergence, it still contains the p(z|x) which is intractable. So it is difficult to solve the optimization problem directly. Fortunately, we can convert solve above optimization problem by converting minimizing KL-divergence to maximizing ELBO which does not contain any intractable part.

1. Relationship between VI and MCMC?

In practice, the probabilistic models that we use are often quite complex. In fact, many interesting classes of models may not admit exact polynomial-time solutions at all, and for this reason, much research effort in machine learning is spent on developing algorithms that yield *approximate* solutions to the inference problem. This section begins our study of such algorithms.

There exist two main families of approximate algorithms: *variational* methods Variational inference methods take their name from the *calculus of variations*, which deals with optimizing functions that take other functions as arguments., which formulate inference as an optimization problem, as well as *sampling* methods, which produce answers by repeatedly generating random numbers from a distribution of interest.

MCMC is sampling method. VI is variational method.

1. Application of VI?