Variational Inference

Anna-Lena Popkes

February 23, 2019

Introduction

Variational inference is an important topic that is widely used in machine learning. For example, it's the basis for variational autoencoders. Also Bayesian learning often makes use variational of inference. To understand what variational inference is, how it works and why it's useful we will go through each point step by step.

What are latent variables?

A latent variable is the opposite of an *observed* variable. This means that a latent variable is not directly observed but inferred from other variables which are observed. This book provides a nice conceptual example:

Example:
Consider your overall health. There are multiple measurements we can use to assess health, were each measurement is looking at a certain physical property, for example blood pressure or body temperature. However, 'health' remains an abstract concept that cannot be measured directly. In this sense, health is a latent variable, whereas blood pressure and body temperature are observable variables.

What is variational inference?

Variational inference is a machine learning method which allows us to approximate probability distributions. In many real world problems we are faced with probability distributions that can't be computed. This especially often happens when a distribution involves latent variables. Therefore, we need strategies to approximate such distributions. Variational inference is one method for doing this. Several other methods exist which broadly fall into two classes: methods that rely on *stochastic* approximations (like Markov chain Monte Carlo sampling) and those that rely on deterministic approximations (like variational inference).

A key characteristic of variational inference is that it *reframes* the original problem (computing some probability distribution) into a simpler problem which can be solved. Different to this, Markov chain Monte Carlo sampling directly approximates the target distribution by sampling from it.

A high level example for variational inference is given here:

> **Example:**
> Variational inference is similar to what often happens when attending a presentation or lecture. Someone in the audience asks the presenter a very difficult question which she can't answer. Instead of answering the original, difficult question, the presenter reframes the question into an easier one which can be answered exactly.

Problem set-up

Suppose we have the following set-up:

- A set of observations $\boldsymbol{x} = x_1, ..., x_n$
- A set of latent variables $\boldsymbol{z} = z_1, ..., z_l$
- The joint probability distribution is given by $p(\boldsymbol{x}, \boldsymbol{z})$
- In many probabilistic models we are interested in the posterior distribution $p(\boldsymbol{z} \,|\, \boldsymbol{x})$ of the latent variables given the observed data $\boldsymbol{x}$:

$$p(\boldsymbol{z} \,|\, \boldsymbol{x}) = \frac{p(\boldsymbol{z}, \boldsymbol{x})}{p(\boldsymbol{x})}$$

This posterior distribution can be used for several purposes. For example, it can be used to provide point estimates for the latent variables.

Problem: For many models it's impossible to evaluate the posterior distribution or even to compute expected values with respect to the distribution. To evaluate $p(\boldsymbol{z} \,|\, \boldsymbol{x})$ we need to compute the denominator $p(\boldsymbol{x})$ which is called *evidence*. The evidence is given by $p(\boldsymbol{x}) = \int p(\boldsymbol{z}, \boldsymbol{x}) d\boldsymbol{z}$. For many models this integral cannot be computed or takes exponential time to compute (for example, the dimensionality of the latent variables might be too high).

Solution: We approximate $p(\boldsymbol{z} \,|\, \boldsymbol{x})$ using variational inference.

How does variational inference work?

To approximate $p(\boldsymbol{z}\,|\,\boldsymbol{x})$ we introduce a *variational distribution* over the latent variables $q(\boldsymbol{z})$. More precisely, we choose a *family of distributions* $\mathcal{Q}$ characterized by some parameters θ. For example, we could decide that our variational distribution belongs to the family of Gaussian distributions. In this case the parameters θ would be the mean and standard deviation of the Gaussian distribution. Each member $q(\boldsymbol{z}) \in \mathcal{Q}$ represents a candidate approximation to the true posterior $p(\boldsymbol{z}\,|\,\boldsymbol{x})$.

Our goal is to find the best candidate distribution. Or, to be more precise, to find the setting of the parameters θ that make our candidate $q(\boldsymbol{z})$ as similar as possible to $p(\boldsymbol{z}\,|\,\boldsymbol{x})$.

Of course the choice of the variational family $\mathcal{Q}$ has a large impact on the final result. The true posterior is often not contained in the variational family. However, we don't need to find the exact posterior. We just want to find a (very) good estimate.

KL divergence

We evaluate our candidate variational distribution using the *Kullback-Leibler divergence* which was introduced in this post. The KL divergence can be used to measure how similar $q(\boldsymbol{z})$ is to the target distribution $p(\boldsymbol{z}\,|\,\boldsymbol{x})$. To find the best variational distribution we minimize the KL divergence:

$$q_{\text{best}}(\boldsymbol{z}) = \arg \min_{q(\boldsymbol{z}) \in \mathcal{Q}} D_{KL}\big(q(\boldsymbol{z})\,||\,p(\boldsymbol{z}\,|\,\boldsymbol{x})\big)$$

The KL divergence is defined as:

$$D_{KL}\big(q(\boldsymbol{z})\,||\,p(\boldsymbol{z}\,|\,\boldsymbol{x})\big) = \int_{-\infty}^{\infty} q(\boldsymbol{z}) \log\big(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z}\,|\,\boldsymbol{x})}\big) = \mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z}\,|\,\boldsymbol{x})}\big)\big]$$

Problem: The KL divergence can't be computed. To see why we reformulate the definition of the KL divergence:

$$\begin{aligned}
\mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z}\,|\,\boldsymbol{x})}\big)\big] &= \mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(q(\boldsymbol{z})\big)\big] - \mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(p(\boldsymbol{z}\,|\,\boldsymbol{x})\big)\big] \\
&= \mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(q(\boldsymbol{z})\big)\big] - \mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(p(\boldsymbol{z},\boldsymbol{x})\big)\big] + \log\big(p(\boldsymbol{x})\big)
\end{aligned}$$

The last term in this equation is exactly the evidence we came across earlier $p(\boldsymbol{x}) = \int p(\boldsymbol{z},\boldsymbol{x}) d\boldsymbol{z}$ and which can't be computed.

Solution: Instead of minimizing the KL divergence, we minimize an alternative quantity which is equivalent up to an added constant. This is the so called *evidence lower bound* or *ELBO*:

$$\mathrm{ELBO}(q) = -\,\mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(q(\boldsymbol{z})\big)\big] + \mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(p(\boldsymbol{z},\boldsymbol{x})\big)\big]$$

When comparing the ELBO with the KL divergence we can see that the ELBO is simply the negative KL divergence plus our problematic evidence term $p(\boldsymbol{x})$. Maximizing the ELBO is equivalent to minimizing the KL divergence.

Evidence lower bound

To gain a deeper understanding of what it means to find the optimal candidate $q(\boldsymbol{z})$ we can rewrite the evidence lower bound:

$$\begin{aligned}\mathrm{ELBO}(q) &= -\,\mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(q(\boldsymbol{z})\big)\big] + \mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(p(\boldsymbol{z},\boldsymbol{x})\big)\big]\\ &= -\,\mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(q(\boldsymbol{z})\big)\big] + \mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(p(\boldsymbol{z})\big)\big] + \mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(p(\boldsymbol{x}\,|\,\boldsymbol{z})\big)\big]\\ &= \mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(p(\boldsymbol{x}\,|\,\boldsymbol{z})\big)\big] - D_{KL}\big(q(\boldsymbol{z})\,||\,p(\boldsymbol{z})\big)\end{aligned}$$

Let's take a closer look at the individual terms:

1. The first term $\mathbb{E}$_{$q(\boldsymbol{z})$}$\big[\log\big(p(\boldsymbol{x}\,|\,\boldsymbol{z})\big)\big]$ describes the probability of the data given the latent variables. By maximizing the ELBO, we encourage the optimization process to choose a candidate distribution $q(\boldsymbol{z})$ which explains the observed data well.

2. The second term $-D_{KL}\big(q(\boldsymbol{z})\,||\,p(\boldsymbol{z})\big)$ is the negative KL divergence between our variational distribution $q(\boldsymbol{z})$ and the prior distribution over the latent variables $p(\boldsymbol{z})$. Maximizing this term corresponds to minimizing the KL divergence. So the optimization process is encouraged to make the variational distribution similar to the prior distribution over the latent variables.

Why it's called evidence lower bound

The name 'evidence lower bound' comes from an important property of the ELBO: it provides a lower bound on the (log) evidence $p(\boldsymbol{x})$.

We already determined that

1. $D_{KL}\big(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})\big) = \mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(q(\boldsymbol{z})\big)\big] - \mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(p(\boldsymbol{z},\boldsymbol{x})\big)\big] + \log\big(p(\boldsymbol{x})\big)$

2. $\mathrm{ELBO}(q) = -\,\mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(q(\boldsymbol{z})\big)\big] + \mathbb{E}_{q(\boldsymbol{z})}\big[\log\big(p(\boldsymbol{z},\boldsymbol{x})\big)\big]$

Combining 1 and 2 gives us:

$$D_{KL}\big(q(\boldsymbol{z})\,||\,p(\boldsymbol{z}\,|\,\boldsymbol{x})\big) = \log\big(p(\boldsymbol{x})\big) - \mathrm{ELBO}(q)$$
$$\Leftrightarrow \log\big(p(\boldsymbol{x})\big) = D_{KL}\big(q(\boldsymbol{z})\,||\,p(\boldsymbol{z}\,|\,\boldsymbol{x})\big) + \mathrm{ELBO}(q)$$

Because the KL divergence is always non-negative, i.e. $D_{KL}(\cdot) \geq 0$ we know that

$$log\big(p(\boldsymbol{x})\big) \geq \mathrm{ELBO}(q)$$

Hence, the ELBO provides a lower bound on the (log) evidence $p(\boldsymbol{x})$.

Example variational family

The choice of the variational family $\mathcal{Q}$, or rather its complexity determines how complex it will be to optimize the ELBO. In simple terms: if the variational family is very complex, it will be more difficult to solve our optimization problem. One way of restricting the variational family $\mathcal{Q}$ is to choose a parametric distribution $q(\boldsymbol{z}\,|\,\theta)$ which is governed by a set of parameters θ. For example, we could choose a Gaussian distribution.

Another popular approach is the so called *mean field approximation*. This approach assumes that the variational distribution *factorizes* with respect to some partition of the latent variables. Mean field approximation works as follows:

1. We partition the latent variables $\boldsymbol{z}$ into M disjoint groups $\boldsymbol{z}_i$ with $i = 1, ..., M$.
2. We then assume that the variational distribution factorizes with respect to these groups: $q(\boldsymbol{z}) = \prod_{i=1}^{M} q_i(\boldsymbol{z}_i)$

We don't make any further assumptions about the form of the different q_i. They might all be Gaussian distributions or a combination of different distributions. This offers a lot of flexibility.

The goal of variational inference is now to find the distribution $q(\boldsymbol{z})$ of the form $q(\boldsymbol{z}) = \prod_{i=1}^{M} q_i(\boldsymbol{z}_i)$ which maximizes the ELBO. To be more precise, we need to optimize the ELBO with respect to all distributions $q_i(\boldsymbol{z}_i)$. This is done by optimizing with respect to each of the factors in turn. More details on this procedure can be found in Bishop's 'Pattern Recognition and Machine Learning' book in chapter 10.1 (a link to the book is given below).

Sources and further reading

- Book: Pattern Recognition and Machine Learning
- Paper: Variational Inference: A Review for Statisticians